

Josafhat Salinas Ruíz
Osva Antonio Montesinos López
Gabriela Hernández Ramírez
Jose Crossa Hiriart

Generalized Linear Mixed Models with Applications in Agriculture and Biology

OPEN ACCESS


 Springer

Generalized Linear Mixed Models with Applications in Agriculture and Biology


Josafhat Salinas Ruíz • Osva! Antonio Montesinos
López • Gabriela Hernández Ramírez
Jose Crossa Hiriart


Generalized Linear Mixed Models with Applications in Agriculture and Biology

 Springer

Josafhat Salinas Ruíz 
Innovación Agroalimentaria Sustentable
Colegio de Postgraduados
Córdoba, Mexico

Osva Antonio Montesinos López 
Facultad de Telemática
University of Colima
Colima, Mexico

Gabriela Hernández Ramírez 
Ciencia de los Alimentos y Biotecnología
Tecnológico Nacional de México/ITS
de Tierra Blanca, Mexico

Jose Crossa Hiriart 
International Maize and Wheat Improvement
Center (CIMMYT)
Texcoco, Mexico



ISBN 978-3-031-32799-5 ISBN 978-3-031-32800-8 (eBook)
<https://doi.org/10.1007/978-3-031-32800-8>

The translation was done with the help of artificial intelligence (machine translation by the service DeepL.com). A subsequent human revision was done primarily in terms of content.

© The Editor(s) (if applicable) and The Author(s) 2023. This book is an open access publication.

Open Access This book is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this book are included in the book's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the book's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors, and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Foreword

This book is another example of CIMMYT's commitment to scientific advancement, and comprehensive and inclusive knowledge sharing and dissemination. By using alternative statistical models and methods to describe and analyze data sets from different disciplines, such as biology and agriculture, the book facilitates the adoption and effective use of these tools by publicly funded researchers and practitioners of national agricultural research extension systems (NARES) and universities across the Global South.

The authors aim to offer different and new models, methods, and techniques to agricultural scientists who often lack the resources to adopt these tools or face practical constraints when analyzing different types of data.

This work would not be possible with the continuous support of CIMMYT's outstanding partners and donors who invest in non-profit frontier research for the benefit of millions of farmers and low-income communities worldwide. For that reason, it could not be more fitting for this book to be published as an open access resource for the international community to benefit from. I trust that this publication will greatly contribute to accelerate the development and deployment of resource-efficient and nutritious crops for a food secure future.

Bram Govaerts
Director General, CIMMYT

Acknowledgments

The work presented in this book described models and methods for improving the statistical analyses of continuous, ordinal, and count data usually collected in agriculture and biology. The authors are grateful to the past and present CIMMYT Directors General, Deputy Directors of Research, Deputy Directors of Administration, Directors of Research Programs, and other Administration offices and Laboratories of CIMMYT for their continuous and firm support of biometrical genetics, and statistics research, training, and service in support of CIMMYT's mission: "maize and wheat science for improved livelihoods."

This work was made possible with support from the CGIAR Research Programs on Wheat and Maize (wheat.org, maize.org), and many funders including Australia, United Kingdom (DFID), USA (USAID), South Africa, China, Mexico (SAGARPA), Canada, India, Korea, Norway, Switzerland, France, Japan, New Zealand, Sweden, and the World Bank. We thank the financial support of the Mexico Government throughout MASAGRO and several other regional projects and close collaboration with numerous Mexican researchers.

We acknowledge the financial support provided by the (1) Bill and Melinda Gates Foundation (INV-003439 BMGF/FCDO Accelerating Genetic Gains in Maize and Wheat for Improved Livelihoods [AG2MW]) as well as (2) USAID projects (Amend. No. 9 MTO 069033, USAID-CIMMYT Wheat/AGGMW, AGG-Maize Supplementary Project, AGG [Stress Tolerant Maize for Africa]).

Very special recognition is given to Bill and Melinda Gates Foundation for providing the Open Access fee of this book.

We are also thankful for the financial support provided by the (1) Foundations for Research Levy on Agricultural Products (FFL) and the Agricultural Agreement Research Fund (JA) in Norway through NFR grant 267806, (2) Sveriges Llantbruksuniversitet (Swedish University of Agricultural Sciences) Department of

The original version of this book has been revised. The Acknowledgment section which was inadvertently omitted after the Foreword has now been included.

Plant Breeding, Sundsvägen 10, 23053 Alnarp, Sweden, (3) CIMMYT CRP, (4) the Consejo Nacional de Tecnología y Ciencia (CONACYT) of México, and (5) Universidad de Colima of Mexico.

We highly appreciate and thank the several students and professors at the Universidad de Colima and students as well professors from the Colegio de Post-Graduados (COLPOS) who tested and made suggestions on early version of the material covered in the book; their many useful suggestions had a direct impact on the current organization and the technical content of the book.

Contents

1	Elements of Generalized Linear Mixed Models	1
1.1	Introduction to Linear Models	1
1.2	Regression Models	2
1.2.1	Simple Linear Regression	2
1.2.2	Multiple Linear Regression	4
1.3	Analysis of Variance Models	6
1.3.1	One-Way Analysis of Variance	6
1.3.2	Two-Way Nested Analysis of Variance	9
1.3.3	Two-Way Analysis of Variance with Interaction	13
1.4	Analysis of Covariance (ANCOVA)	18
1.5	Mixed Models	22
1.5.1	Introduction	22
1.5.2	Mixed Models	23
1.5.3	Distribution of the Response Variable Conditional on Random Effects ($y \mathbf{b}$)	25
1.5.4	Types of Factors and Their Related Effects on LMMs	26
1.5.5	Nested Versus Crossed Factors and Their Corresponding Effects	27
1.5.6	Estimation Methods	27
1.5.7	One-Way Random Effects Model	30
1.5.8	Analysis of Variance Model of a Randomized Block Design	31
1.6	Exercises	35
	Appendix	40
2	Generalized Linear Models	43
2.1	Introduction	43
2.2	Components of a GLM	44
2.2.1	The Random Component	44

2.2.2	The Systematic Component	44
2.2.3	Predictor's Link Function η	45
2.3	Assumptions of a GLM	48
2.4	Estimation and Inference of a GLM	48
2.5	Specification of a GLM	49
2.5.1	Continuous Normal Response Variable	49
2.5.2	Binary Logistic Regression	51
2.5.3	Poisson Regression	59
2.5.4	Gamma Regression	67
2.5.5	Beta Regression	72
2.6	Exercises	76
	Appendix	83
3	Objectives of Inference for Stochastic Models	85
3.1	Three Aspects to Consider for an Inference	86
3.1.1	Data Scale in the Modeling Process Versus Original Data	87
3.1.2	Inference Space	87
3.1.3	Inference Based on Marginal and Conditional Models	88
3.2	Illustrative Examples of the Data Scale and the Model Scale	88
3.3	Fixed and Random Effects in the Inference Space	101
3.3.1	A Broad Inference Space or a Population Inference	101
3.3.2	Mixed Models with a Normal Response	103
3.4	Marginal and Conditional Models	105
3.4.1	Marginal Versus Conditional Models	105
3.4.2	Normal Distribution	107
3.4.3	Non-normal Distribution	108
3.5	Exercises	111
4	Generalized Linear Mixed Models for Non-normal Responses	113
4.1	Introduction	113
4.2	A Brief Description of Linear Mixed Models (LMMs)	114
4.3	Generalized Linear Mixed Models	114
4.4	The Inverse Link Function	116
4.5	The Variance Function	117
4.6	Specification of a GLMM	117
4.7	Estimation of the Dispersion Parameter	119
4.8	Estimation and Inference in Generalized Linear Mixed Models	123
4.8.1	Estimation	123
4.8.2	Inference	125
4.9	Fitting the Model	126
4.10	Exercises	126

- 5 Generalized Linear Mixed Models for Counts** 129
 - 5.1 Introduction 129
 - 5.2 The Poisson Model 129
 - 5.2.1 CRD with a Poisson Response 131
 - 5.2.2 Example 2: CRDs with Poisson Response 134
 - 5.2.3 Example 3: Control of Weeds in Cereal Crops
in an RCBD 138
 - 5.2.4 Overdispersion in Poisson Data 142
 - 5.2.5 Factorial Designs 149
 - 5.2.6 Latin Square (LS) Design 157
 - 5.3 Exercises 170
 - Appendix 1 178
- 6 Generalized Linear Mixed Models for Proportions
and Percentages** 209
 - 6.1 Response Variables as Ratios and Percentages 209
 - 6.2 Analysis of Discrete Proportions: Binary and Binomial
Responses 210
 - 6.2.1 Completely Randomized Design (CRD): Methylation
Experiment 210
 - 6.3 Factorial Design in a Randomized Complete Block Design
(RCBD) with Binomial Data: Toxic Effect of Different
Treatments on Two Species of Fleas 217
 - 6.4 A Split-Plot Design in an RCBD with a Normal Response 219
 - 6.4.1 An RCBD Split Plot with Binomial Data: Carrot Fly
Larval Infestation of Carrots 221
 - 6.5 A Split-Split Plot in an RCBD:- In Vitro Germination
of Seeds 232
 - 6.6 Alternative Link Functions for Binomial Data 237
 - 6.6.1 Probit Link: A Split-Split Plot in an RCBD
with a Binomial Response 238
 - 6.6.2 Complementary Log-Log Link Function: A Split Plot
in an RCBD with a Binomial Response 239
 - 6.7 Percentages 242
 - 6.7.1 RCBD: Dead Aphid Rate 242
 - 6.7.2 RCBD: Percentage of Quality Malt 245
 - 6.7.3 A Split Plot in an RCBD: Cockroach Mortality
(*Blattella germanica*) 247
 - 6.7.4 A Split-Plot Design in an RCBD: Percentage Disease
Inhibition 250
 - 6.7.5 Randomized Complete Block Design with a Binomial
Response with Multiple Variance Components 253
 - 6.8 Exercises 257
 - Appendix 265

7 Time of Occurrence of an Event of Interest 279

7.1 Introduction 279

7.2 Generalized Linear Mixed Models with a Gamma Response 280

7.2.1 CRD: Estrus Induction in Pelibuey Ewes 280

7.2.2 Randomized Complete Block Design (RCBD):
Itch Relief Drugs 282

7.2.3 Factorial Design: Insect Survival Time 284

7.2.4 A Split Plot with a Factorial Structure on a Large
Plot in a Completely Randomized Design (CRD) 287

7.3 Survival Analysis 291

7.3.1 Concepts and Definitions 293

7.3.2 CRD: *Aedes aegypti* 295

7.3.3 RCBD: *Aedes aegypti* 297

7.4 Exercises 301

Appendix 1 306

**8 Generalized Linear Mixed Models for Categorical and Ordinal
Responses 321**

8.1 Introduction 321

8.2 Concepts and Definitions 322

8.3 Cumulative Logit Models (Proportional Odds Models) 323

8.3.1 Complete Randomize Design (CRD)
with a Multinomial Response: Ordinal 324

8.3.2 Randomized Complete Block Design (RCBD)
with a Multinomial Response: Ordinal 328

8.4 Cumulative Probit Models 336

8.5 Effect of Judges' Experience on Canned Bean Quality
Ratings 338

8.6 Generalized Logit Models: Nominal Response Variables 348

8.6.1 CRDs with a Nominal Multinomial Response 350

8.6.2 CRD: Cheese Tasting 353

8.7 Exercises 357

Appendix 374

9 Generalized Linear Mixed Models for Repeated Measurements 377

9.1 Introduction 377

9.2 Example of Turf Quality 378

9.3 Effect of Insecticides on Aphid Growth 382

9.4 Manufacture of Livestock Feed 387

9.5 Characterization of Spatial and Temporal Variations
in Fecal Coliform Density 390

9.6 Log-Normal Distribution 395

9.6.1 Emission of Nitrous Oxide (N₂O) in Beef Cattle
Manure with Different Percentages of Crude Protein
in the Diet 397

9.7 Effect of a Chemical Salt on the Percentage Inhibition of the *Fusarium sp.* 398

9.8 Carbon Dioxide (CO₂) Emission as a Function of Soil Moisture and Microbial Activity 402

9.9 Effect of Soil Compaction and Soil Moisture on Microbial Activity 407

9.10 Joint Model for Binary and Poisson Data 409

9.11 Exercises 413

Appendix 416

References 425

Chapter 1

Elements of Generalized Linear Mixed Models



1.1 Introduction to Linear Models

Linear models are commonly used to describe and analyze datasets from different research areas, such as biological, agricultural, social, and so on. A linear model aims to best represent/describe the nature of a dataset. A model is usually made up of factors or a series of factors that can be nominal or discrete variables (sex, year, etc.) or continuous variables (age, height, etc.), which have an effect on the observed data. Linear models are the most commonly used statistical models for estimating and predicting a response based on a set of observations.

Linear models get their name because they are linear in the model parameters. The general form of a linear model is given by

$$y = X\beta + \epsilon \tag{1.1}$$

where y is the vector of dimension $n \times 1$ observed responses, X is the design matrix of $n \times (p + 1)$ fixed constants, β is the vector of $(p + 1) \times 1$ parameters to be estimated (unknown), and ϵ is the vector of $n \times 1$ random errors. Linearity arises because the mean response of vector y is linear to the vector of unknown parameters β . Mathematically, this is demonstrated by obtaining the first derivative of the predictor with respect to β , and, if after derivation it is still a function of any of the beta parameters, then the model is said to be nonlinear; otherwise, it is a linear model. In this case, the derivative of the predictor (1.1) with respect to beta is equal to X , so, mathematically, the model in (1.1) is linear, since after derivation, the predictor no longer depends on the β parameters.

Several models used in statistics are examples of the general linear model $y = X\beta + \epsilon$. These include regression models and analysis of variance (ANOVA) models. Regression models generally refer to those in which the design matrix X is

of a full column rank, whereas in analysis of variance models, the design matrix \mathbf{X} is not of a full column rank. Some linear models are briefly described in the following sections.

1.2 Regression Models

Linear models are often used to model the relationship between a variable, known as the response or dependent variable, y , and one or more predictors, known as independent or explanatory variables, X_1, X_2, \dots, X_p .

1.2.1 Simple Linear Regression

Consider a model in which a response variable y is linearly related to an explanatory variable X_1 via

$$y_i = \beta_0 + \beta_1 X_{1i} + \varepsilon_i$$

where ε_i are uncorrelated random errors ($i = 1, 2, \dots, n$) which are commonly assumed to be normally distributed with mean 0 and variance constant $\sigma^2 > 0$, $\varepsilon_i \sim N(0, \sigma^2)$. If $X_{11}, X_{12}, \dots, X_{1n}$ are constant (fixed), then this is a general linear model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ where

$$\mathbf{y}_{n \times 1} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad \mathbf{X}_{n \times 2} = \begin{pmatrix} 1 & X_{11} \\ 1 & X_{12} \\ \vdots & \vdots \\ 1 & X_{1n} \end{pmatrix}, \quad \boldsymbol{\beta}_{2 \times 1} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}, \quad \boldsymbol{\epsilon}_{n \times 1} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

Example Let us consider the relationship between the performance test scores and tissue concentration of lysergic acid diethylamide commonly known as LSD (from German Lysergsäure-diethylamid) in a group of volunteers who received the drug

Table 1.1 Average mathematical test scores and LSD tissue concentrations

Tissue concentration of LSD	Mathematical average
1.17	78.93
2.97	58.20
3.26	67.47
4.69	37.47
5.83	45.65
6.00	32.92
6.41	29.97

Table 1.2 Results of the simple regression analysis

(a) Type III tests of fixed effects					
Effect	Num DF	Den DF	F-value	Pr > F	
Conc	1	5	35.93	0.0019	
(b) Parameter estimates					
Effect	Estimate	Standard error	DF	t-value	Pr > t
Intercept	89.1239	7.0475	5	12.65	<0.0001
Conc	-9.0095	1.5031	5	-5.99	0.0019
Scale ($\hat{\sigma}^2$)	50.7763	32.1137	.	.	.

(Wagner et al. 1968). The average scores on the mathematical test and the LSD tissue concentrations are shown in Table 1.1.

The components of this regression model are as follows:

$$\text{Distribution: } y_i \sim N(\eta_i, \sigma^2)$$

$$\text{Linear predictor: } \eta_i = \beta_0 + \beta_1 \times \text{Conc}_i$$

$$\text{Link function: } \mu_i = \eta_i \text{ (identity)}$$

The syntax for performing a simple linear regression using the GLIMMIX procedure in Statistical Analysis Software (SAS) is as follows:

```
proc glimmix;
model y= X1/solution;
run;
```

Part of the results is shown in Table 1.2. The analysis of variance (item a) indicates that drug concentration has a significant effect on average mathematical performance ($P = 0.0019$). The estimates of the regression model parameters (item b) are β_0 and β_1 , and the mean squared error (MSE scale) is shown in Table 1.2(b) under “Parameter estimates.”

With these results, the linear predictor ($\hat{\eta}_i$) that predicts the average mathematical performance as a function of LSD concentration is as follows:

$$\hat{\eta}_i = 89.124 - 9.01 \times \text{Conc}_i$$

This means that we can predict the average mathematical performance of an individual for whom we need to know the LSD concentration (Conc_i) to be applied. From the estimated parameters, we can say that there is a negative relationship between LSD concentration and mathematical score. Figure 1.1 clearly shows that an increase in drug supply has a negative effect on the mathematical score of the youth. This fitted model explains 87.7% of the variability in the data (Fig. 1.1).

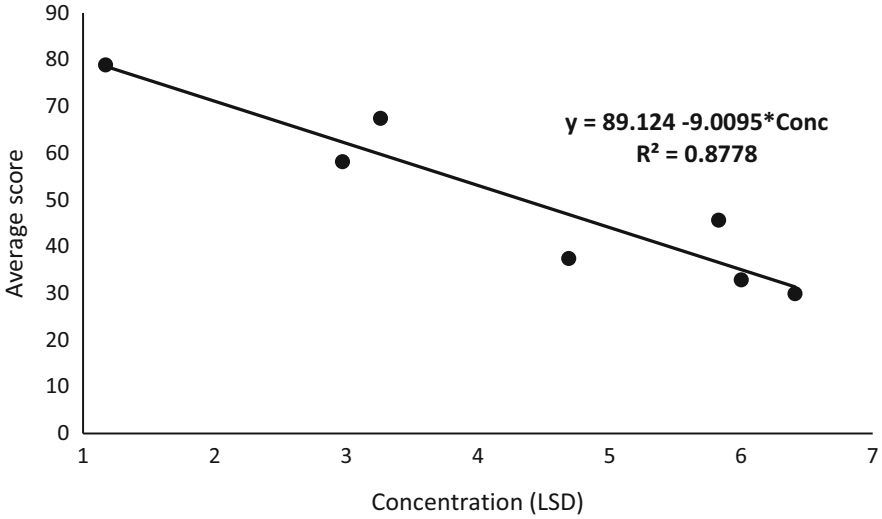


Fig. 1.1 Relationship between applied drug concentration and the mathematical score of the youth

Adjusted model of the relationship between the average score and LSD concentration.

1.2.2 Multiple Linear Regression

Suppose that a response variable y is linearly related to several independent variables X_1, X_2, \dots, X_p such that

$$y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_p X_{ip} + \varepsilon_i$$

for $i = 1, 2, \dots, n$. Here, ε_i are uncorrelated random errors ($i = 1, 2, \dots, n$) normally distributed with a zero mean and constant variance σ^2 , i.e., $\varepsilon_i \sim N(0, \sigma^2)$. If the explanatory variables are fixed constants, then the above model belongs to a general linear model of the form $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, as can be seen below:

$$\mathbf{y}_{n \times 1} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad \mathbf{X}_{n \times (p+1)} = \begin{pmatrix} 1 & X_{11} & X_{12} \cdots & X_{1p} \\ 1 & X_{21} & X_{22} \cdots & X_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & X_{n1} & X_{n2} \cdots & X_{np} \end{pmatrix}, \quad \boldsymbol{\beta}_{p+1} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix},$$

$$\boldsymbol{\varepsilon}_{n \times 1} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

Table 1.3 Body weight (kilograms) and its relationship with circumference (centimeters) and heart length (centimeters) of seven young bulls

Bull	1	2	3	4	5	6	7
Weight (kilograms)	480	450	480	500	520	510	500
Circumference (centimeters)	175	177	178	175	186	183	185
Length (centimeters)	128	122	124	131	131	130	124

A regression analysis can be used to assess the relationship between explanatory variables and the response variable. It is also a useful tool for predicting future observations or simply describing the structure of the data.

Example Let us to fit a regression model of the relationship between body weight and heart girth and length of the hearts of seven young bulls from the data shown in Table 1.3.

The components of this multiple regression model are as follows:

$$\text{Distribution: } y_i \sim N(\eta_i, \sigma^2)$$

$$\text{Linear predictor: } \eta_i = \beta_0 + X_{i1}\beta_1 + X_{i2}\beta_2$$

$$\text{Link function: } \mu_i = \eta_i(\text{identity})$$

The syntax for performing a multiple regression using the GLIMMIX procedure in SAS, assuming that there is no interaction between bull heart girth (X_1) and length (X_2), is shown below:

```
proc glimmix;
model y = X1 X2/solution cl;
run;
```

Based on the regression model specifications, the option “solution cl” prompts GLIMMIX to provide the value of the estimated parameters and their respective confidence intervals. Other useful options available are “htype = 1, 2, and 3,” which refer to the sum of squares of types I, II, and III. The type III fixed effects tests in (a) of Table 1.4 indicate that there is a linear relationship between heart length (size) and weight in young bulls. The estimated parameters with their respective confidence intervals $(\widehat{\beta}_0, \widehat{\beta}_1, \widehat{\beta}_2)$ as well as the MSE (scale) of the fitted regression model are listed below in (b).

Note that in a linear model, the parameters are linearly entered, but the variables do not necessarily have to be linear. For example, consider the following two examples:

$$y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 \log(X_{i2}) + \cdots + \beta_k X_{ik} + \epsilon_i$$

Table 1.4 Results of the multiple regression analysis

(a) Type III tests of fixed effects								
Effect	Num DF	Den DF	F-value			Pr > F		
X1	1		4.42			0.1034		
X2	1		9.51			0.0368		
(b) Parameter estimates								
Effect	Estimate	Standard error	DF	t-value	Pr > t	α	Lower	Upper
Intercept	-495.01	225.87		-2.19	0.0935	0.05	-1122.13	132.10
X1	2.2573	1.0739		2.10	0.1034	0.05	-0.7243	5.2388
X2	4.5808	1.4855		3.08	0.0368	0.05	0.4564	8.7053
Scale	139.51	98.6518

$$y_i = \beta_0 + X_{i1}^{\beta_1} + \beta_2 X_{i2} + \cdots + \beta_k \exp(X_{ik}) + \epsilon_i$$

The first example is a linear model, whereas the second one is not, since its derivatives do not depend on the beta coefficients, with the exception of the term $X_{i1}^{\beta_1}$ whose derivative is equal to $X_{i1}^{\beta_1} \log(X_{i1})$. This clearly shows that the second example is a nonlinear model because the derivative of the predictor depends on β_1 .

1.3 Analysis of Variance Models

1.3.1 One-Way Analysis of Variance

Consider an experiment in which you want to test t treatments ($t > 2$), to the level of the i th treatment with n_i experimental units that are selected and randomly assigned to the i th treatment. The model describing this experiment is as follows:

$$y_{ij} = \mu + \tau_i + \epsilon_{ij}$$

for $i = 1, 2, \dots, t$ and $j = 1, 2, \dots, n_i$. Here, ϵ_{ij} are the uncorrelated random errors with normal distribution with a zero mean and a variance constant σ^2 ($\epsilon_{ij} \sim N(0, \sigma^2)$). If the treatment effects are considered as fixed constants (drawn from a finite number), then this model is a special case of the general linear model (1), with the total number of experimental units $n = \sum_i^t n_i$.

Table 1.5 Biomass production of the three types of bacteria

Bacteria A	Bacteria B	Bacteria C
12	20	40
15	19	35
9	23	42

Table 1.6 Analysis of variance

Sources of variation	Degrees of freedom
Bacteria type	$t - 1 = 3 - 1 = 2$
Error	$t(r - 1) = 3 \times 2 = 6$
Total	$tr - 1 = 3 \times 3 - 1 = 8$

In matrix terms, the information under this design of experiment is equal to:

$$\mathbf{y}_{n \times 1} = \begin{pmatrix} y_{11} \\ y_{12} \\ \vdots \\ y_{tn_t} \end{pmatrix}, \quad \mathbf{X}_{n \times (t+1)} = \begin{pmatrix} \mathbf{1}_{n_1} & \mathbf{1}_{n_1} & \mathbf{0}_{n_1} & \cdots & \mathbf{0}_{n_1} \\ \mathbf{1}_{n_2} & \mathbf{0}_{n_2} & \mathbf{1}_{n_2} & \cdots & \mathbf{0}_{n_2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{1}_{n_t} & \mathbf{0}_{n_t} & \mathbf{0}_{n_t} & \cdots & \mathbf{1}_{n_t} \end{pmatrix}, \quad \boldsymbol{\beta}_{t+1 \times 1} = \begin{pmatrix} \mu \\ \tau_1 \\ \tau_2 \\ \vdots \\ \tau_t \end{pmatrix},$$

$$\boldsymbol{\epsilon}_{n \times 1} = \begin{pmatrix} \epsilon_{11} \\ \epsilon_{12} \\ \epsilon_{13} \\ \vdots \\ \epsilon_{tn_t} \end{pmatrix}$$

where $\mathbf{1}_{n_i}$ is the vector of ones of order n_i and $\mathbf{0}_{n_i}$ is the vector of zeros of order n_i . Note that the matrix $\mathbf{X}_{n \times (t+1)}$ is not of a full column rank because its first column can be obtained as a linear combination of its remaining columns.

Example Assume that measurements of the biomass produced by three different types of bacteria are collected in three separate Petri dishes (replicates) in a glucose broth culture medium for each bacterium (Table 1.5).

The sources of variation and degrees of freedom (DFs) for this experiment are shown in Table 1.6.

The components for this one-way model, assuming that each of the response variable y_{ij} is normally distributed, are as follows:

Distribution: $y_{ij} \sim N(\mu_{ij}, \sigma^2)$

Linear predictor: $\eta_i = \alpha + \tau_i$

Link function: $\mu_i = \eta_i$ (identity)

where y_{ij} is the response observed at the j th repetition in the i th bacterium, η_i is the linear predictor, α is the intercept (the grand mean), and τ_i is the fixed effect due to the type of bacterium.

Table 1.7 Results of the one-way analysis of variance

(a) Fit statistics				
-2 Res log likelihood				33.36
AIC (Akaike information criterion) (smaller is better)				41.36
AICC (Corrected Akaike information criterion) (smaller is better)				81.36
BIC (Bayesian information criterion) (smaller is better)				40.52
CAIC (Consistent Akaike's information criterion) (smaller is better)				44.52
HQIC (Hannan and Quinn information criterion) (smaller is better)				38.02
Pearson's chi-square				52.67
Pearson's chi-square / DF				8.78
(b) Type III tests of fixed effects				
Effect	Num DF	Den DF	F-value	Pr > F
Bacteria			64.95	<0.0001

The SAS syntax for a one-way analysis of variance (ANOVA) is as follows:

```
proc glimmix data=biomass;
class bacteria;
model y = bacteria;
lsmeans bacteria/lines;
run;
```

Similar to “proc glm” or “proc mixed,” the “class” command allows to define the type of class variables (categorical or nominal) to be included in the model; in this case, for the class variable “bacteria,” the “model” command allows to declare (list) the response variable “y” and all the class or continuous variables that enter the model, whereas the “lsmeans” command asks GLIMMIX to estimate the means of the treatments and the “lines” option allows to make a comparison of means. Part of the results is presented below.

By default, “proc GLIMMIX” provides the fit statistics (information criteria), which are extremely useful for comparing or choosing a model that explains the largest possible proportion of variation present in a dataset, i.e., the best-fit model (part (a) of Table 1.7). The statistic “-2 res log likelihood” is most useful when comparing nested models, and the rest of the statistics is useful for comparing models that are not necessarily nested. The mean squared error (MSE) in GLIMMIX is given as the statistic “Pearson's chi - square/DF.” In this analysis, this value is 8.78. ($\hat{\sigma}^2 = \text{MSE} = 8.78$). In part (b), the analysis of variance indicates that at least one type of bacterium produces a different biomass ($P < 0.0001$). That is, the null hypothesis is rejected ($H_0 : \tau_A = \tau_B = \tau_C$) at a significance level of 5%.

The estimated least squares (LS) means obtained with “lsmeans” are tabulated under the “Estimate” column with their standard errors in the “Standard error” column of Table 1.8. These estimated means were obtained (by default) with Fisher's LSD (least significant difference).

Table 1.8 Means and estimated standard errors of the one-way model

Least squares means of bacteria					
Bacteria	Estimate	Standard error	DF	t-value	Pr > t
A	12.0000	1.7105		7.02	0.0004
B	20.6667	1.7105		12.08	<0.0001
C	39.0000	1.7105		22.80	<0.0001

Table 1.9 Comparison of the means (LSD) in the one-way model

T grouping of the least squares means of bacteria ($\alpha = 0.05$)		
LS means with the same letter are not significantly different		
Bacteria	Estimate	
C	39.0000	A
B	20.6667	B
A	12.0000	C

Finally, Table 1.9 presents a comparison of the means obtained with “lines” and indicates that bacteria type C has a better fermentative conversion of glucose to lactic acid compared to bacteria types B and A. Equal letters per column indicate that they are statistically equal.

1.3.2 Two-Way Nested Analysis of Variance

Let us consider an experiment with two factors, A and B, in which each level of B is nested within a level of factor A, that is, each level of factor B appears within a level of factor A. Then, the model that describes this experiment is as follows:

$$y_{ijk} = \mu + \alpha_i + \beta_{j(i)} + \epsilon_{ijk}$$

for $i = 1, 2, \dots, a; j = 1, 2, \dots, b_i; \text{ and } k = 1, 2, \dots, n_{ij}$. In this model, μ is the overall mean, α_i represents the effect due to the i th level of factor A, and $\beta_{j(i)}$ represents the effect of the j th level of factor B nested within the i th level of factor A. Assuming that all factors are fixed, and that the errors ϵ_{ijk} are normally distributed, that is $\epsilon_{ijk} \sim N(0, \sigma^2)$, this model is the general linear model of the form $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$. For example, suppose that you have $a = 3$ levels of factor A, $b = 2$ levels of factor B, and $n_{ij} = 2$, then the vectors and matrices have the following form:

$$\mathbf{y} = \begin{pmatrix} y_{111} \\ y_{112} \\ y_{121} \\ y_{122} \\ y_{211} \\ y_{212} \\ y_{221} \\ y_{222} \\ y_{311} \\ y_{312} \\ y_{321} \\ y_{322} \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} 1 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix},$$

$$\boldsymbol{\beta} = \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \beta_{11} \\ \beta_{12} \\ \beta_{21} \\ \beta_{22} \\ \beta_{31} \\ \beta_{32} \end{pmatrix}, \quad \boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_{111} \\ \epsilon_{112} \\ \epsilon_{121} \\ \epsilon_{122} \\ \epsilon_{211} \\ \epsilon_{212} \\ \epsilon_{221} \\ \epsilon_{222} \\ \epsilon_{311} \\ \epsilon_{312} \\ \epsilon_{321} \\ \epsilon_{322} \end{pmatrix}.$$

Example Suppose that a researcher was studying the assimilation of fluorescently labeled proteins in rat kidneys and wanted to know whether his two technicians, technician A and technician B, were performing the procedure consistently. Technician A randomly chose three rats, and technician B randomly chose three other rats, and each technician measured the protein assimilation in each rat. Since rats are expensive and measurements are cheap, both technicians measured protein assimilation at various random locations in the kidneys of each rat (Table 1.10).

When performing a nested ANOVA, we are often interested in testing the null hypothesis ($H_0 : \tau_A = \tau_B$). As in this example, we do not wish to test whether the subgroups (rats within technicians) are significantly different, since the goal is to prove that both technicians are performing their jobs adequately. The sources of variation and degrees of freedom are shown in Table 1.11.

The components of this two-way model, assuming that the response variable y_{ij} is normally distributed, are as follows:

Table 1.10 Levels of protein assimilation in the rat kidneys measured by both technicians

Technician A			Technician B		
Rat1	Rat2	Rat3	Rat4	Rat5	Rat6
1.119	1.045	0.9873	1.3883	1.3952	1.2574
1.2996	1.1418	0.9873	1.104	0.9714	1.0295
1.5407	1.2569	0.8714	1.1581	1.3972	1.1941
1.5084	0.6191	0.9452	1.319	1.5369	1.0759
1.6181	1.4823	1.1186	1.1803	1.3727	1.3249
1.5962	0.8991	1.2909	0.8738	1.2909	0.9494
1.2617	0.8365	1.1502	1.387	1.1874	1.1041
1.2288	1.2898	1.1635	1.301	1.1374	1.1575
1.3471	1.1821	1.151	1.3925	1.0647	1.294
1.0206	0.9177	0.9367	1.0832	0.9486	1.4543

Table 1.11 Sources of variation and degrees of freedom of the two-way nested design

Sources of variation	Degrees of freedom
Technician	$a - 1 = 2 - 1 = 1$
Rat (technical)	$a(b - 1) = 2(3 - 1) = 4$
Error	$ab(r - 1) = 2 \times 3(10 - 1) = 54$
Total	$abr - 1 = 2 \times 3 \times 10 - 1 = 59$

$$\text{Distribution: } y_{ij} \sim N(\mu_{ij}, \sigma^2)$$

$$\text{Linear predictor: } \eta_{ij} = \alpha + \tau_i + \beta(\tau)_{j(i)}$$

$$\text{Link function: } \mu_i = \eta_i \text{ (identity)}$$

where y_{ij} is the level of assimilation of the fluorescent protein obtained from rat j by technician i , α is the intercept, τ_i is the fixed effect due to the technician, and $\beta(\tau)_{j(i)}$ is the nested effect of rat j within technician i .

The SAS commands for the main effects of factor A and factor B nested within A are as follows:

```
proc glimmix data=rata nobound;
class technician rat rep;
model protein=technical rat(technical);
lsmeans technician rat(technician)/lines;
run;
```

Part of the results is shown in Table 1.12. The results indicate that there is minimum variability of the technicians since the value of the mean squared error (Pearson's chi-square/DF) is 0.04 (part (a)). This means that the variance between group means is smaller than would be expected. The analysis of variance in part (b) indicates that there is no difference in the measurement of fluorescent proteins in the rats between technicians ($P = 0.3065$). Since there is variation between rats in the

Table 1.12 Fit statistics of the two-way nested design

(a) Fit statistics				
-2 Res log likelihood				-12.39
AIC (smaller is better)				1.61
AICC (smaller is better)				4.04
BIC (smaller is better)				15.53
CAIC (smaller is better)				22.53
HQIC (smaller is better)				6.98
Pearson's chi-square				1.95
Pearson's chi-square / DF				0.04
(b) Type III tests of fixed effects				
Effect	Num DF	Den DF	F-value	Pr > F
Technician	1		1.07	0.3065
Rat (technical)			3.98	0.0067

Table 1.13 Comparison of the means (LSD) in the nested model

(a) Technical least squares means					
Technician	Estimate	Standard error	DF	t-value	Pr > t
A	1.2110	0.03466		34.94	<0.0001
B	1.1604	0.03466		33.48	<0.0001
(b) T grouping of the technical least squares means ($\alpha = 0.05$)					
LS means with the same letter are not significantly different					
Technician	Estimate				
A	1.2110			A	
				A	
B	1.1604			A	

average protein uptake, it is to be expected that between rats within technicians, there are mean differences in the protein uptake ($P = 0.0067$).

In Table 1.13 part (a), the values of the least squares means tabulated under the "Estimate" column are shown with their respective "Standard errors." It can be seen that rats under technician A have statistically the same mean protein uptake as do rats under technician B (part (b)).

Comparison of means for rat subgroups under both technicians showed similar means for rats under technician A but different means for rats under technician B (part (a) and (b), Table 1.14).

Table 1.14 Comparison of the means (LSD) of the subgroups nested within technicians

(a) Least squares means of rats (technical)

Technician	Rat	Estimate	Standard error	DF	t-value	Pr > t
A	5	1.2187	0.06003		20.30	<0.0001
A		1.2302	0.06003		20.49	<0.0001
A		1.1841	0.06003		19.72	<0.0001
B	1	1.3540	0.06003		22.56	<0.0001
B		1.0670	0.06003		17.77	<0.0001
B		1.0602	0.06003		17.66	<0.0001

(b) T grouping of the least squares means ($\alpha = 0.05$) of rats (technical)

LS means with the same letter are not significantly different

Technician	Rat	Estimate		
B	1	1.3540		A
A		1.2302	B	A
A	5	1.2187	B	A
A		1.1841	B	A
B		1.0670	B	
B		1.0602	B	

1.3.3 Two-Way Analysis of Variance with Interaction

This experiment is used when one wishes to test two factors A and B, with a levels of factor A and b levels of factor B. In this experiment, both factors are crossed, this means that each level of A occurs in combination with each level of factor B. The model with interaction is given by:

$$y_{ijk} = \mu + \alpha_i + \beta_{ij} + \gamma_{ij} + \epsilon_{ijk}$$

for $i = 1, 2, \dots, a; j = 1, 2, \dots, b; k = 1, 2, \dots, n_{ij};$ and $\epsilon_{ijk} \sim N(0, \sigma^2)$. If all the parameters of the model are fixed, then this model can be expressed as $y = X\beta + \epsilon$. For this model with $a = 3, b = 2,$ and $n_{ij} = 3,$ the matrix expression has the form:

$$\mathbf{y} = \begin{pmatrix} y_{111} \\ y_{112} \\ y_{113} \\ y_{121} \\ y_{122} \\ y_{123} \\ y_{211} \\ y_{212} \\ y_{213} \\ y_{221} \\ y_{222} \\ y_{223} \\ y_{311} \\ y_{312} \\ y_{313} \\ y_{321} \\ y_{322} \\ y_{323} \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} 1 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix},$$

$$\boldsymbol{\beta} = \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \beta_1 \\ \beta_2 \\ \gamma_{11} \\ \gamma_{12} \\ \gamma_{21} \\ \gamma_{22} \\ \gamma_{31} \\ \gamma_{32} \end{pmatrix}, \quad \boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_{111} \\ \epsilon_{112} \\ \epsilon_{113} \\ \epsilon_{121} \\ \epsilon_{122} \\ \epsilon_{123} \\ \epsilon_{211} \\ \epsilon_{212} \\ \epsilon_{213} \\ \epsilon_{221} \\ \epsilon_{222} \\ \epsilon_{223} \\ \epsilon_{311} \\ \epsilon_{312} \\ \epsilon_{313} \\ \epsilon_{321} \\ \epsilon_{322} \\ \epsilon_{323} \end{pmatrix}$$

Example This experiment consisted of developing an in vitro efficacy test for self-tanning formulations. Two brands, 1 = erythrulose, 2 = dihydroxyacetone (factor A), and three formulations, 1 = solution, 2 = gel, and 3 = cream (factor B), were tested with four replicates for each condition according to Jermann et al. (2001). Total color change was measured for each of the combination conditions. The dataset is shown in Table 1.15.

Table 1.15 Color change (Y) in each of the brands and formulations

Brand	Formulation	Y	Brand	Formulation	Y
1	1	16.79		1	32.85
1	1	12.68		1	38.08
1	1	12.47		1	30.25
1	1	11.67		1	28.41
1		10.23			25.06
1		10.29			21.66
1		8.97			19.86
1		8.51			18.62
1		9.43			25.89
1		9.45			22.96
1		8.86			24.55
1		8.66			24.59

Table 1.16 Analysis of variance of the two-way model with interaction

Sources of variation	Degrees of freedom
Brand	$a - 1 = 2 - 1 = 1$
Formulation	$a - 1 = 3 - 1 = 2$
Brand \times formulation	$(a - 1)(b - 1) = 1 \times 2 = 2$
Error	$ab(r - 1) = 2 \times 3 \times 3 = 18$
Total	$abr - 1 = 2 \times 3 \times 4 - 1 = 23$

For this two-way model, assuming that the response variable y_{ijk} has a normal distribution, the components are as follows:

$$\text{Distribution: } y_{ijk} \sim N(\mu_{ijk}, \sigma^2)$$

$$\text{Linear predictor: } \eta_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij}$$

$$\text{Link function: } \mu_{ij} = \eta_{ij} \text{ (identity)}$$

where y_{ijk} is the color change observed at the k th repetition at the i th level of factor A and at the j th level of factor B, μ is the intercept (the overall mean), α_i is the fixed effect due to the level of factor A (mark), β_j represents the fixed effect of the level of factor B (type of formulation), and γ_{ij} is the fixed effect due to the interaction between the brand and formulation. Table 1.16 shows the sources of variation and degrees of freedom.

The following code in GLIMMIX in SAS allows us to estimate the main effects and the interaction:

```
proc glimmix;
class brand formulation;
model y = brand|formulacion;
lsmeans brand|formulacion/lines;
run;
```

Table 1.17 Results of the analysis of variance of the two-way model with interaction

(a) Fit statistics				
–2 Res log likelihood				90.20
AIC (smaller is better)				104.20
AICC (smaller is better)				115.40
BIC (smaller is better)				110.43
CAIC (smaller is better)				117.43
HQIC (smaller is better)				105.06
Pearson’s chi-square				99.61
Pearson’s chi-square / DF				5.53
(b) Type III tests of fixed effects				
Effect	Num DF	Den DF	<i>F</i> -value	Pr > <i>F</i>
Brand	1		257.04	<0.0001
Formulation			22.99	<0.0001
Brand × formulation			4.68	0.0231

Table 1.18 Means and standard errors of the tanning brand

Least squares means of the brand					
Brand	Estimate	Standard error	DF	<i>t</i> -value	Pr > <i>t</i>
1	10.6675	0.6791		15.71	<0.0001
	26.0650	0.6791		38.38	<0.0001
T grouping of the least squares means ($\alpha = 0.05$)					
LS means with the same letter are not significantly different					
Brand	Estimate				
	26.0650		A		
1	10.6675		B		

Part of the results is shown below. Of all the fit statistics in (a) of Table 1.17, the value that we are interested in highlighting in this analysis is “Pearson’s chi – square/DF,” which corresponds to the mean squared error (MSE), even though we are evaluating different possible models for this given dataset. The value of the MSE is 5.53. The type III fixed effects tests, in part (b) of Table 1.17, indicate that the type of brand ($P < 0.0001$), formulation ($P < 0.0001$), and the interaction between both factors ($P = 0.0231$) all have a significant effect on the change of self-tanning color.

The least mean squares obtained with “lsmeans” are shown in the Table 1.18 for the levels of tanning brand factor in Table 1.19 for the levels of tanning brand formulation and in Table 1.20 for the interaction of both factors. The “lines” option allows us to make a comparison of means using the LSD method.

The least squares means for the tanning brand factor are given in Table 1.18.

The least squares means for the type of tanning brand formulation are given in Table 1.19.

Table 1.19 Means and standard errors of the tanning brand formulation

Least squares means for the tanning brand formulation					
Formulation	Estimate	Standard error	DF	t-value	Pr > t
1	22.9000	0.8317		27.53	<0.0001
	15.4000	0.8317		18.52	<0.0001
	16.7988	0.8317		20.20	<0.0001
T grouping of the least squares means ($\alpha = 0.05$) of the tanning brand formulation					
LS means with the same letter are not significantly different					
Formulation	Estimate				
1	22.9000		A		
	16.7988		B		
			B		
	15.4000		B		

Table 1.20 Comparison of the means of the interaction of both factors

T grouping of the least squares means ($\alpha = 0.05$) of the marca*formulation			
LS means with the same letter are not significantly different			
Brand	Formulation	Estimate	
	1	32.3975	A
		24.4975	B
		21.3000	B
1	1	13.4025	C
1		9.5000	D
1		9.1000	D

The hypothesis test for the interaction should be tested first, and only if the interaction effect is not significant, should the main effects be tested. If the interaction is significant, then tests for the main effects are meaningless. The interaction analysis shows that brand 2 (dihydroxyacetone), in all three formulations, shows a greater change compared to brand 1 (erythrose).

Now, considering the previous model without interaction ($\gamma_{11} = \gamma_{12} = \dots = \gamma_{32} = 0$) where factor A has a levels and factor B has b levels, the model without interaction is given by:

$$y_{ijk} = \mu + \alpha_i + \beta_j + \epsilon_{ijk}$$

for $i = 1, 2, \dots, a; j = 1, 2, \dots, b; k = 1, 2, \dots, n_{ij}$; and $\epsilon_{ijk} \sim N(0, \sigma^2)$. The model without interaction with $a = 3, b = 2$, and $n_{ij} = 3$ reduces to:

$$\mathbf{y} = \begin{pmatrix} y_{111} \\ y_{112} \\ y_{113} \\ y_{121} \\ y_{122} \\ y_{123} \\ y_{211} \\ y_{212} \\ y_{213} \\ y_{221} \\ y_{222} \\ y_{223} \\ y_{311} \\ y_{312} \\ y_{313} \\ y_{321} \\ y_{322} \\ y_{323} \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} 1 & 1 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 & 1 \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \beta_1 \\ \beta_2 \end{pmatrix}, \quad \boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_{111} \\ \epsilon_{112} \\ \vdots \\ \epsilon_{322} \\ \epsilon_{323} \end{pmatrix}$$

Note that the design matrix for the model without interaction is the same as that for the model with interaction, except that the last six columns are removed.

Let us assume that the interaction effect is not significant. The following SAS code estimates the main effects of both factors. Running the program and analysis is left as practice for the readers.

```

proc glimmix;
class brand formulation;
model y = formula brand;
lsmeans brand formulation/lines;
run;

```

1.4 Analysis of Covariance (ANCOVA)

Consider an experiment to compare $t \geq 2$ treatments after adjusting for the effects of a covariate x . The model for an analysis of covariance is given by:

$$y_{ij} = \mu + \tau_i + \beta_i x_{ij} + \epsilon_{ij}$$

for $i = 1, 2, \dots, t$ and $j = 1, 2, \dots, n_i$. Here, ϵ_{ij} are the independent normally distributed random errors with a zero mean and a variance constant $\sigma^2 > 0$. In this model, μ is the overall mean, τ_i is the fixed effect of the i th treatment (ignoring the covariates x 's), β_i denotes the slope of the line that relates the response variable y to x for the i th treatment, and x_{ij} are fixed covariates. Assuming $t = 3$, $n_1 = n_2 = n_3 = 3$, we have:

$$\mathbf{y} = \begin{pmatrix} y_{11} \\ y_{12} \\ y_{13} \\ y_{21} \\ y_{22} \\ y_{23} \\ y_{31} \\ y_{32} \\ y_{33} \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} 1 & 1 & 0 & 0 & x_{11} & 0 & 0 \\ 1 & 1 & 0 & 0 & x_{12} & 0 & 0 \\ 1 & 1 & 0 & 0 & x_{13} & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & x_{21} & 0 \\ 1 & 0 & 1 & 0 & 0 & x_{22} & 0 \\ 1 & 0 & 1 & 0 & 0 & x_{23} & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & x_{31} \\ 1 & 0 & 0 & 1 & 0 & 0 & x_{32} \\ 1 & 0 & 0 & 1 & 0 & 0 & x_{33} \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix},$$

$$\boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_{11} \\ \epsilon_{12} \\ \epsilon_{13} \\ \epsilon_{21} \\ \epsilon_{22} \\ \epsilon_{23} \\ \epsilon_{31} \\ \epsilon_{32} \\ \epsilon_{33} \end{pmatrix}$$

The analysis of covariance (ANCOVA), as can be seen, obeys a general linear model of the form $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$.

For example consider a hypothetical study of flower production in two subspecies of plants. The number of flowers per plant may vary between the subspecies, but, within each subspecies, flower production may also vary with the size of each plant, and this relationship may be positive or negative. A positive relationship might arise if plants with more resources (sunlight, water, nutrients) could invest more energy in both growth and flower production. A negative relationship could arise if there was a trade-off between the energy invested in growth and the energy invested in flower production. In this study, subspecies is a categorical variable and plant size is a continuous variable (the covariate). Measuring plant size and flower production in the two subspecies allows the investigation of three different questions:

Is flower production influenced by subspecies?

Is flower production influenced by plant size?

Is the effect of flower production on plant size influenced by subspecies?

Example 1. The central question in plant reproductive ecology is how hermaphroditic plant species allocate resources to male and female structures. A study conducted to address this question counted the number of stamens (male structures that produce pollen) and ovules (female structures that when fertilized by a pollen grain will become seeds) in the flowers of “prairie larkspur” plants in two populations in southeastern Minnesota. The total number of flowers produced by each plant was also determined to assess whether plant size affected ovule production per flower. The dataset for this example can be found in the [Appendix](#) (Data: Larkspur plants).

An ANCOVA is appropriate for this study to test the following three null hypotheses from these data:

- (a) There is no difference in the average number of ovules per flower between the two populations (the main effect).
- (b) There is no effect of plant size on the average number of ovules per flower (the covariate effect).
- (c) The effect of plant size on the mean number of ovules per flower did not differ between the study sites (the interaction effect).

The components of the ANCOVA model, assuming that the response variable y_{ijk} is normally distributed, are as follows:

$$\text{Distribution: } y_{ij} \sim N(\mu_{ijk}, \sigma^2)$$

$$\text{Linear predictor: } \eta_{ij} = \mu + \tau_i + \text{planta}(\tau)_{j(i)} + \beta_i(X_{ij} - \bar{X}_{..})$$

$$\text{Link function : } \mu_{ij} = \eta_{ij} \text{ (identity)}$$

where y_{ij} is the number of ovules observed in the j th plant of the i th population, μ is the overall mean, τ_i is the fixed effect due to the population i , $\text{planta}(\tau)_{j(i)}$ is the random effect due to the plant j in the population i , β_i is the slope of the population i , $\bar{X}_{..}$ is the overall mean of the size of all plants, and X_{ij} is the plant size i in the population j . The ANCOVA results (sources of variation and degrees of freedom) are shown in Table 1.21.

The basic syntax in GLIMMIX for analysis of covariance with different slopes is as follows:

```
proc glimmix;
class poblacion plant;
model ovules = population xbar population*xbar/ddfm=satterthwaite;
random plant (population);
lsmeans population/lines;
run;
```

Table 1.21 Analysis of covariance

Sources of variation	Degrees of freedom
Population	$t - 1 = 2 - 1 = 1$
β (depending on the size of the plant)	1
Population β	1
Error	$\left(\sum_{i=1}^t r_i - 1\right) - t - 1 = 75$
Total	$\sum_{i=1}^t r_i - 1 = 79 - 1 = 78$

Table 1.22 Results of the analysis of covariance for the two populations of larkspur plants

(a) Covariance parameter estimates					
Cov Parm	Estimate	Standard error			
Plant (population)	12.7951	2.2416			
Residual	0.9321	.			
(b) Type III tests of fixed effects					
Effect	Num DF	Den DF	F-value	Pr > F	
Population	1		7.32	0.0084	
Center	1		16.39	0.0001	
Center x population	1		7.81	0.0066	
(c) Least squares means of population					
Population	Estimate	Standard error	DF	t-value	Pr > t
Cedar.cr	20.3538	0.6062		33.58	<0.0001
St. Croix	22.7596	0.6502		35.00	<0.0001
(d) T grouping of the least squares means ($\alpha = 0.05$) of population					
LS means with the same letter are not significantly different					
Population	Estimate				
St. Croix	22.7596	A			
Cedar.cr	20.3538	B			

In the above syntax, the “class” command lists all classes or categorical variables, except the covariate (continuous variable), which – in this case – is a variable centered by the average of the size of all plants ($xbar = (X_{ij} - \bar{X}_{..})$). The options “ddfm” and “lines” invoke proc GLIMMIX to do a degree-of-freedom correction using the Satterthwaite method and a comparison of the means using the LSD method. Part of the results is shown in Table 1.22.

The estimates of the variance components (part (a)) due to plant and within-treatment variability are $\hat{\sigma}_{\text{planta(poblacion)}}^2 = 12.795$ and $\hat{\sigma}^2 = MSE = 0.9321$, respectively. The analysis of variance in (b) showed that there is a significant effect between the two populations ($P = 0.0084$), plant size ($P = 0.0001$) and plant size is influenced by subspecies (interaction) on the average number of ovules ($P = 0.0066$) per flower. The estimated means and their respective standard errors of the average number of ovules for both populations are tabulated in the “Estimate” column in part (c), as well as the comparison of the means in part (d).

If in the previous model we assume that the slopes were equal ($\beta_1 = \beta_2$), then the ANCOVA reduces to:

$$y_{ij} = \mu + \tau_i + \beta(X_{ij} - \bar{X}_{..}) + \epsilon_{ij}$$

The ANCOVA model, with $t = 3$, $n_1 = n_2 = n_3 = 3$ for this case (equality of slopes) reduces to:

$$\mathbf{y} = \begin{pmatrix} y_{11} \\ y_{12} \\ y_{13} \\ y_{21} \\ y_{22} \\ y_{23} \\ y_{31} \\ y_{32} \\ y_{33} \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} 1 & 1 & 0 & 0 & x_{11} \\ 1 & 1 & 0 & 0 & x_{12} \\ 1 & 1 & 0 & 0 & x_{13} \\ 1 & 0 & 1 & 0 & x_{21} \\ 1 & 0 & 1 & 0 & x_{22} \\ 1 & 0 & 1 & 0 & x_{23} \\ 1 & 0 & 0 & 1 & x_{31} \\ 1 & 0 & 0 & 1 & x_{32} \\ 1 & 0 & 0 & 1 & x_{33} \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \beta \end{pmatrix}, \quad \boldsymbol{\varepsilon} = \begin{pmatrix} \epsilon_{11} \\ \epsilon_{12} \\ \epsilon_{13} \\ \epsilon_{21} \\ \epsilon_{22} \\ \epsilon_{23} \\ \epsilon_{31} \\ \epsilon_{32} \\ \epsilon_{33} \end{pmatrix}.$$

The basic syntax using GLIMMIX for an analysis of covariance with equal slopes is as follows:

```

proc glimmix;
class poblacion plant;
model ovules = population xbar/ddfm=satterthwaite;
random plant (population);
lsmeans population/lines;
run;

```

So far, we have exemplified the general linear model of the form $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$. In the following, some characteristics of a linear mixed model (LMM) will be described.

1.5 Mixed Models

1.5.1 Introduction

Linear mixed models (LMMs) are appropriate for analyzing continuous response variables in which the residuals are normally distributed. These types of models are well suited for studies of grouped datasets such as (1) students in classrooms, animals in herds, people grouped by municipality or geographic region, or randomized block experimental designs such as batches of raw materials for an industrial process and (2) longitudinal or repeated measures studies, in which subjects are measured repeatedly over time or under different conditions. These designs occur in a wide variety of settings: biology, agriculture, industry, and socioeconomic sciences. LMMs provide researchers with powerful and flexible analytical tools for these types of data.

The name linear mixed models comes from the fact that these models are linear in the parameters and that the covariates, or independent variables, may involve a combination of fixed and random effects. “Fixed effects” can be associated with continuous covariates, such as weight in kilograms of an animal, maize yield in tons per hectare, and reference test score or socioeconomic status, which will carry a continuous range of values, or with factors, such as gender, variety, or group

treatment, which are categorical. Fixed effects are unknown constant parameters associated with continuous covariates or levels of the categorical factors in an LMM. The estimation of these parameters in LMMs is generally of intrinsic interest because they indicate the relationship of the covariates with the continuous response variable.

When the levels of a factor are drawn from a large enough sample such that each particular level is not of interest (e.g., classrooms, regions, herds, or clinics that are randomly sampled from a population), the effects associated with the levels of those factors can be modeled as random effects in an LMM. “Random effects” are represented by random (unobserved) variables that we generally assume to have a particular distribution, with normal distribution being the most common.

Mixed models are extremely useful because they allow us to work on (address) two important aspects:

1. From a statistical point of view, biological data are often structured in a way that does not satisfy the assumption of independence of the dataset. Examples include the following:
 - (a) Multiple measurements of the same subject/organism
 - (b) Experiments organized into spatial blocks
 - (c) Observational data in which multiple investigations were conducted in different locations
 - (d) Synthesis of data from similar experiments that were performed by different researchers
2. From a biological perspective, the processes being measured can be affected by multiple sources of variation, often occurring at different spatial or temporal scales. We are interested in using statistical methods that can model multiple sources of stochasticity, at multiple scales, so that we can measure the relative magnitude of the different sources of variation and determine which predictors explain variation at different scales.

1.5.2 Mixed Models

The matrix notation for a mixed model is highly similar to that for a fixed effects (systematic) model. The main difference is that, instead of using only one design matrix to explain the entire model in its systematic part, the matrix notation for a mixed model uses at least two design matrices: a design matrix X to describe the fixed effects in the model and a design matrix Z to describe the random effects in the model. The fixed effects design matrix X is constructed in the same way as a general linear fixed effects model ($y = X\beta + \epsilon$). X has a dimension of $n \times (p + 1)$, where n is the number of observations in the dataset and $p + 1$ is the number of parameters of fixed effects in the model to be estimated. The design matrix for the random effects Z is constructed in the same way as the construction of the design matrix for fixed effects, but now for the random effects. The Z matrix has a dimension of $n \times q$, where q is the number of coefficients of random effects in the model.

In matrix notation, a linear mixed model can be represented as

$$\mathbf{y} = \overbrace{\mathbf{X}\boldsymbol{\beta}}^{\text{Sistematic}} + \overbrace{\mathbf{Z}\mathbf{b}}^{\text{random}} + \overbrace{\boldsymbol{\varepsilon}}^{\text{Experimental Error}} \quad (1.2)$$

$$\mathbf{b} \sim N(\mathbf{0}, \mathbf{G}) \text{ and } \boldsymbol{\varepsilon} \sim N(\mathbf{0}, \mathbf{R})$$

where \mathbf{y} is the vector of $n \times 1$ observations, $\boldsymbol{\beta}$ is the vector of $(p + 1) \times 1$ fixed effects, \mathbf{b} is the vector of random effects of $q \times 1$, $\boldsymbol{\varepsilon}$ is the vector of $n \times 1$ random error terms, \mathbf{X} is the design matrix of $n \times (p + 1)$ for fixed effects related to observations at $\boldsymbol{\beta}$, and \mathbf{Z} is the design matrix $n \times q$ for the random effects (\mathbf{b}) related to observations at \mathbf{b} .

Assuming that both \mathbf{b} and $\boldsymbol{\varepsilon}$ are uncorrelated random variables with a zero mean and variance–covariance matrices \mathbf{G} and \mathbf{R} , respectively, we have

$$E(\mathbf{b}) = \mathbf{0}, E(\boldsymbol{\varepsilon}) = \mathbf{0}$$

$$\text{Var}(\mathbf{b}) = \mathbf{G}, \text{Var}(\boldsymbol{\varepsilon}) = \mathbf{R}$$

$$\text{Cov}(\mathbf{b}, \boldsymbol{\varepsilon}) = \mathbf{0}$$

It is not difficult to verify that $\text{Var}(\mathbf{y}) = \text{Var}(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b} + \boldsymbol{\varepsilon})$ is

$$\text{Var}(\mathbf{y}) = \mathbf{Z}\mathbf{G}\mathbf{Z}' + \mathbf{R} = \mathbf{V}$$

Matrix \mathbf{V} is an important component when working with linear mixed models (LMMs) because it contains random sources of variation and also defines how such models differ from ordinary least squares estimation. If the model contains only random effects, such as a randomized complete block design (RCBD), then matrix \mathbf{G} is the first point of attention. On the other hand, for repeated measures or for spatial analysis, matrix \mathbf{R} is extremely important. Assuming that the random effects (blocks) have a normal distribution,

$$\mathbf{b} \sim N(\mathbf{0}, \mathbf{G}) \text{ and } \text{Var}(\boldsymbol{\varepsilon}) = \mathbf{R}$$

Then, the vector of observations \mathbf{y} will have a normal distribution, that is, $\mathbf{y} \sim N(\mathbf{X}\boldsymbol{\beta}, \mathbf{V})$. The same model can be written in the probability distribution form in two different but equivalent ways. The first is the marginal model

$$\mathbf{y} \sim N(E[\mathbf{y}] = \mathbf{X}\boldsymbol{\beta}, \mathbf{V} = \mathbf{Z}\mathbf{G}\mathbf{Z}' + \mathbf{R}) \quad (1.3)$$

In this marginal model, the mean is based only on fixed effects and the parameters describing the random effects appear (are contained) in the variance and covariance matrix \mathbf{V} (Littell et al. 2006). In general, a structure is imposed in \mathbf{b} in terms of $\text{Var}(\mathbf{b}) = \mathbf{G}$, and, therefore, marginally, the components of \mathbf{y} depend on the structure in $\mathbf{V} = \mathbf{Z}\mathbf{G}\mathbf{Z}' + \mathbf{R}$.

The second model is the conditional model

$$\mathbf{y} \mid \mathbf{b} \sim N(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b}, \mathbf{R}) \quad (1.4)$$

In this conditional model, \mathbf{b} is distributed as shown in Eq. (1.2) for this parameter. For LMMs, the two models are exactly the same; but if the response variable is modeled under a non-normal distribution, then the models are different (Stroup, 2012) and generalized linear mixed models are required.

The fixed effects estimator ($\boldsymbol{\beta}$) is useful to obtain the best linear unbiased estimators (commonly known as BLUEs), whereas the estimator \mathbf{b} is useful for computing the best linear unbiased predictors (commonly known as BLUPs) for the random effects \mathbf{b} . The estimation of the expected value of the marginal LMM (1.3) allows the estimation of the BLUEs and that of the conditional LMM (1.4), the BLUPs. The estimators for the BLUEs of $\boldsymbol{\beta}$ and the BLUPs of \mathbf{b} are as follows:

$$\begin{aligned} \hat{\boldsymbol{\beta}} &= (\mathbf{X}^T \mathbf{V}^{-1} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{V}^{-1} \mathbf{y} \\ \hat{\mathbf{b}} &= \mathbf{GZ}^T \mathbf{V}^{-1} (\mathbf{y} - \hat{\boldsymbol{\beta}}) \end{aligned}$$

This solution is efficient when working with small datasets because, in the context of big data, it is computationally highly demanding since the inverse of matrix \mathbf{V} has to be estimated. For this reason, it is normally used to obtain the solution of the BLUEs of $\boldsymbol{\beta}$ and the BLUPs of \mathbf{b} , also known as Henderson's mixed model equations, which are presented later in this chapter.

1.5.3 *Distribution of the Response Variable Conditional on Random Effects ($\mathbf{y} \mid \mathbf{b}$)*

The distribution selected by the researcher from the population under study should be true or a good approximation that represents the likely distribution of the response variable. A good representation of the population distribution of a response variable should not only take into account the nature of the response variable (e.g., continuous, discrete, etc.) and the shape of the distribution but should also provide a good model for the relationship between the mean and variance. For the distribution of the dataset, in this chapter, we assume that it is normally distributed with a mean μ and a variance σ^2 $\{y_{ij} \sim (\mu, \sigma^2)\}$ and, for the random effects, it will assume a normal distribution with mean 0 and constant variance σ_b^2 $\{b_j \sim (0, \sigma_b^2)\}$.

1.5.4 Types of Factors and Their Related Effects on LMMs

In an LMM, there are two types of factors, namely, fixed factors that make up the systematic part and random factors that are the stochastic part, and their related effects on the dependent variable (response). In the following sections, we provide a brief description of these factors and their implications in the context of an LMM.

1.5.4.1 Fixed Factors

A fixed factor is commonly used in standard analysis of variance (ANOVA) or analysis of covariance (ANCOVA) models. It is defined as a categorical or classification variable, for which the researcher has included all levels (or conditions) in the model that are of interest in the study. Fixed factors may include qualitative covariates, such as gender; classification variables implied by a sampling design, such as a region or a stratum, or by a study design, such as the method of treatment in a randomized clinical trial; and so on. The levels of a fixed factor are chosen to represent specific conditions so that they can be used to define contrasts (or sets of contrasts) of interest in the research study.

1.5.4.2 Random Factors

A random factor is a classification variable with levels that can be randomly sampled from a population with different levels of study. All possible levels of a random factor are not present in the dataset, but this is the intention of the researcher, i.e., to make inference about the entire population of levels from the selected sample of these factor levels. Random factors are considered in an analysis such that the change in the dependent variable across random factor levels can be evaluated and the results of the data analysis can be generalized to all random factor levels in the population.

1.5.4.3 Fixed Versus Random Factors

In contrast to fixed factor levels, random factor levels do not represent conditions specifically chosen to meet the objectives of the study. However, depending on the objectives of the study, the same factor may be considered as either a fixed factor or a random factor.

Fixed effects, commonly referred to as regression coefficients or fixed effect parameters, describe the relationships between the dependent variable and predictor variables (i.e., fixed factors or continuous covariates) for an entire population of units of analysis or for a relatively small number of subpopulations defined by the levels of a fixed factor. Fixed effects may describe the contrasts or differences

between levels of a fixed factor (e.g., sex between males and females) in the mean responses for a continuous dependent variable or may describe the effect of a continuous covariate on the dependent variable. Fixed effects are assumed to be unknown fixed quantities in an LMM and are estimated based on analysis of the data collected in a study.

Random effects are random values associated with the levels of a random factor (or factors) in an LMM. These values, which are specific to a given level of a random factor, generally represent random deviations from the relationships described by fixed effects. For example, random effects associated with levels of a random factor may enter an LMM as random intercepts (random deviations for a given subject or group as an overall intercept) or as random coefficients (random deviations for a given subject or group from the total fixed effects) in the model. In contrast to fixed effects, random effects are represented as stochastic variables in an LMM.

1.5.5 Nested Versus Crossed Factors and Their Corresponding Effects

When a given level of one factor (random or fixed) can be measured only at a single level of another factor and not across multiple levels, then the levels of the first factor are said to be nested within the levels of the second factor. The effects of the nested factor on the response variable are known as nested effects. For example, suppose that you want to conduct a particular study at the primary level in a school zone, you would select schools and classrooms at random. Classroom levels (one of the random factors) are nested within school levels (another random factor), since each classroom can appear within a single school.

When a given level of one factor (random or fixed) can be measured across multiple levels of another factor, one factor is said to be crossed with the other and the effects of these factors on the dependent variable are known as crossover effects.

1.5.6 Estimation Methods

Standard methods of estimation in mixed models with a normal response are maximum likelihood (ML) and restricted maximum likelihood (REML). The linear mixed effects model is as follows:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b} + \boldsymbol{\epsilon}$$

The variance–covariance matrix \mathbf{V} for a one-way analysis of variance (ANOVA) with a randomized block effect and with six observations is equal to:

$$V = \text{Var}(\mathbf{y}) = \mathbf{ZGZ}' + \sigma^2 \mathbf{I} = \begin{pmatrix} \sigma^2 + \sigma_b^2 & \sigma_b^2 & 0 & 0 & 0 & 0 \\ \sigma_b^2 & \sigma^2 + \sigma_b^2 & 0 & 0 & 0 & 0 \\ 0 & 0 & \sigma^2 + \sigma_b^2 & 0 & 0 & 0 \\ 0 & 0 & \sigma_b^2 & \sigma^2 + \sigma_b^2 & 0 & 0 \\ 0 & 0 & 0 & 0 & \sigma^2 + \sigma_b^2 & \sigma_b^2 \\ 0 & 0 & 0 & 0 & \sigma_b^2 & \sigma^2 + \sigma_b^2 \end{pmatrix}$$

The variance of y_{11} is $V_{11} = \sigma^2 + \sigma_b^2$ and the covariance between y_{11} and y_{21} is $V_{12} = V_{21} = \sigma_b^2$. These two observations come from the same block. The covariance between y_{11} and other observations is zero. In matrix V , all possible covariances can be found.

1.5.6.1 Maximum Likelihood

The likelihood function l is a function of the observations and the model parameters. It gives us a measure of the probability of looking at a particular observation \mathbf{y} , given a set of model parameters $\boldsymbol{\beta}$ and \mathbf{b} . The likelihood function for $\mathbf{y} \mid \boldsymbol{\beta}$ and \mathbf{b} for a mixed model is given by:

$$l(\mathbf{y} \mid \mathbf{b}) = -\frac{n}{2} \log(2\pi) - \frac{1}{2} \log|\mathbf{R}| - \frac{1}{2} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{b})' \mathbf{R}^{-1} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{b})$$

and

$$l(\mathbf{b}) = -\frac{N_b}{2} \log(2\pi) - \frac{1}{2} \log|\mathbf{G}| - \frac{1}{2} \mathbf{b}^T \mathbf{G}^{-1} \mathbf{b}$$

where N_b represents the total number of random effect levels. Therefore, the joint distribution of \mathbf{y} and \mathbf{b} is equal to:

$$l(\mathbf{y}, \mathbf{b}) = -\left(\frac{1}{2}\right) (\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{b})^T \mathbf{R}^{-1} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{b}) - \left(\frac{1}{2}\right) \mathbf{b}^T \mathbf{G}^{-1} \mathbf{b}$$

Now, after deriving the above expression with respect to $\boldsymbol{\beta}$ and \mathbf{b} and then setting it to zero and solving the resulting equations with respect to $\boldsymbol{\beta}$ and \mathbf{b} , the maximum likelihood estimators are obtained:

$$\frac{\partial l(\mathbf{y}, \mathbf{b})}{\partial \boldsymbol{\beta}^T} = \mathbf{X}^T \mathbf{R}^{-1} \mathbf{y} - \mathbf{X}^T \mathbf{R}^{-1} \mathbf{X} \boldsymbol{\beta} - \mathbf{Z}^T \mathbf{R}^{-1} \mathbf{Z} \mathbf{b}$$

$$\frac{\partial l(\mathbf{y}, \mathbf{b})}{\partial \mathbf{b}^T} = \mathbf{Z}^T \mathbf{R}^{-1} \mathbf{y} - \mathbf{X}^T \mathbf{R}^{-1} \mathbf{Z} \boldsymbol{\beta} - \mathbf{Z}^T \mathbf{R}^{-1} \mathbf{Z} \mathbf{b}$$

Setting them to zero and solving for $\boldsymbol{\beta}$ and \mathbf{b} , we obtain the following linear mixed equations:

$$\begin{pmatrix} X^T R^{-1} X & X^T R^{-1} Z \\ Z^T R^{-1} X & Z^T R^{-1} Z + G^{-1} \end{pmatrix} \begin{pmatrix} \hat{\beta} \\ \hat{b} \end{pmatrix} = \begin{pmatrix} X^T R^{-1} y \\ Z^T R^{-1} y \end{pmatrix}$$

The solution can be written as:

$$\begin{pmatrix} \hat{\beta} \\ \hat{b} \end{pmatrix} = \begin{pmatrix} X^T R^{-1} X & X^T R^{-1} Z \\ Z^T R^{-1} X & Z^T R^{-1} Z + G^{-1} \end{pmatrix}^{-1} \begin{pmatrix} X^T R^{-1} y \\ Z^T R^{-1} y \end{pmatrix}$$

Here, β is the vector of fixed effects parameters and b is the vector of random effects parameters. The information of these parameters is related to the two covariance matrices G and R , and it no longer depends on V as in the previous solution. Moreover, this solution, which is known as Henderson's (1950) mixed model equations, is computationally much more efficient than the previous one given for the parameters (β and b) since it does not need to obtain the inverse of the matrix $V = ZGZ' + R$. The solution to these mixed model linear equations is based on the assumption that we know the components of G and R , which, in practice, need to be estimated. Therefore, the following is a popular method for estimating the variance components of G and R , which is extremely versatile and powerful.

1.5.6.2 Restricted Maximum Likelihood Estimation

The restricted maximum likelihood method is also known as the residual maximum likelihood method and is extremely useful, among other things, for estimating variance components. This method is also based on the maximum likelihood method, but, instead of maximizing the likelihood function of the original data, it maximizes the likelihood function over a set of errors obtained by removing the variables from the original response to fixed effects, which are assumed to be known. That is, now instead of maximizing over y is maximized over Ky but to obtain the variance components, it is assumed that K is a matrix of constants, such that $KX = \mathbf{0}$, which implies that:

$$E(Ky) = (KX\beta + KZb + K\epsilon) = \mathbf{0}$$

$$\text{Var}(Ky) = (K^T VK)$$

This implies that Ky is distributed over $N(\mathbf{0}, K^T VK)$ and the likelihood of Ky is called the restricted maximum likelihood (REML). There are many options to choose K and typically $K = I - X(X^T X)^{-1} X^T$, which is the ordinary least squares residual operator used. Therefore, the log likelihood of Ky is equal to

$$l(V|Ky) = -\frac{n-p}{2} \log(2\pi) - \frac{1}{2} \log|K^T VK| - \frac{1}{2} (y^T K^T)' K^T VK^{-1} (Ky)$$

This log likelihood after some algebra, according to Stroup (2012), is equal to:

$$l(\mathbf{V}|\mathbf{K}\mathbf{y}) = -\frac{n-p}{2} \log(2\pi) - \frac{1}{2} \log|\mathbf{V}| - \frac{1}{2} \log(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}) - \frac{1}{2} \mathbf{r}'\mathbf{r}$$

where $p = \text{rank}(\mathbf{X})$ and $\mathbf{r} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_{ML}$, where $\hat{\boldsymbol{\beta}}_{ML} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1} \mathbf{X}'\mathbf{V}^{-1}\mathbf{y}$

The variance components of \mathbf{G} and \mathbf{R} are estimated with iterative methods such as the Newton–Raphson or Fisher’s scoring method, which maximizes the likelihood function $l(\mathbf{V}|\mathbf{K}\mathbf{y})$ with respect to the variance components. The maximization process starts with starting values for the variance components to estimate \mathbf{G} and \mathbf{R} , and, with these values of \mathbf{G} and \mathbf{R} , it is possible to estimate a new, more refined version of the parameters $\boldsymbol{\beta}$ and \mathbf{b} ; then, these values are used to update the estimates of the variance components of the matrices \mathbf{G} and \mathbf{R} , and this process continues until the established convergence is met.

1.5.7 One-Way Random Effects Model

Suppose that we randomly select a possible levels from a sufficiently large set of levels of the factor of interest. In this case, we say that the factor is random. Random factors are usually categorical. Continuous covariates that cannot be measured at random levels are generally known as “systematic” or “fixed” effects (e.g., linear, quadratic, or even exponential terms). Random effects are not systematic. Let us assume a simple one-way model:

$$y_{ij} = \mu + \tau_i + \varepsilon_{ij}; \quad i = 1, 2, \dots, a; \quad j = 1, 2, \dots, n_i$$

However, in this case, the treatment effects and the error term are random variables, i.e., $\tau_i \sim N(0, \sigma_\tau^2)$ and $\varepsilon_{ij} \sim N(0, \sigma^2)$, respectively. The terms τ_i and ε_{ij} are uncorrelated, commonly referred to as “variance components.”

There can be some confusion about the differences between noise factors and random factors. Noise factors can be fixed or random.

Factors are random when we think of them as being/coming from a random sample of a larger population, and their effect is not systematic. It is not always clear when a factor is random. For example, suppose that the vice president of a chain of stores is interested in the effects of implementing a management policy in his stores and the experiment includes all five existing stores, he might consider “the store” as a fixed factor because the levels of the factor “store” do not come from a random sample. However, if the store chain has 100 stores and takes 5 stores for the experiment, as the company is considering rapid expansion and plans to implement the selected new policy at the new locations, then “store” could be considered as a random factor.

In fixed effects models, the researcher’s interest would focus on testing the equality of means of treatments (stores). This would not be appropriate, however, for the case in which 5 stores are randomly selected out of 100 because the

treatments are randomly selected and we are interested in the population of treatments (stores), not in a particular store or group of stores. The appropriate hypothesis test for this random effect model would be

$$H_0 : \sigma_\tau^2 = 0 \quad \text{vs} \quad H_a : \sigma_\tau^2 > 0$$

Partitioning a standard analysis of variance from the total sum of squares still works; however, the form of the appropriate test statistic depends on the expected mean squares. In this case, the appropriate test statistic would be

$$F_c = \frac{\text{Mean Square}_{\text{Treatments}}}{\text{Mean Square}_{\text{Error}}},$$

F_c follows an F -distribution (Fisher–Snedecor) with degrees of freedom $a - 1$ in the numerator and $N - a$ in the denominator, where $N = \sum_{i=1}^a n_i$.

In a completely random model, we are interested in estimating the variance components. σ_τ^2 and σ^2 . To do so, we use the analysis of variance method, which consists of equating the expected mean squares with the observed values as follows:

$$\hat{\sigma}^2 + n\hat{\sigma}_\tau^2 = \text{Mean Square}_{\text{Treatments}}$$

where $\hat{\sigma}^2 = \text{Mean Square}_{\text{Error}}$

$$\hat{\sigma}_\tau^2 = \frac{\text{Mean Square}_{\text{Error}} - \hat{\sigma}^2}{n}$$

1.5.8 Analysis of Variance Model of a Randomized Block Design

Consider a one-way analysis of variance model with a randomized block additive effect. Assume two treatments and three blocks,

$$y_{ij} = \mu + \tau_i + b_j + \epsilon_{ij}$$

where $b_j \sim N(0, \sigma_b^2)$ and $\epsilon_{ij} \sim N(0, \sigma^2)$ with $i = 1, 2, 3$ and $j = 1, 2, 3$. The random effects b_j and ϵ_{ij} are independent and uncorrelated. In addition, treatment effects are assumed to be fixed. The matrix notation of this model is as follows:

$$\begin{pmatrix} y_{11} \\ y_{21} \\ y_{12} \\ y_{22} \\ y_{13} \\ y_{23} \end{pmatrix} = \underbrace{\begin{pmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \end{pmatrix}}_{\mathbf{X}} \underbrace{\begin{pmatrix} \mu \\ \tau_1 \\ \tau_2 \end{pmatrix}}_{\boldsymbol{\beta}} + \underbrace{\begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix}}_{\mathbf{Z}} \underbrace{\begin{pmatrix} b_1 \\ b_2 \\ b_3 \end{pmatrix}}_{\mathbf{b}} + \underbrace{\begin{pmatrix} \varepsilon_{11} \\ \varepsilon_{21} \\ \varepsilon_{12} \\ \varepsilon_{22} \\ \varepsilon_{13} \\ \varepsilon_{23} \end{pmatrix}}_{\boldsymbol{\varepsilon}}$$

where $\mathbf{b} \sim N(\mathbf{0}, \mathbf{G})$ and $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \mathbf{R})$. The variance–covariance matrix \mathbf{G} for the random effects in this case is a diagonal matrix 3×3 with diagonal elements σ_b^2 . Note how the matrix representation of this model exactly corresponds to the mixed model formulation. That is,

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b} + \boldsymbol{\varepsilon}, \quad \text{where } \mathbf{b} \sim N(\mathbf{0}, \mathbf{G}) \text{ and } \boldsymbol{\varepsilon} \sim N(\mathbf{0}, \mathbf{R}).$$

Example An animal nutritionist is interested in comparing the effect of three diets on weight gain in piglets. To conduct the experiment, the nutritionist randomly selects 3 litters from a set of 20, each containing 3 healthy, similar-sized, recently weaned piglets. In each litter, three piglets are selected and each piglet is randomly assigned to a treatment.

A randomized complete block design (RCBD) is a variation of the completely randomized design (CRD). In this design, blocks of experimental units are chosen in such a way that the units within the blocks are as homogeneous as possible with respect to each other (homogeneous) and different between blocks. In a randomized complete block design, generally in each block, there is one experimental unit for each treatment, but this does not limit having more than one experimental unit for each treatment in each block.

An RCBD has two sources of variation: the factor of interest that includes the treatments to be studied and the “block factor” that identifies the litters used in the experiment.

Assumptions in RCBD:

1. Sampling: Blocks (litters) are independently randomly selected and treatments are randomly assigned to each of the experimental units within each block.
2. Errors are normal, independent, and identically normally distributed with a zero mean and a constant variance σ^2 .

Table 1.23 lists the weight in kilograms of piglets from three different litters under three different diets. To make inferences about the pattern of weight gain for the entire population (all litters) of piglets, the litters must be considered in the model as a random effect. Thus, the linear mixed model describing the variability of piglet weight gain in this research, as a function of diets, is as follows:

Table 1.23 Weight gain (kilograms) of the three litters of piglets

Litter	Diet1	Diet2	Diet3
1	54.3	53.1	59.7
	53.6	52.4	59.7
	55.2	57.1	67.2

Table 1.24 Analysis of variance of the randomized complete block design

Sources of variation	Degrees of freedom
Blocks	$b - 1 = 3 - 1 = 2$
Diet	$t - 1 = 3 - 1 = 2$
Error	$(t - 1)(b - 1) = 4$
Total	$tb - 1 = 8$

$$y_{ij} = \mu + \tau_i + b_j + \varepsilon_{ij} \text{ for } i = 1, 2, 3; j = 1, 2, 3$$

where y_{ij} is the weight observed in the ij th piglet, μ is the overall mean, τ_i is the fixed effect due to i th diet, b_j is the random effect due to the j th block (litter) assuming $b_j \sim N(0, \sigma_b^2)$, and ε_{ij} is the independent and identically distributed, approximately normal, observed error term with mean 0 and variance σ^2 , i.e., $\varepsilon_{ij} \sim N(0, \sigma^2)$.

Random effects, b_j and ε_{ij} , are assumed to be independent and uncorrelated. Table 1.24 shows an outline of the analysis of variance for this dataset.

The SAS program to analyze this dataset is as follows:

```
proc glimmix data=piglets;
class litter diet;
model gain=diet/ddfm=satterthwaite;
random litter;
lsmeans diet/lines;
contrast "Diet1 vs Diet2" diet 1 -1 0;
contrast "Diet2 vs Diet3" diet 1 0 -1;
run; quit;
```

In the previous syntax, we can mention two commands of great importance in this example: (1) the “`ddfm = satterthwaite`” command allows to make a correction of the degrees of freedom, and this correction is of great importance when the number of experimental units (UE) is different in each one of the treatments and (2) the command “`lines`” serve to obtain the means of “`lsmeans`” but are grouped with letters, and, if these averages appear with different letters, then they reflect significant differences.

The output for this code is shown in Table 1.25. Subsection (a) of this table shows the estimated variance due to litter ($\hat{\sigma}_{\text{litter}}^2 = 5.3117$) and the mean squared error ($\hat{\sigma}^2 = 3.2961$). The analysis of variance, part (b), shows that there is a highly significant effect of diet on piglet weight gain ($P = 0.0091$). In the results (part c), we also observe the estimated means and its standard errors (obtained with “`lsmeans diet/lines`”) and the grouping of means that are statistically different (part d). In these last results, we can observe that the weight gain of piglets under treatments I and II

Table 1.25 Results of the analysis of variance of the three different diets tested on piglet weight gain

(a) Covariance parameter estimates					
Cov Parm		Estimate	Standard error		
Litter		5.3117	6.4573		
Residual		3.2961	2.3307		
(b) Type III tests of fixed effects					
Effect	Num DF	Den DF	F-value	Pr > F	
Diet			19.02	0.0091	
(c) Dietary least squares means					
Diet	Estimate	Standard error	DF	t-value	Pr > t
I	54.3667	1.6939	3.406	32.10	<0.0001
II	54.2000	1.6939	3.406	32.00	<0.0001
III	62.2000	1.6939	3.406	36.72	<0.0001
(d) T grouping of the dietary least squares means ($\alpha = 0.05$)					
LS means with the same letter are not significantly different					
Diet	Estimate				
III	62.2000	A			
I	54.3667	B			
II	54.2000	B			

Table 1.26 Analysis of variance under a CRD

Type III tests of fixed effects				
Effect	Num DF	Den DF	F-value	Pr > F
Diet			7.28	0.0248

are not statistically different from each other, but they are statistically different with respect to treatment III.

Since the researcher wishes to make an inference about the entire population of litters, the factor “litter” must be entered as a random effect; otherwise, the ability of the *F*-test to detect differences between treatments is diminished because the *P*-value changes from 0.0091 to 0.0248. Another way to see the importance of including random effects in an ANOVA is to calculate the relative efficiency (RE) between the two models.

Table 1.26 shows the results of the analysis of variance under a completely randomized design (CRD), i.e., $y_{ij} = \mu + \text{litter}_i + e_{ij}$ is as follows:

In this case, if the experiment had been analyzed under a CRD, then the relative efficiency (RE) between an RCBD and a CRD would be:

$$RE = \frac{CME_{CRD}}{CME_{RCBD}} = \frac{\frac{(SSB_{RCBD} + SCE_{RCBD})}{t(b-1)}}{CME_{RCBD}} = \frac{(b-1)MSB_{RCBD} + b(t-1)CME_{RCBD}}{(bt-1)CME_{RCBD}}$$

where CME_{DCA} is the mean squared error under a CRD, CME_{RCBD} is the mean squared error under an RCBD, SSB_{DBCA} is the sum of squares due to blocks in an RCBD, SSE_{DBCA} is the sum of squares of errors in an RCBD, MSB_{DBCA} is the mean

Table 1.27 Fit statistics of a CRD and RCBD

Fit statistics	CRD	RCBD
-2 Res log likelihood	33.24	31.01
AIC (smaller is better)	41.24	35.01
AICC (smaller is better)	81.24	39.01
BIC (smaller is better)	40.41	33.20
CAIC (smaller is better)	44.41	35.20
HQIC (smaller is better)	37.90	31.38
Pearson's chi-square	51.65	19.78
Pearson's chi-square / DF	8.61	3.30

square due to blocks, and t and b are the number of treatments and blocks, respectively. If blocks are not useful, then the RE would be equal to 1. The higher the RE, the more effective the blocking is in reducing the error variance. This value can be interpreted as the relationship r/b , where r is the number of experimental units that would have to be assigned to each treatment if a CRD were used instead of an RCBD.

In Table 1.27, we can observe the mean squared error (MSE) of a CRD and RCBD (Pearson's chi-square / DF) obtained with the GLIMMIX procedure in SAS as well as a series of fit statistics.

The MSE for a CRD and an RCBD are 8.61 and 3.3, respectively. Substituting these values into the above equation, we obtain

$$ER = \frac{CME_{CRD}}{CME_{RCBD}} = \frac{8.61}{3.3} = 2.609.$$

This value indicates that, an RCBD is 2.609 times more efficient than a CRD. In other words, this implies that it should have taken, at least, 8 ($2.609 \times 3 \approx 8$) more experimental units \times treatment units in a CRD to obtain the same MSE as that obtained in an RCBD.

1.6 Exercises

Exercise 1.6.1 The following dataset corresponds to the growth of pea plants, in eye units, in tissue culture with auxins (0.114 mm). The purpose of this experiment was to test the effects of the addition of various types of sugars to the culture medium on growth in length. Pea plants were randomly assigned to one of five treatments: control (no sugar), 2% of glucose, 2% of fructose, 1% of glucose + 1% of fructose, and 2% sucrose. A total of 10 observations were taken in each of the treatments, assuming that the measurements are approximately normally distributed with constant variance. Here, the individual plants to which the treatments were applied are the experimental units. The data from this experiment are shown below (Table 1.28):

Table 1.28 Growth of pea plants in the culture medium with auxins with different types of sugars

Plant	Control	2% Glucose	2% Fructose	1% Glucose +1% fructose	2% Sucrose
1	75	57	58	58	62
2	67	58	61	59	66
3	70	60	56	58	65
4	75	59	58	61	63
5	65	62	57	57	64
6	71	60	56	56	62
7	67	60	61	58	65
8	67	57	60	57	62
9	76	59	57	57	62
10	68	61	58	59	67

Table 1.29 Growth (height in centimeters) of the two forage species with three types of fertilizers plus a control

	Fertilizer			
	Control	F1	F2	F3
Species A	21	32	22.5	28
	19.5	30.5	26	27.5
	22.5	25	28	31
	21.5	27.5	27	29.5
	20.5	28	26.5	30
	21	28.6	25.2	29.2
Species B	23.7	30.1	30.6	36.1
	23.8	28.9	30.6	36.1
	23.8	30.9	28.1	38.7
	23.7	34.4	34.9	37.1
	22.8	32.7	30.1	36.8
	24.4	32.7	25.5	37.1

- (a) Write the statistical model that best describes this dataset, indicating its components.
- (b) Calculate the analysis of variance for this experiment.
- (c) Is there any significant difference between treatments on average plant growth?

Exercise 1.6.2 A forage company wants to test three different types of fertilizers (F1, F2, and F3) for the production of two forage species (A and B) for cattle and compare them with a fertilizer they usually apply, which we will call control. For this, he decides to use 48 pots with 6 replications in the greenhouse to test the combinations of fertilizers and forage species. The data from this experiment are shown in Table 1.29:

- (a) Write and describe the statistical model of the experimental design with all its components.
- (b) Calculate the analysis of variance for this experiment.
- (c) Is there any significant difference between treatments on average plant growth?

Exercise 1.6.3 The data in this experiment are the number of plants regrown after grazing with sheep–goats. The initial size of the plant at the top of its rootstock is recorded, and the weight of seeds (g) that it produces at the end of the season is the response or dependent variable. The data for this experiment are as follows (Table 1.30):

- (a) List and describe all the components of the linear mixed model.
- (b) Calculate the ANOVA for this dataset and answer the following questions:

Is seed weight influenced by the type of grazing?

Is seed weight influenced by the plant size?

Is the effect of grazing type on plant size influenced by the initial plant size?

Exercise 1.6.4 An experiment was conducted to study the effect of supplementation of weaned lambs on health and growth rate when exposed to helminthiasis. A total of 16 Dorper (breed 1) and 16 Red Maasai (breed 2) lambs were treated with an anthelmintic at 3 months of age (after weaning) and randomly allocated into “blocks” of 4 per breed, classified on the basis of 3-month body weight for supplemented and unsupplemented groups. Therefore, two lambs in each block were randomly allocated to supplemented (night-fed cotton seed meal and wheat bran) and unsupplemented groups. All lambs were kept on grazing for a further 3 months. Data recorded included the initial body weight (kilograms) at weaning and weight at 3 months after weaning, percentage red blood cell volume (RBCV), and fecal egg count (FEC) at 6 months of age. Data from this experiment are shown below (Table 1.31):

- (a) List and describe all the components of the linear mixed model.
- (b) Calculate the ANOVA for this dataset and answer the following questions:

Did supplementation improve weight gain? Did supplementation affect PRBC and FEC, and were there differences in weight gain, PRBC, or FEC between breeds?

Table 1.30 Fruit production after grazing

Size	Fruit	Grazing
6.225	59.77	No Grazing
6.487	60.98	No Grazing
4.919	14.73	No Grazing
5.13	19.28	No Grazing
5.417	34.25	No Grazing
5.359	35.53	No Grazing
7.614	87.73	No Grazing
6.352	63.21	No Grazing
4.975	24.25	No Grazing
6.93	64.34	No Grazing
6.248	52.92	No Grazing
5.451	32.35	No Grazing
6.013	53.61	No Grazing
5.928	54.86	No Grazing
6.264	64.81	No Grazing
7.181	73.24	No Grazing
7.001	80.64	No Grazing
4.426	18.89	No Grazing
7.302	75.49	No Grazing
5.836	46.73	No Grazing
10.253	116.05	Grazing
6.958	38.94	Grazing
8.001	60.77	Grazing
9.039	84.37	Grazing
8.91	70.11	Grazing
6.106	14.95	Grazing
7.691	70.7	Grazing
8.988	80.31	Grazing
8.975	82.35	Grazing
9.844	105.07	Grazing
8.508	73.79	Grazing
7.354	50.08	Grazing
8.643	78.28	Grazing
7.916	41.48	Grazing
9.351	98.47	Grazing
7.066	40.15	Grazing
8.158	52.26	Grazing
7.382	46.64	Grazing
8.515	71.01	Grazing
8.53	83.03	Grazing

Table 1.31 Supplementation trial in Dorper (breed 1) and Red Maasai (breed 2) lambs

Id	Race	Sex	Supplement	Block	IW	FW	PRBC	FEC	WG
349	1	2	1	1	8	8.9	10	6500	0.9
326	1	2	1	1	9	10.1	11	2650	1.1
393	1	1	1	2	12	12.6	22	750	0.6
71	1	1	1	2	12.3	14.6	15	5200	2.3
271	1	1	1	3	13	13.7	19	4800	0.7
382	1	2	1	3	15.5	16.8	24	2450	1.3
85	1	2	1	4	16.3	18.2	27	200	1.9
176	1	2	1	4	15.9	17.7	21	3000	1.8
286	1	2	2	1	11	13.6	21	1600	2.6
183	1	1	2	1	9.9	11.7	21	450	1.8
21	1	2	2	2	11.6	13.1	25	2900	1.5
122	1	1	2	2	12.5	14.8	25	300	2.3
374	1	1	2	3	14.6	17.9	19	2250	3.3
32	1	2	2	3	14.2	16.9	22	2800	2.7
282	1	2	2	4	16.3	20.2	20	750	3.9
94	1	1	2	4	16.7	17.7	13	5600	1
127	2	2	1	1	7.5	8.1	26	1350	0.6
216	2	2	1	1	8.2	9.3	19	1150	1.1
133	2	1	1	2	10.1	11.7	30	200	1.6
249	2	1	1	2	8.8	10.4	28	0	1.6
123	2	2	1	3	1.6	12.6	23	600	1
222	2	2	1	3	11.3	13.5	24	1500	2.2
290	2	2	1	4	12.3	14.3	22	1950	2
148	2	1	1	4	13.1	14.9	26	500	1.8
142	2	2	2	1	8.2	11.5	25	850	3.3
154	2	2	2	1	9.5	12.2	35	700	3.7
166	2	1	2	2	9.7	12.8	29	400	3.1
322	2	1	2	2	8.6	12	26	800	3.4
156	2	1	2	3	10.2	13	28	1550	2.8
161	2	2	2	3	11.2	14.6	22	550	3.4
321	2	1	2	4	12.1	15.9	25	1250	3.8
324	2	1	2	4	13.8	18.1	24	1100	4.3

IW initial weight, *FW* final weight, *PRBC* percentage of red blood cells, *FEC* fecal egg count, *WG* weight gain

Appendix

Population	Plant	Stamens	Eggs	Total no. of flowers	Ratio (stamens/ovules)
St. Croix	1	30.75	13.75	8	2.24
St. Croix	2	33.83	16.17	12	2.09
St. Croix	3	35.67	16.33	6	2.18
St. Croix	4	35.40	17.40	14	2.03
St. Croix	5	33.50	23.50	13	1.43
St. Croix	6	37.40	18.40	10	2.03
St. Croix	7	33.57	21.29	25	1.58
St. Croix	8	29.86	28.71	20	1.04
St. Croix	9	33.80	29.60	17	1.14
St. Croix	10	31.60	25.80	14	1.22
St. Croix	11	32.57	27.50	21	1.18
St. Croix	12	31.80	24.00	13	1.33
St. Croix	13	35.25	17.75	8	1.99
St. Croix	14	30.00	16.83	13	1.78
St. Croix	15	30.50	18.75	9	1.63
St. Croix	16	32.20	21.40	13	1.50
St. Croix	17	32.40	26.25	12	1.23
St. Croix	18	38.50	17.75	8	2.17
St. Croix	19	37.00	25.83	16	1.43
St. Croix	20	33.00	25.25	8	1.31
St. Croix	21	31.40	25.20	15	1.25
St. Croix	22	31.80	25.60	14	1.24
St. Croix	23	30.40	19.20	15	1.58
St. Croix	24	35.20	22.40	22	1.57
St. Croix	25	27.80	20.80	10	1.34
St. Croix	26	31.29	22.71	14	1.38
St. Croix	27	32.83	22.33	20	1.47
St. Croix	29	31.20	17.40	14	1.79
St. Croix	30	33.00	19.20	13	1.72
St. Croix	31	33.80	22.20	13	1.52
St. Croix	32	32.22	27.63	31	1.17
St. Croix	33	32.91	28.73	18	1.15
St. Croix	34	34.50	15.75	9	2.19
St. Croix	35	28.33	17.33	8	1.63
St. Croix	36	30.71	23.14	14	1.33
St. Croix	37	33.00	24.00	14	1.38
St. Croix	38	31.00	20.50	4	1.51
St. Croix	39	35.00	21.83	15	1.60

(continued)

Population	Plant	Stamens	Eggs	Total no. of flowers	Ratio (stamens/ovules)
St. Croix	40	35.00	18.00	10	1.94
Cedar Creek	1	30.17	18.67	16	1.62
Cedar Creek	2	32.43	15.14	23	2.14
Cedar Creek	3	28.00	14.00	15	2.00
Cedar Creek	4	29.22	16.89	35	1.73
Cedar Creek	5	36.00	17.14	20	2.10
Cedar Creek	6	30.83	20.17	15	1.53
Cedar Creek	7	31.75	18.00	18	1.76
Cedar Creek	8	29.25	19.00	8	1.54
Cedar Creek	9	32.78	24.44	24	1.34
Cedar Creek	10	32.67	22.83	17	1.43
Cedar Creek	11	31.43	21.00	28	1.50
Cedar Creek	15	33.50	29.50	4	1.14
Cedar Creek	16	32.83	15.17	20	2.16
Cedar Creek	17	35.00	15.00	9	2.33
Cedar Creek	18	33.17	13.83	15	2.40
Cedar Creek	19	33.29	27.14	28	1.23
Cedar Creek	20	35.50	19.83	16	1.79
Cedar Creek	21	35.71	18.86	21	1.89
Cedar Creek	23	31.38	25.63	5	1.22
Cedar Creek	25	28.25	17.50	11	1.61
Cedar Creek	27	31.82	24.91	37	1.28
Cedar Creek	28	35.13	26.88	23	1.31
Cedar Creek	32	33.75	21.63	26	1.56
Cedar Creek	33	32.00	20.80	14	1.54
Cedar Creek	34	36.29	17.00	18	2.13
Cedar Creek	35	28.60	16.40	11	1.74
Cedar Creek	36	33.00	20.80	14	1.59
Cedar Creek	37	34.90	25.11	49	1.39
Cedar Creek	38	34.80	19.60	18	1.78
Cedar Creek	40	30.00	21.17	16	1.42
Cedar Creek	41	34.50	20.50	16	1.68
Cedar Creek	42	37.75	29.00	18	1.30
Cedar Creek	43	33.50	20.75	10	1.61
Cedar Creek	44	33.00	22.40	12	1.47
Cedar Creek	45	35.50	21.50	16	1.65
Cedar Creek	46	32.50	22.00	14	1.48
Cedar Creek	47	32.67	16.67	8	1.96
Cedar Creek	48	35.75	21.50	26	1.66
Cedar Creek	49	31.38	22.88	22	1.37
Cedar Creek	50	33.83	20.50	17	1.65

Data: Larkspur plants from two populations in the state of Minnesota

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



Chapter 2

Generalized Linear Models



2.1 Introduction

In the generalized linear model (GLM) (which is not highly general) $y = X\beta + \epsilon$, the response variables are normally distributed, with constant variance across the values of all the predictor variables, and are linear functions of the predictor variables. Transformations of data are used to try to force the data into a normal linear regression model or to find a non-normal-type response variable transformation (discrete, categorical, positive continuous scale, etc.) that is linearly related to the predictor variables; however, this is no longer necessary. Instead of using a normal distribution, a positively skewed distribution with values that are positive real numbers can be selected. Generalized linear models (GLMs) go beyond linear mixed models, taking into account that the response variables are not of continuous scale (not normally distributed), GLMs are heteroscedastic, and there is a linear relationship between the mean of the response variable and the predictor or explanatory variables.

Nelder and Wedderburn (1972) implemented a unified methodology for linear models, thus opening a window for researchers to design models that can explain the variation of the phenomenon under study. Later, McCullagh and Nelder (1989) proposed an extension of linear models, called generalized linear models (GLMs). They pointed out that the key elements of a classical linear model are as follows: (i) the observations are independent, (ii) the mean of the observation is a linear function of some covariates, and (iii) the variance of the observation is a constant. To further extend these, points (ii) and (iii) are modified as follows: (ii') the mean of the observation is associated with a linear function of some covariates via a link function and (iii') the variance of the observation is a function of the mean. For more details, see the study by McCullagh and Nelder (1989). GLMs can be adapted to a wide variety of response variables. Special cases of GLMs include not only regression and analysis of variance (ANOVA) but also logistic regression, probit models, Poisson regression, log-linear models, and many more.

2.2 Components of a GLM

The construction of a GLM begins with choosing the distribution of the response variable, the predictor or explanatory variables to include in the systematic component, and how to connect the mean of the response to the systematic component. The three important components are described in the following sections:

2.2.1 *The Random Component*

The first component to specify is the random component, which consists of choosing a probability distribution for the response variable. This can be any member of the exponential family of distributions, such as normal, binomial, Poisson, gamma, and so on.

2.2.2 *The Systematic Component*

The second component of a GLM is the systematic component or linear predictor, which consists of a linear combination of explanatory variables (the predictor). The systematic component of a model is the fixed structural part of the model that explains the systematic variability between means. The linear predictor is found on the right-hand side of the equation in the specification of a linear or nonlinear regression model. Let x_1, x_2, \dots, x_p be the numerical (dummy) or discrete (category) predictor (explanatory) variables, then the linear predictor is

$$\eta_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_p x_{pi} = \mathbf{x}_i^T \boldsymbol{\beta}$$

where $\boldsymbol{\beta}^T = (\beta_0, \beta_1, \beta_2, \dots, \beta_p)$ is the vector of regression parameters and $\mathbf{x}_i^T = (1, x_{1i}, x_{2i}, \dots, x_{pi})$ is the vector of predictor variables. Although η is a linear function, the x 's can be nonlinear in form. For example, η can be a quadratic, cubic, or higher-order polynomial. The expected value of y_i and the linear predictor η_i are related through the link function. For example, in a Poisson GLM, the predictor is equal to $\log(\lambda_i) = \mathbf{x}_i^T \boldsymbol{\beta}$, since the link is a natural logarithm, better known as the link log.

In normal linear regression models, the focus is on η and finding the predictors or explanatory variables that best explain or predict the mean of the response variable. This is also important in a GLM. Problems such as multicollinearity in normal linear regression are also problems in generalized linear models.

2.2.3 Predictor’s Link Function η

Finally, we will look at the specification of the link function that maps the mean of the response variable to the linear predictor. The link function allows a nonlinear relationship between the mean of the response variable and the linear predictor, and this link $g(\cdot)$ connects the mean of the response variable with the linear predictor. That is,

$$g(\mu) = \eta$$

The function must be monotonous (and differentiable). The mean is equal, in turn, to the inverse transformation of $g(\cdot)$, that is,

$$\mu = g^{-1}(\eta)$$

The most natural and meaningful way to interpret the model parameters is in terms of the scale of the data. In other words,

$$\mu = g^{-1}(\eta) = g^{-1}(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_p x_{pi})$$

It is important to note that the link relates the mean of the response to the linear predictor and that this is different from transforming directly to the response variable. If the response variables are transformed (i.e., y 's), then a distribution must be selected, which describes the population distribution of the transformed data, thus making the original interpretation of the data more difficult. A transformation of the mean is generally not equal to the mean of the transformed values, that is, $g(E[y]) \neq E(g[y])$. For example, suppose we have a distribution with the following values (and probabilities):

y_i	1	2	3	4
$\text{prob}(Y = y_i)$	0.125	0.375	0.375	0.125

The mean of this distribution is $E[y] = 1 \times 0.125 + 2 \times 0.375 + 3 \times 0.375 + 4 \times 0.125 = 2.5$. Therefore, the logarithm of the mean of this distribution is $\ln(E[y]) = \ln(2.5) = 0.916$, whereas the mean of the logarithm is equal to $E(\ln[y]) = 0.845$. The value of the linear predictor η could potentially equal any value, but the expected values of the response variable – as in the case of counts or proportions – can be bounded. If there are no restrictions on the response variable (positive or negative real numbers), then the “identity link” function could be used, where the mean is identical to the linear predictor, that is,

$$\mu = \eta.$$

As mentioned before, the link function establishes a connection between the linear predictor η and the mean of the distribution μ . It is important to note that the link function in some cases is in a sense similar to a function transformation, in that it establishes only a mathematical connection between the parameters of the model. A function transformation is applied to the observations to better understand the relationship between the mean and the response variables or, in some cases, to stabilize the variance. Special cases are mentioned below:

- (a) For a normal distribution, the link function is the identity function, $\eta = \mu$, the variance function is constant. i.e., $\text{Var}(\mu) = 1$, and the scale parameter is the variance, i.e., $\phi = \sigma^2$, allowing the use of ordinary least squares in parameter estimation in procedures such as linear regression, analysis of variance (ANOVA) models, or analysis of covariance (ANCOVA) models.
- (b) In a binomial distribution, the response variable takes binary values like 0 and 1 or represents the relative frequency, i.e., $y_i = e_i/n_i$, where e_i is the number of successes and n_i is the number of trials. The mean is a probability (π) and therefore must be between 0 and 1. The linear predictor is not bounded. Therefore, the link function must map the real line within the interval $[0, 1]$. A natural link function for binomial data is the logit link:

$$\eta = \log\left(\frac{\pi}{(1 - \pi)}\right) \rightarrow \pi = \frac{e^\eta}{(1 + e^\eta)}$$

Another useful alternative for these types of data is the probit link function:

$$\eta = \Phi^{-1}(\pi) \rightarrow \pi = \Phi(\eta)$$

where Φ is the cumulative distribution function of a standard normal distribution.

The variance of the function has the form $\text{Var}(\pi) = (\pi/(1 - \pi))$ and the scale parameter ϕ is known and is equal to 1 ($\phi = 1$). The difference between the logit and probit estimators is important if the estimated probabilities are extremely small or extremely close to 1, indicating that large sample sizes are required for an effective inference. Both the logit and probit functions produce extremely close or equivalent results, especially with probability values around 0.5.

- (c) For a Poisson distribution, the link function is the natural log:

$$\eta = \log(\lambda) \rightarrow \lambda = e^\eta$$

The variance of the function has the form $\text{Var}(\lambda) = \lambda$, and, similar to the binomial distribution, the scale parameter is 1. Poisson models with a log link function are often referred to as log-linear models, commonly used when there are contingency (data frequency) tables with at least two entries.

(d) A gamma distribution has a link function of the form:

$$\eta = \frac{1}{\mu} \rightarrow \mu = \frac{1}{\eta}$$

The variance of the function is given by $\text{Var}(\mu) = \mu^2$ and the scale parameter ϕ is usually unknown. In some cases, the log link function is commonly used, which results in an exponential inverse link. It should be noted that the link function does not map the range of the means contained within the linear predictor. Therefore, given its limitations, the theory only provides reasonable approximations for most applications. An exponential distribution is a special case of the gamma distribution.

Previously, the classical methods for working with non-normal data – before the advances in computational methods – consisted of using direct transformation of the response variable, that is, the data were transformed using the function $t(y)$ before being analyzed. The goal of the transformation was to obtain a simple connection between the mean and the linear predictor. However, obtaining a consistent scale of variation when selecting a transformation is vitally important. The usual way for selecting a suitable transformation is based on the assumption that, within the region of variation of the random variable, the transformation can capture the variability adequately through a simple linear approximation of the mean. That is, if the random variable y has a distribution with a mean μ and variance $\sigma^2(\mu)$, we want to find a transformation $t(y)$ such that it is forced to have a constant variance (stabilizes the variance). The commonly used functions to stabilize variance are the square root (\sqrt{y}) when data have a Poisson distribution; the arcsine square root when data are binomial; and the logarithmic transformation for data with a constant coefficient of variation.

Table 2.1 provides an overview of the most common link functions that will give admissible values for certain types of response variables and the corresponding inverse of the link function.

Table 2.1 Common link functions for different response variables

Type of response	Media	Variance	$g(\mu) = \eta$	$g^{-1}(\eta)$
Normal	μ	σ^2		$\mu = \eta$
Poisson	c	λ	$\log(\lambda)$	$\lambda = e^{(\eta)}$
Binomial ratio	π	$\pi(1 - \pi)/N$	(logit) $\log(\pi/1 - \pi)$ (probit) $\Phi^{-1}(\pi)$	$\pi = e^{(\eta)}/(1 + e^{(\eta)})$ $\pi = \Phi(\eta)$
Exponential	μ	μ^2	$\text{Iog}(\mu)$	$\mu = e^{(\eta)}$
Gamma	μ		(inverse) $1/\mu$	$\mu = 1/\eta$
Negative binomial	λ	$\lambda + \lambda^2 c$	$\log(\lambda)$	$\lambda = e^{(\eta)}$

Note: Φ is the cumulative distribution function of a standard normal distribution; μ and π are the expected values of the response; η is the linear predictor; and ϕ is the scale parameter

2.3 Assumptions of a GLM

According to McCullagh and Nelder (1989) and Agresti (2013) in Chap. 4, a GLM is defined under the following assumptions:

- (a) The data y_1, y_2, \dots, y_n are independent.
- (b) The response variable y_i does not necessarily have to have a normal distribution, but we usually assume a distribution from an exponential family (e.g., binomial, Poisson, multinomial, gamma, etc.).
- (c) A GLM does not assume a linear relationship between the dependent variable and the independent variables, but it does assume a linear relationship between the response transformed in terms of the link function and the explanatory variables; for example, for $\text{logit}(\pi)$ from a binary logistic regression, $\text{logit}(\pi) = \beta_0 + \beta x$.
- (d) The predictor (explanatory) variables may be in terms of power or some other nonlinear transformations of the original independent variables.
- (e) The assumption of homogeneity of variance need not be satisfied. In fact, it is not possible in many cases, given the structure of the model and the presence of overdispersion (when the observed variance is larger than what the model assumes).
- (f) Errors are independent but are not normally distributed.
- (g) The estimation method is maximum likelihood (ML) or other methods instead of ordinary least squares (OLS) to estimate the parameters.

2.4 Estimation and Inference of a GLM

Estimators of the regression coefficients for linear models with a normal response are obtained using least squares or ML, and significance tests are generally used to compare the sum of least squares under different hypothesis tests using the F -test. It is worth mentioning that these tests are exact, and, so, no approximations are required for their implementation.

GLMs offer a natural extension of this situation in the sense that: (1) The computational calculations used to determine the ML estimations of the regression parameters/coefficients are highly similar to those used in cases when the response is normal, with the difference being that the estimation process is iterative, which produces successive approximations that converge to the ML estimates. (2) In the inference procedures, the test statistic commonly used is the likelihood ratio test, which is parallel to the F -tests in linear models with a normal response. Thus, GLMs provide a uniform method of estimation and inference. Estimation of parameter β is highly similar to the ML method, whereas the inference methods are generally approximations since they are based on the theory of the distribution of a sufficiently large sample, as in the case of the likelihood ratio method. There are several alternative tests such as the Wald test, test scores, and the likelihood ratio test.

2.5 Specification of a GLM

In the following examples, we will describe the components of a GLM for some normal, gamma, binomial, and Poisson regression models.

2.5.1 Continuous Normal Response Variable

In simple linear regression models, the expected mean value of a continuous response variable depends on a set of explanatory variables, as follows:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2)$$

Equivalently,

$$E(y_i | x_i) = \beta_0 + \beta_1 x_i$$

This GLM can be expressed in terms of its three components:

$$\text{Distribution: } y_i \sim N(\mu_i, \sigma^2)$$

$$E(y_i) = \mu_i$$

$$\text{Var}(y_i) = \sigma^2$$

$$\text{Linear predictor: } \eta_i = \beta_0 + \beta_1 x_i$$

$$\text{Link function: } \eta_i = \mu_i \quad (\text{identity link})$$

where β_0 and β_1 are the intercept and slope, respectively. This means that we are expressing the linear model as a GLM.

Example 1 A simple linear regression analysis was performed on the diamond price (y) as a function of the number of carats (Table 2.2) and assuming that the response variable “ y ” has a normal distribution with a mean $\beta_0 + \beta_1 x_i$ and variance σ^2 .

The basic Statistical Analysis Software (SAS) syntax for simple linear regression is as follows:

```
proc reg ;
model price=weight/clb p r;
output out=diag p=pred r=resid;
id weight;
run;
```

In the above program, “proc reg” invokes a linear regression procedure in SAS. The “clb” option generates a confidence interval for the slope and intercept. The “p”

Table 2.2 Diamond price (dollars) based on weight (carats)

Weight	Price	Weight	Price	Weight	Price	Weight	Price	Weight	Price
0.17	355	0.18	462	0.18	468	0.17		0.25	655
0.16	328	0.28	823	0.16	345	0.32	918	0.35	1086
0.17		0.16	336	0.17	352	0.32	919	0.18	443
0.18		0.2	498	0.16	332	0.15	298	0.25	678
0.25	642	0.23	595	0.17	353	0.16	339	0.25	675
0.16	342	0.29	860	0.18	438	0.16	338	0.15	287
0.15	322	0.12	223	0.17	318	0.23	595	0.26	693
0.19	485	0.26	663	0.18	419	0.23	553	0.15	316
0.21	483	0.25		0.17	346	0.17	345	0.43	.
0.15		0.27		0.15		0.33	945		

Table 2.3 Regression analysis results

		Estimated parameters				
	Effect	Estimate	Standard error	Degree of freedom (DF)	t- value	Pr > t
$\hat{\beta}_0$	Intercept	-259.63	17.3189	46	-14.99	<0.0001
$\hat{\beta}_1$	Weight	3721.02	81.7859	46	45.50	<0.0001
$\hat{\sigma}^2$	Scale	1013.82	211.40			

option generates fitted values and standard errors. The “r” option performs a residual analysis (i.e., checks assumptions). The “output out” statement generates a new dataset called “diag” containing the residuals and the predicted/adjusted values. The “id weight” statement adds the specified variable to the fitted values output.

Part of the results is shown in Table 2.3. The estimated parameters, obtained from “proc reg,” are shown below:

Note that the estimated parameters are all statistically significantly different from zero. Then, the linear predictor takes the form:

$$\eta = - 259.63 + 3721.02 \times \text{weight},$$

If the response variable “y” does not fit the data well, then the normal distribution may barely represent the response distribution; that is, it would weakly explain the variability of the data and, consequently, the “identity” may not be the best link function, since the linear predictor would not include all the relevant information or some combination of the three components of the GLM. Although other fit measure statistics exist in the linear regression model, such as the coefficient of determination (R^2), the residual analysis is used to determine whether there is a good fit of the model or whether the assumptions of a Gaussian model are met. In this example, the value of R^2 is $R^2 = 0.9783$, and this value indicates that the model used explains 97.83% of the total variability of the dataset. In Fig. 2.1, we can see that the simple linear regression model provides a good fit to this dataset.

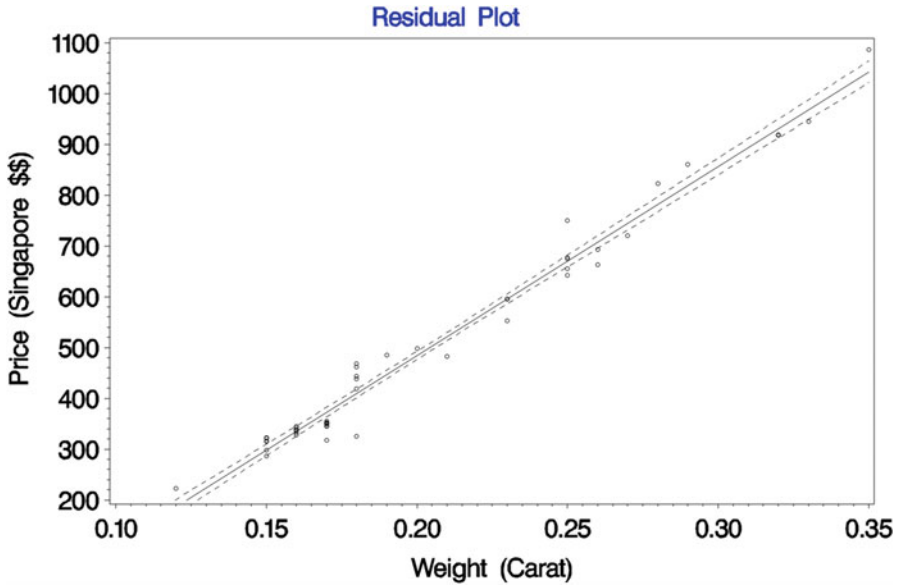


Fig. 2.1 A dot plot of price vs. weight (carat) and fitted model

2.5.2 Binary Logistic Regression

Logistic regression and other binomial response models are widely used in research areas like biological sciences and agriculture. Given their importance in this section, some relevant features of these models are mentioned.

Let y_i be the observed response on a set of p explanatory variables x_1, x_2, \dots, x_p whose distribution y_i is binomial with n_i independent Bernoulli trials and probability of success π_i on each trial, i.e.,

$$y_i \sim \text{Binomial}(n_i, \pi_i)$$

Then, we can model the response using a GLM with a binomial response. The linear predictor in this case will be equal to

$$\log\left(\frac{\pi_i}{1 - \pi_i}\right) = \beta_0 + \beta_1 x_{1i} + \dots + \beta_p x_{pi}$$

commonly known as “logit” because logit is defined as:

$$\text{logit}(\pi_i) = \log\left[\frac{\pi_i}{(1 - \pi_i)}\right]$$

which models the logarithm of the odds ratio, $\frac{\pi_i}{(1-\pi_i)}$, as a function of the predictor variables. The components of this GLM for binomial data are:

Distribution: $y_i \sim \text{Binomial}(n_i, \pi_i)$, with mean and variance

$$E(y_i) = n_i \pi_i \text{ and } \text{Var}(y_i) = n_i \pi_i (1 - \pi_i)$$

Linear predictor: $\eta_i = \beta_0 + \beta_1 x_{1i} + \dots + \beta_p x_{pi}$

Link function : $\eta_i = \text{logit}(\pi_i) = \log \left[\frac{\pi_i}{(1 - \pi_i)} \right]$ (logit link)

Another highly useful link function – when you have experiments – is the “probit” link $\eta_i = \Phi^{-1}(\pi_i)$, which was mentioned before.

The basic GLM for this dataset, under the probit link, is almost identical to the logit link as seen below:

Distribution: $y_i \sim \text{Binomial}(n_i, \pi_i)$

Linear predictor: $\eta_i = \beta_0 + \beta_1 x_i + \dots + \beta_p x_p$

Link function : $\eta_i = \text{probit}(\pi_i) = \Phi^{-1}[\pi_i]$.

Example 1 An engineer is interested in studying the effect of temperature (Temp) from 0 to 40 °C and time in days from 0 to 15 days on the germination of seeds of a certain crop. For this reason, he placed seeds in different pots containing moist soil. After a certain number of days, the number of germinated seeds was counted. If the seeds germinated, then $y = 1$; otherwise, $y = 0$. The probability of germination π_{ij} can be modeled through

$$\eta_{ij} = \beta_0 + \beta_1 \text{Day}_i + \beta_2 \text{Temp}_j$$

where η_{ij} is the linear predictor and β_0 , β_1 , and β_2 are the parameters to be estimated. In this GLM, the link function is

$$\eta_{ij} = \text{logit}(\pi_{ij}) = \log \left[\frac{\pi_{ij}}{(1 - \pi_{ij})} \right]$$

and the probability in the interval (0, 1) is computed through the inverse of the link function

$$\pi_{ij} = \frac{1}{1 + \exp^{-\eta_{ij}}} = g^{-1}(\eta_{ij})$$

This last expression allows to estimate the probability of germination (π_{ij}) under different temperature conditions ($^{\circ}\text{C}$) and time periods (days). Note that the nonlinear relationship between the result π_{ij} and the linear predictor η_{ij} is modeled by the inverse of the link function. In this particular case, the link function is the logit.

$$\eta_{ij} = \log \left[\frac{\pi_{ij}}{(1 - \pi_{ij})} \right] = g(\pi_{ij})$$

For the illustration of this example, a set of data was simulated using the values $\beta_0 = 8$, $\beta_1 = -0.19$, and $\beta_2 = -0.37$ in the linear predictor and the inverse of the linear function by varying the temperature from 0 to 40 $^{\circ}\text{C}$ and time from 0 to 15 days, i.e.,

$$\hat{\pi}_{ij} = \frac{1}{1 + e^{(8 - 0.19 \times \text{Temp}_i - 0.37 \times \text{Day}_j)}}$$

Part of the simulated data is shown below:

Temp	Days	Germ
0	0	0.000335
0	0.5	0.000403
0	1	0.000485
0	1.5	0.000584
0		0.000703
.	.	.
.	.	.
.	.	.
40	13	0.987991
40	13.5	0.989999
40	14	0.991674
40	14.5	0.99307
40	15	0.994234

The following commands allow us to perform a binomial regression using the “logit,” “probit,” and linear regression with the “identity” link. It is important to mention that we denote temperature as t and days as d in the codes used below.

Logit Regression

```
proc glimmix data=germ;
model p = t d/solution dist=binomial link=logit cl;
output out=logitout pred(noblp ilink)=predicted resid=residual;
run;
```

Probit Regression

```
proc glimmix data=germ;
model p = t d/solution dist=binomial link=probit cl;
output out=probitout pred(noblup ilink)=predicted resid=residual;
run;
```

Linear Regression (Identity)

```
proc glimmix data=germ ;
model p = t d/solution cl dist=normal;
output out=identity out pred(noblup ilink)=predicted resid=residual;
run;
```

“proc GLIMMIX” in SAS uses complex models without modifying the response variable as occurs when a direct transformation is applied to the response variable. Instead, GLIMMIX uses a link function of the response variable that is modeled as having a linear relationship with the explanatory variables. The “model” command specifies the response variable p as a function of the explanatory variables t and d , which define $X\beta$. The “solution” option in the model specification invokes the regression procedure to list the fixed effects parameter estimates of the model (β_0, β_1 , and β_2). The “dist” option is used to specify the distribution of the response variable, and the “link” option is used to specify the link function.

To get predicted probability values for each observation, the “output” option in proc GLIMMIX is used. Two types of predicted values can be obtained with the “output” option. The first type is the solution for the random effects (best linear unbiased predictors (BLUPs)) in the linearized model, and the second type is the predictions based on the fixed effects (best linear unbiased estimators (BLUEs)) (pred(noblup ilink) = predicted). The “ilink” sub-option in the “pred” option asks for the inverse function of the predicted values, that is, the probabilities of the predictions that are stored under the predicted file name. Finally, the “resid” option is used to request the residuals of the regression, which are stored in the residual.

Table 2.4 shows part of the output (analysis of variance (part (a)) and estimation and significance of fixed effects (part (b)) of the regression procedure using the logit link function.

Table 2.4 Estimation and significance of fixed effects using the logit link function

(a) Type III tests of fixed effects					
Effect	Num DF	Den DF	F-value	Pr > F	
T	1	2508	551.28	<0.0001	
D	1	2508	407.19	<0.0001	

(b) Parameters estimates					
Effect	Estimate	Standard error	DF	t-value	Pr > t
Intercept	-8.0000	0.3189	2508	-25.08	<0.0001
T	0.1900	0.008092	2508	23.48	<0.0001
D	0.3700	0.01834	2508	20.18	<0.0001

Table 2.5 Parameter estimates, linear predictor, and probability of linear, logit, and probit models

Link function	Parameter	Estimated value	$\hat{\eta}^*$	$\hat{\pi}^*$
Linear	$\hat{\beta}_0$	-0.417	1.149	0.873
	$\hat{\beta}_1$	0.022		
	$\hat{\beta}_2$	0.0412		
Logit	$\hat{\beta}_0$	-8.00	5.95	0.962
	$\hat{\beta}_1$	0.190		
	$\hat{\beta}_2$	0.370		
Probit	$\hat{\beta}_0$	-4.483	3.362	0.965
	$\hat{\beta}_1$	0.106		
	$\hat{\beta}_2$	0.207		

*The linear predictor $\hat{\eta}$ and the probability $\hat{\pi}$ were estimated using $D = 15$ and $T = 30$

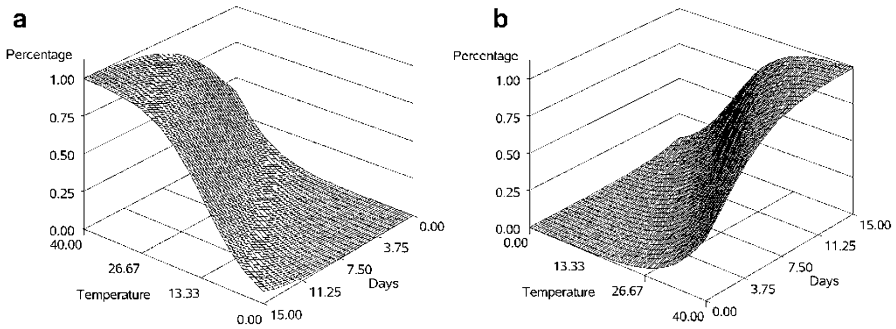


Fig. 2.2 (a, b) Probability of seed germination as a function of temperature and day

In Table 2.5, parameter estimates of the linear predictor for the generalized linear, logit, and probit models are presented. The probabilities estimated by the probit and logit models are almost identical to each other, but those of the linear probability model are different; this is because the data were generated with a binomial distribution, whereas the estimated linear predictor differs substantially from the linear predictor under the link probit and logit.

In Fig. 2.2a, b, we observe that in an interval between 3 and 7 days and 0 and 15 °C, there is approximately 20% seed germination, but, while both factors increase, the germination percentage also increases substantially.

2.5.2.1 Model Diagnosis

For a linear model, a plot of the predicted values against the residuals is probably the simplest way to decide whether the model used provides a good fit to the data; but, for a GLM, we must decide on the appropriate scale to use for the fitted values.

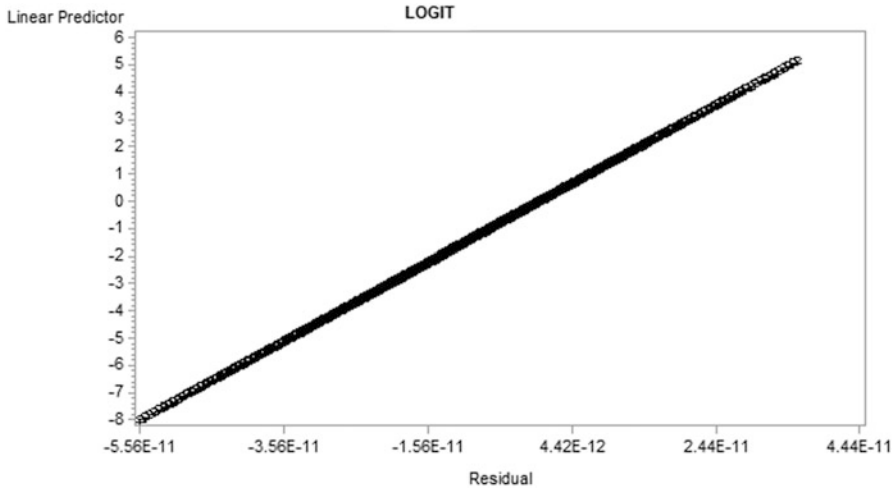


Fig. 2.3 Predicted vs. residual values using the logit link

Generally, it is better to use linear predictors $\hat{\eta}$ in the plot rather than the predicted responses $\hat{\mu}$. If there is no linear relationship between the linear predictors and the residuals, then it could indicate a lack of fit in the model. For a linear model, we could perform a transformation of the response variable, but this is not highly recommended for a GLM as this could change the response distribution. Another alternative would be to change the link function, but since there are not many link functions that allow interpreting a model easily, this is not a good option. Moreover, changing the linear predictor or transforming the predictor variables would not be the best way to go.

Figures 2.3, 2.4, and 2.5 show the linear predictor versus residual (we can also see the predicted value versus the residual). By investigating the nature of the relationship between the predictors and the residuals in Fig. 2.3, we can see that there is a linear relationship between the predictor and the residual, using the logit function, whereas the probit and identity functions do not show this linear relationship. However, with the probit link function, we observe a curvilinear relationship between the predictor and the residual, which may be because homogeneity of variance is not satisfied under this link function. Therefore, the logit link is shown to be the best choice.

Example 2 Fruit flies can be a year-round problem in fruit-growing areas in many regions of the world, such as in Mexico, and are most common especially in late summer and fall because ripe or fermented fruits and vegetables attract insects by serving as a natural host. If these insects are not controlled, economic losses in fruit-growing areas could be large and devastating to the producers. In response to this, entomologists have implemented experiments to help mitigate the damages caused by these insects. One such experiment attempted to establish the relationship between the concentration of a toxic agent (nicotine) for 5 hours and the number

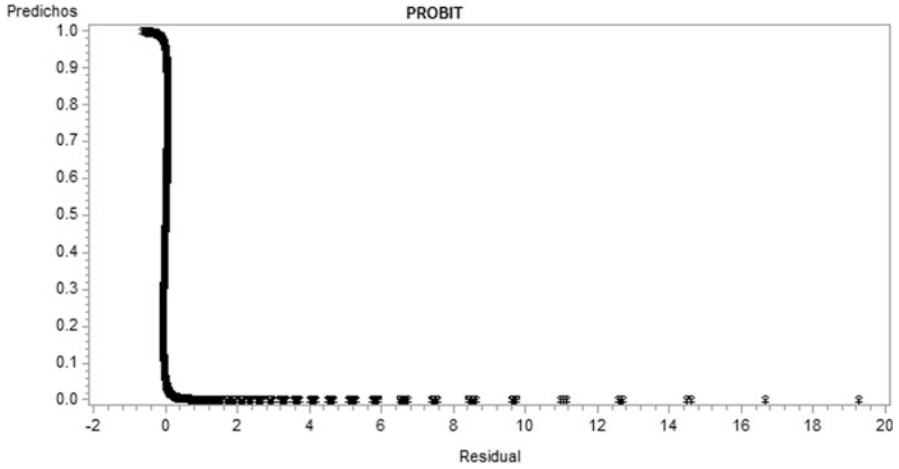


Fig. 2.4 Predicted values vs. residuals using the probit link

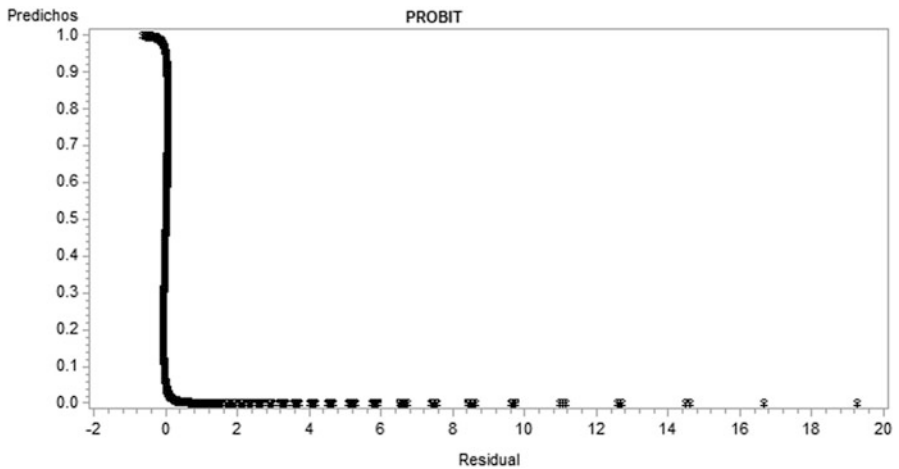


Fig. 2.5 Predicted vs. residual values using the identity link

of insects killed (common fruit fly); the data are shown in Table 2.6, and, for more information, see the study by Myers et al. (2002).

The number of dead insects can be modeled under a binomial distribution (n, π) . Let y_i denote the number of dead insects at a concentration i . The GLM components for this dataset are:

Table 2.6 Ratio of the concentration of a toxic agent to the number of fruit flies killed

Concentration (g/100 cc)	Number of insects (n)	Dead insects (y)	Proportion of dead insects
0.1	47	8	0.17
0.15	53	14	0.264
0.20	55	24	0.436
0.30	52	32	0.615
0.50	46	38	0.826
0.70	54	50	0.926
0.95	52	50	0.962

Distribution: $y_i \sim \text{Binomial}(n_i, \pi_i)$, with mean and variance
 $: E(y_i) = n_i \pi_i$ and $\text{Var}(y_i) = n_i \pi_i (1 - \pi_i)$

Linear predictor: $\eta_i = \beta_0 + \beta_1 \text{conc}_i$

Link function : $\eta_i = \text{logit}(\pi_i) = \log \left[\frac{\pi_i}{1 - \pi_i} \right]$ (logit link)

Note that we are using conc_i to denote the independent variable nicotine toxicant concentration. The following SAS code allows us to perform a binomial regression for the fruit fly dataset:

```
fly data;
input conc n y;
datalines;
0.1 47 8
0.15 53 14
0.2 55 24
0.3 52 32
0.5 46 38
0.7 54 50
0.95 52 50
;
proc glimmix data=nobound fly;
model y/n = conc/dist=binomial link=logit solution;
run;
```

The above syntax produces the following output:

The analysis of variance (Table 2.7 a) shows that there is a highly significant effect of nicotine concentration on the number of flies killed ($P = 0.0004$). From the results obtained, we can observe that, in part (b), the maximum likelihood estimator for the intercept and slope are $\hat{\beta}_0 = -1.7361$ and $\hat{\beta}_1 = 6.2954$, respectively, which are used to construct the linear predictor:

$$\hat{\eta}_i = -1.7361 + 6.2954 \times \text{conc}_i$$

Table 2.7 Results of the analysis of variance with the logit link

(a) Type III tests of fixed effects					
Effect	Num DF	Den DF	F-value	Pr > F	
Conc	1	5	71.94	0.0004	
(b) Parameter estimates					
Effect	Estimate	Standard error	DF	t-value	Pr > t
Intercept	-1.7361	0.2420	5	-7.17	0.0008
Conc	6.2954	0.7422	5	8.48	0.0004

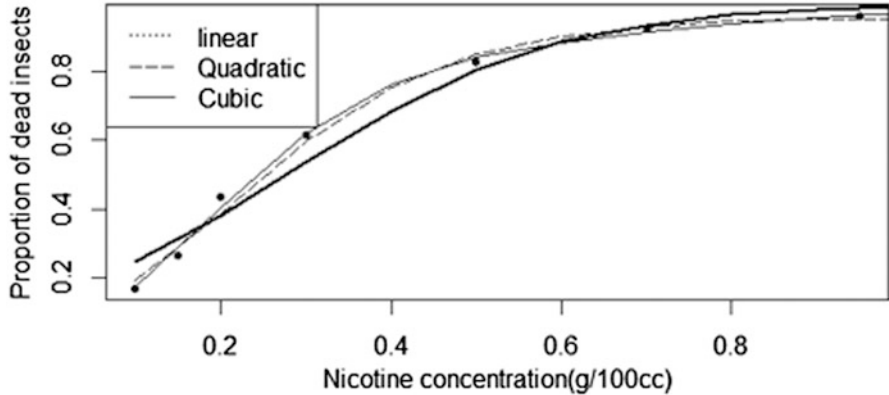


Fig. 2.6 Proportion of dead insects as a function of nicotine concentration

Therefore, with the logistic regression model, we can estimate the probability that an insect dies when exposed to a certain concentration i of nicotine using the following expression:

$$\hat{\pi}(\text{conc}_i) = \frac{e^{\hat{\eta}_i}}{1 + e^{\hat{\eta}_i}} = \frac{e^{-1.7361 + 6.2954 \times \text{conc}_i}}{1 + e^{-1.7361 + 6.2954 \times \text{conc}_i}}$$

A plot of the mean proportion of dead insects exposed to a certain concentration of nicotine and the regression curve (linear, quadratic, and cubic) is shown in Fig. 2.6. In this figure, we observe that as the nicotine concentration increases, the mean proportion of dead insects increases. The best linear predictor is of a quadratic order.

2.5.3 Poisson Regression

Often, the outcome of a variable is numerical in the form of counts. Sometimes it is a count of rare events such as, for example, (1) the number of plants infected by a

certain disease in a population over a period of time, (2) the number of insects surviving after the application of an insecticide over time, (3) the number of dead fish found per cubic kilometer due to a certain pollutant, (4) the number of sick animals occurring in a given month in a given country, and so on. The Poisson probability distribution is perhaps the most widely used for modeling count-type response variables. As λ (the average count) increases, the Poisson distribution grows symmetrically and eventually approaches a normal distribution.

The Poisson likelihood function is appropriate for nonnegative integer data and this process assumes that events occur randomly over time, so the following conditions must be met:

- (a) The probability of at least one occurrence of an event in a given time interval is proportional to the length of the interval.
- (b) The probability of two or more occurrences of an event within an extremely small interval is negligible.
- (c) The number of occurrences of an event in disjoint time intervals are mutually independent.

The probability distribution of a Poisson random variable "y," which represents the number of successes occurring in a given time interval or in a given region of space, is given by the expression

$$P(y = k) = \frac{e^{-\lambda} \lambda^k}{k!}, \quad \lambda > 0, \quad k = 1, 2, \dots$$

where λ is the average number of successes (the average count) in a time or space interval. The mean and variance of this distribution are the same, that is,

$$E(y) = \text{Var}(y) = \lambda$$

Poisson regression belongs to a GLM and is appropriate for analyzing count data or contingency tables. A Poisson regression assumes that the response variable "y" has a Poisson distribution and that the logarithm of its expected value can be modeled by a linear combination of unknown parameters and independent variables. As in a standard linear regression, the predictors, weighted by the coefficients of x_1, x_2, \dots, x_p , are summed to form the linear predictor,

$$\eta_i = \beta_0 + \sum_{p=1}^P x_{pi} \beta_p$$

where β_0 is the intercept and β_p is the slope of the covariates x_p ($p = 1, \dots, P$). Thus, the expected value of y_i and the linear predictor η_i are related through the link function. The components of a GLM with a Poisson response ($y_i \sim \text{Poisson}(\lambda_i)$), where λ_i is the expected value of y_i , are as follows:

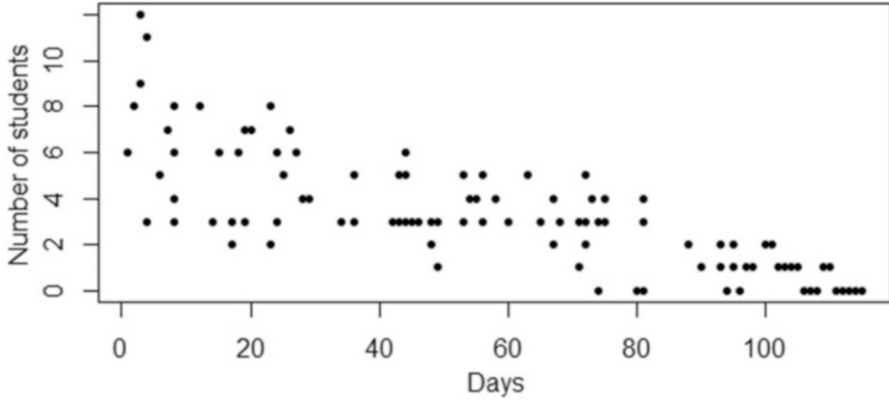


Fig. 2.7 Students infected with the disease

Distribution: $y_i \sim \text{Poisson}(\lambda_i)$, with $E(y_i) = \text{Var}(y_i) = \lambda_i$

Linear predictor: $\eta_i = \beta_0 + \beta_1 x_{1i} + \dots + \beta_p x_{pi}$

Link function: $\eta_i = \log(\lambda_i) = g(\lambda_i)$ (log link)

Example 1 The following dataset corresponds to the number of students diagnosed (Fig. 2.7) with a certain infectious disease within a period of days of an initial outbreak. We will fit a generalized linear model for “count” data assuming a Poisson distribution.

Note that the response distribution is skewed to the right and that the responses are positive integers. Since the response variable is count, the initial choice of a Poisson distribution is reasonable for this dataset with its canonical link, the natural logarithm. The number of “days elapsed” after the initial disease outbreak is the predictor variable in the systematic component. Thus, the GLM for this dataset (Appendix: Data: Infected students) is:

Distribution: Infected students_{*i*} $\sim \text{Poisson}(\lambda_i)$

Linear predictor: $\eta_i = \beta_0 + \beta_1 \text{Days}$

Link function: $\eta_i = \log(\lambda_i)$ (log link)

Part of the data is shown below:

Days elapsed	Infected students
1	6
2	8
3	12
.	.

(continued)

Days elapsed	Infected students
.	.
.	.
109	1
110	1
112	0

For the purposes of implementation, we use days to denote elapsed days and students to denote infected students. We can employ the Poisson regression model using GLIMMIX in SAS, as shown below:

```
proc glimmix data=students method=laplace;
model students=days/solution dist=poisson link=log;
output out=sal_infection pred(noblup ilink)=predicted
resid=residual;
run;
```

The “proc GLIMMIX” statement invokes the SAS generalized linear mixed model (GLMM) procedure. The “model” command specifies the response variable and the predictor variable, whereas the “solution” option in the model specification requests a listing of the fixed effects parameter estimates. The “dist = poisson” option specifies the distribution of the data, and the “link = log” option declares the link function to be used in the model. The default estimation technique in generalized linear mixed models is restricted pseudo-likelihood (the “RPSL method”); in this example, we use “method = laplace.” The “output” option creates a dataset containing predicted values and diagnostic residuals, calculated after fitting the model. By default, all variables in the original dataset are included in the output dataset, whereas the “out = sal_infection” statement specifies the name of the output dataset. The “pre(noblup ilink) = predicted” option calculates the predicted values without taking into account the random effects of the model, and “ilink” calculates the statistics and predicted values at the scale of the data. Finally, the “resid = residual option” calculates the residuals.

The probability estimation of a GLMM involves an integral, which, in general, cannot be calculated explicitly. “GLIMMIX,” by default, uses the RSPL method, but it also offers different options such as the quadrature and Laplace integration method, among others. These integral approximation methods approximate the probability function of an GLMM, and the optimization of the function is numerically approximated. These methods provide a real objective function for optimization. For more details, see the SAS manual. However, in a GLM, this approximation involving the integral is not necessary since an exact solution can be obtained to estimate the parameters, as there are no random effects. The results of this analysis are shown below (Table 2.8).

The fit statistics in part (a) (“Fit statistics”) give us an idea of the quality of the goodness of the fit of the model; these statistics are very useful when we are proposing different models to try and find the best model for the data. In this case, the value of the generalized chi-squared statistic divided over its degrees of freedom

Table 2.8 Results of the analysis of variance

(a) Fit statistics (Akaike’s information criterion (AIC), a small sample bias corrected Akaike’s information criterion (AICC), Bozdogan Akaike’s information criterion (CAIC), Schwarz’s Bayesian information criterion (BIC), Hannan and Quinn information criterion (HQIC))					
–2 Log likelihood					389.11
AIC (smaller is better)					393.11
AICC (smaller is better)					393.22
BIC (smaller is better)					398.49
CAIC (smaller is better)					400.49
HQIC (smaller is better)					395.29
Pearson’s chi-square					84.95
Pearson’s chi-square / DF					0.78
(b) Type III tests of fixed effects					
Effect	Num DF	Den DF	F-value	Pr > F	
Days	1		102.28	<0.0001	
(c) Parameter estimates					
Effect	Estimate	Standard error	DF	t-value	Pr > t
Intercept	1.9902	0.08394		23.71	<0.0001
Days	– 0.01746	0.001727		– 10.11	<0.0001

is close to 1. This indicates that the variability of these data has been reasonably modeled and that there is no residual overdispersion. The value of the generalized chi-squared statistic divided over its degrees of freedom (Pearson’s chi – square/DF) is the experimental error of the analysis.

The “Type III tests of fixed effects” (in part (b)) and the solution for the intercept and the days effect (“Parameter estimates”) in part (c) are shown in Table 2.8. The negative coefficient of the covariate days indicates that as the number of days increases, the average number of students diagnosed with the disease decreases.

That is, we reject the null hypothesis ($P = 0.0001$) that the expected number of infected students is the same as the number of days increases.

We see that with a 1-day increase in the infection period, the expected (or average) number of students diagnosed with the disease decreases by a factor of $e^{-0.01746} = 0.9827$.

The estimated linear predictor for this GLM is:

$$\hat{\eta}_i = 1.9902 - 0.01746 \times \text{Days}$$

For example, we can calculate the probability of diagnosing “ $k = 2$ ” infected students in a period of 2 days; i.e., “Days = 2” as follows:

$$\hat{P}(Y_i = k) = \frac{\exp(-\hat{\lambda}_i) (\hat{\lambda}_i)^k}{k!}$$

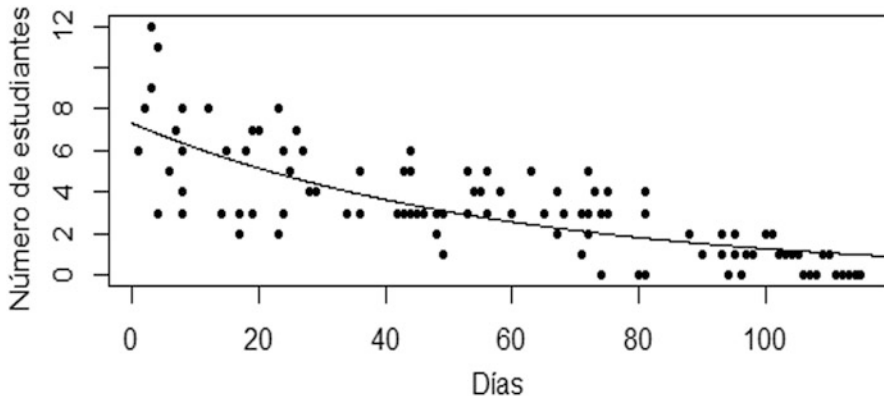


Fig. 2.8 Infected students and a Poisson regression fit

$$\begin{aligned} \hat{P}(Y_i = 2) &= \frac{\exp(-\exp[1.9902 - 0.0146 \times 2]) (\exp[1.9902 - 0.0146 \times 2])^2}{2} \\ &= \frac{\exp(-\exp(1.961)) (\exp(1.961))^2}{2} = 0.0207 \end{aligned}$$

This value indicates that the probability of observing/diagnosing two students with the disease in a 2-day period is 0.0207 (2.0701%).

In Fig. 2.8, we observe that the Poisson model is a good candidate for modeling this dataset, since there is no overdispersion in this regression model.

Example 2 A forest engineer is interested in modeling the number of trees recently infected by a certain virus. The data that he has are age (years), height (meters) of the trees, and the number of infected trees. Using a linear model could result in negative values of the parameter λ , which would not make sense. The link function $g(\lambda)$ for a Poisson error structure is the logarithm. Therefore, the GLM, defining $y_i =$ infected trees_{*i*}, can be as follows:

$$\begin{aligned} \text{Distribution: } &y_i \sim \text{Poisson}(\lambda_i) \\ \text{Linear predictor: } &\eta_i = \beta_0 + \beta_1 \times \text{Age}_i + \beta_2 \times \text{height}_i \\ \text{Link function: } &\eta_i = \log(\lambda_i) = g(\lambda_i) \quad (\log \text{ link}) \end{aligned}$$

For this example, a dataset was simulated using the following parameter values: $\beta_0 = -2$, $\beta_1 = -0.03$, and $\beta_2 = -0.04$. In addition, in order to obtain the linear predictor, the variable age (years) varied from 0 to 50 and height (meters) from 0 to 30, both with increments in one unit. Thus, the values of y_{ij} were simulated with the following expression:

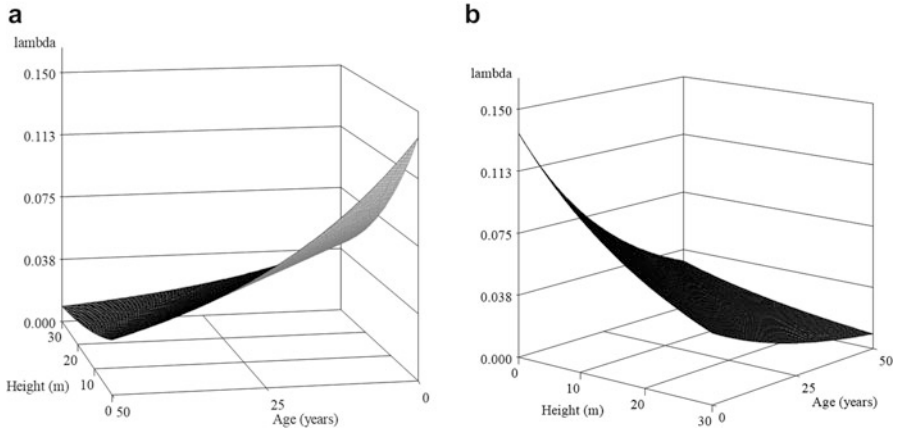


Fig. 2.9 (a, b) Probability of tree infection as a function of tree height and age in years

$$y_{ij} = \exp(-2 - 0.03 \times \text{Age}_i - 0.04 \times \text{Height}_j)$$

In Fig 2.9a, b, we can see that at a young age, between 1 and 10 years and at a height of no more than 10 meters, trees are more susceptible to be infested by the virus. However, as their age increases, trees show greater resistance.

The following SAS code fits a Poisson regression model with two predictor variables, assuming that there is no interaction between the two explanatory variables.

```
proc glimmix data=infection method=laplace;
model infection=age height /solution dist=poisson link=log;
output out=sal_infection pred(noblup ilink)=predicted
resid=residual;
run;
```

In Table 2.9 part (a), the analysis of variance shows that age and tree height are highly significant, indicating that both variables help explain the infection mechanism of the trees through a Poisson model ($P < 0.0001$).

The linear predictor for this GLM, with Poisson distribution, in the response variable is:

$$\eta_{ij} = -2 - 0.03 \times \text{Age}_i - 0.04 \times \text{Height}_j$$

The estimated values of the parameters of each of the explanatory variables indicate that as age (years) and height (meters) increase by one unit, the tree is less susceptible to the virus. If we want to calculate the probability of diagnosing “ $k = 3$ ” infected trees with the virus when they are 2 years old and 3-meters tall, we can use the following equation:

Table 2.9 Part of the results of the analysis of variance under a Poisson distribution

(a) Type III tests of fixed effects					
Effect	Num DF	Den DF	F-value	Pr > F	
Age	1	6158	43.20	<0.0001	
Height	1	6158	29.10	<0.0001	
(b) Parameter estimates					
Effect	Estimate	Standard error	DF	t-value	Pr > t
Intercept	-2.0000	0.1388	6158	-14.41	<0.0001
Age	-0.03000	0.004564	6158	-6.57	<0.0001
Height	-0.04000	0.007415	6158	-5.39	<0.0001

$$\hat{P}(Y_i = k) = \frac{\exp(-\hat{\lambda}_i) (\hat{\lambda}_i)^k}{k!}$$

$$\hat{P}(Y_i = 3)$$

$$= \frac{\exp(-\exp[-2 - 0.03 \times \text{Age} - 0.04 \times \text{Height}]) (\exp[-2 - 0.03 \times \text{Age} - 0.04 \times \text{Height}])^3}{3!}$$

$$= \frac{\exp(-\exp[-2 - 0.03 \times 2 - 0.04 \times 3]) (\exp[-2 - 0.03 \times 2 - 0.04 \times 3])^3}{3!} = 0.000215$$

This value indicates that the probability of observing/diagnosing three trees with the virus causing the disease when they are 2 years old and 3-meters tall is 0.000215 (0.0215%).

A Poisson regression model, sometimes referred to as a log-linear model, is especially useful when it is used in contingency table modeling. Log-linear models are models of associations between variables in a contingency table; they treat variables symmetrically and do not distinguish one variable as a response. They have a formal structure of double or more entries that can be fitted by binomial or Poisson regression. These models for contingency tables have several specific applications in biological and social sciences.

Variables can be nominal or ordinal. A nominal variable has no natural order; for example, gender (male, female, transgender), eye color (blue, brown, green), and type of pet (cat, bird, fish, dog, mouse). An ordinal variable has a range of orders; for example, when you want to measure the degree of consumer satisfaction with the consumption of a product (very dissatisfied, somewhat dissatisfied, neither satisfied nor dissatisfied, somewhat satisfied, very satisfied).

2.5.4 Gamma Regression

A gamma distribution is a distribution that occurs naturally in processes for which waiting times, between events, are relevant. Lifetime data are sometimes modeled with a gamma distribution. This distribution can take a wide range of forms due to the relationship between the mean and variance across its two parameters (α and β) and is suitable for dealing with heteroscedasticity of nonnegative data. The probability of observing a particular value y , given the parameters α and β , is

$$f(y) = \frac{1}{\Gamma(\alpha)\beta^\alpha} y^{\alpha-1} e^{-(y/\beta)}; y, \alpha, \beta > 0$$

where $\Gamma(\cdot)$ is the gamma function. A gamma regression uses the input variables X 's and coefficients to make a prediction about the mean of "y," but it actually focuses more of its attention on the scale parameter β . The mean and variance of a Gamma random variable are:

$$E(Y) = \alpha\beta = \mu \quad \text{and} \quad \text{Var}(Y) = \alpha\beta^2 = \mu^2/\alpha$$

The probability density function gamma can be rewritten in terms of the mean μ and the scale parameter α as follows:

$$f(y) = \frac{1}{\Gamma(\alpha)y} \left(\frac{y\alpha}{\mu}\right)^\alpha \exp\left(-\frac{y\alpha}{\mu}\right), \quad y > 0$$

Plotting the gamma distribution (Fig. 2.10) with three different values of shape $\alpha = (0.75, 1, \text{ and } 2)$, the scale parameter μ has a multiplicative effect. In the gamma density of the first panel $\alpha = 0.75$, we see that the density is infinite at 0, whereas in the second panel $\alpha = 1$, it corresponds to the exponential density, and, in the third panel $\alpha = 2$, we see a skewed distribution.

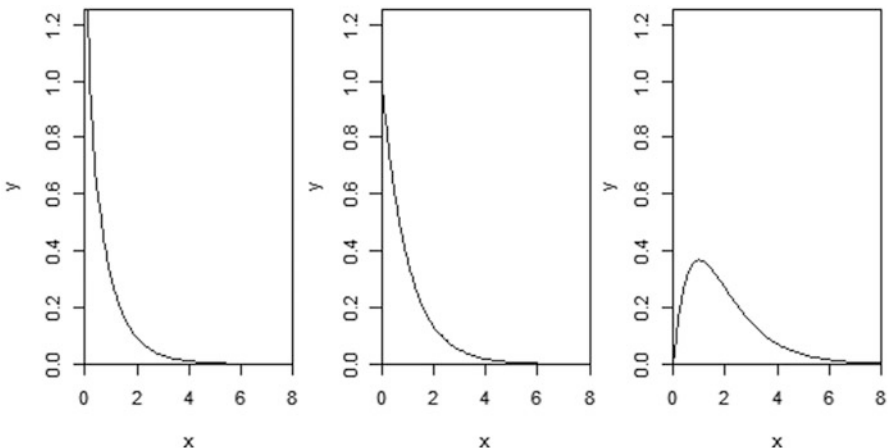


Fig. 2.10 Gamma density: from left to right, $\alpha = 0.75, 1, \text{ and } 2$

A gamma distribution can arise in different forms. The sum of " n " independent and identically distributed exponential random variables with parameter β has a gamma distribution (n, β) . The chi-squared distribution χ^2 is a special case of a gamma distribution with $\beta = 1/2$ and $\alpha/2$ degrees of freedom.

Theoretically, a Gamma distribution should be the best choice when the response variable has a real value in the range of zero to infinity and it is appropriate when a fixed relationship between the mean and variance is suspected. If we expect the values " y " to be small, then we should expect a small amount of variability in the observed values. Conversely, if we expect large values of " y ," then we should expect (observe) a lot of variability.

Models with a gamma distribution with multiplicative covariate effects provide additional support for modeling nonnegative right-skewed continuous responses, such as the gamma variable with the log link function. Whether the data are modeled with an inverse or logarithmic link function will depend on whether the rate of change or the logarithm of the rate of change is a more meaningful measure. For example, in studies of yield density that commonly assume that yield per plant is inversely proportional to plant density (Shinozaki and Kira 1956), the linear predictor is:

$$\eta_i = (\beta_0 + \beta_1 x_i)^{-1}$$

Example 1 In the development of coagulation agents, it is common to perform in vitro clotting time studies. The following data were reported by McCullagh and Nelder (1989). Plasma samples from healthy men were diluted to nine different percentages of prothrombin-free plasma concentration; the greater the dilution, the more interference with the ability of the blood to clot because the natural clotting ability of the blood has been weakened. For each sample, clotting was induced by introducing thromboplastin, a clotting agent, and the time until clotting occurred (in seconds) was recorded. Five samples were measured at each of the nine concentration percentages, and the mean clotting times were averaged; therefore, the response is the mean clotting time across the five samples. In Fig. 2.11, the response variable is plotted against the percentage thromboplastin concentration in which we observe that the longer clotting times tend to be more variable than the smaller clotting times, so a linear regression model may not be appropriate.

In this analysis, we will model clotting times as the response variable (y_i) with plasma concentration percentage as the predictor variable. Conc denotes the independent variable concentration. The GLM for this dataset is:

$$\text{Distribution: } y_i = \text{Clotting time}_i \sim \text{Gamma}(\alpha, \beta)$$

$$\text{Linear predictor: } \eta_i = \beta_0 + \beta_1 \times \text{conc}_i$$

$$\text{Link function : } \mu_i = \frac{1}{\beta_0 + \beta_1 \times \text{conc}_i} \text{ (inverse link)}$$

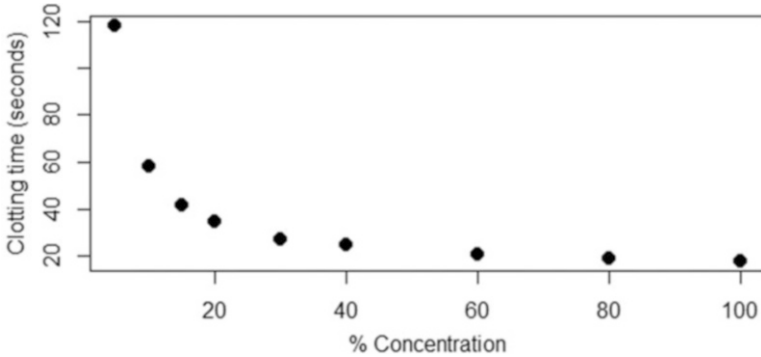


Fig. 2.11 Clotting time (seconds), depending on the thromboplastin concentration

The following syntax allows us to adjust a GLM with gamma errors in GLIMMIX:

```
data coagu;
input num conc y;
datalines;
1 5 118
2 10 58
3 15 42
4 20 35
5 30 27
6 40 25
7 60 21
8 80 19
9 100 18
;
proc glimmix data = coagu;
model y = conc / dist=gamma link=power(-1) solution;
output out=salgamm1 pred(noblup ilink)=predicted resid=residual;
run;
```

Most of the syntax has already been described in the previous examples; the only new one is the **link = power(-1)** option. This statement invokes the inverse link function in the GLIMMIX procedure.

Some of the output from this analysis is shown in Table 2.10.

The dilution percentage, part (a) in Table 2.10, of the blood plasma concentration significantly affects the clotting time ($P = 0.0004$). The values for constructing the fitted linear predictor are tabulated in part (b) of Table 2.10.

$$\hat{\eta}_i = 0.008686 + 0.000658 \times \text{conc}_i$$

Table 2.10 Results of the regression analysis under a gamma distribution

(a) Type III tests of fixed effects					
Effect	Num DF	Den DF	F-value	Pr > F	
Conc	1		41.01	0.0004	

(b) Parameter estimates					
Effect	Estimate	Standard error	DF	t-value	Pr > t
Intercept	0.008686	0.002294		3.79	0.0068
Conc	0.000658	0.000103		6.40	0.0004
Scale	0.05213	0.02436	.	.	.

With the parameterization of the gamma distribution, previously chosen, the intercept and the beta coefficient corresponding to the concentration variable were calculated through GLIMMIX in SAS, as well as the scale parameter(α), which in the SAS output corresponds to the scale. With part of this information, it is possible to calculate the mean ($E[Y] = \mu$) and variance ($\text{Var}[Y] = \mu^2/\alpha$) for a concentration $\text{conc} = 10$ as follows:

$$\hat{y} = \hat{\mu} = \frac{1}{0.008686 + 0.000658 \times \text{conc}} = \frac{1}{0.008686 + 0.000658 \times 10} = 65.505$$

$$\widehat{\text{Var}}(y) = \frac{\hat{\mu}^2}{\alpha} = \frac{65.505^2}{0.052} = 85818.215$$

The average time it takes for blood to clot – when a thromboplastin concentration of 10% is added – is 65.505 seconds with a variance of 85818.215.

2.5.4.1 Model Selection

Selecting a model from a set of candidate models that provides the best fit and largely explains the variability in the data is a necessary but complex task. This process involves trying to minimize information loss. From the field of information theory, several information criteria have been proposed to quantify information, or the expected value of information, and, among these, the most widely used are the Akaike information criterion (AIC) (Akaike 1973, 1974) and the Bayesian information criterion (BIC) (Schwarz 1978). Both AIC and BIC are based on the ML estimates of the model parameters. In a regression fit, the estimates of $\hat{\beta}$'s under the ordinary least method and the ML method are identical. The difference between the two methods comes from estimating the common variance σ^2 of the normal distribution of the errors, around the true mean.

Table 2.11 Goodness-of-fit metrics for each of the three models and regression analysis results for model 3

(a) Fit statistics					
	Model 1	Model 2	Model 3		
-2 Log likelihood	62.15	44.49	27.47		
AIC (smaller is better)	68.15	52.49	37.47		
AICC (smaller is better)	72.95	62.49	57.47		
BIC (smaller is better)	68.74	53.27	38.45		
CAIC (smaller is better)	71.74	57.27	43.45		
HQIC (smaller is better)	66.87	50.78	35.34		
Pearson's chi-square	0.50	0.07	0.01		
Pearson's chi-square / DF	0.07	0.01	0.001		
(b) Type III tests of fixed effects					
Effect	Num DF	Den DF	F-value	Pr > F	
Conc	1	5	476.73	<0.0001	
Conc × conc	1	5	110.78	0.0001	
conc × conc × conc	1	5	50.92	0.0008	
(c) Parameter estimates					
Effect	Estimate	Standard error	DF	t-value	Pr > t
Intercept	-0.00040	0.000576	5	-0.70	0.5177
Conc	0.001946	0.000089	5	21.83	<0.0001
conc × conc	-0.00003	2.576E-6	5	-10.53	0.0001
conc × conc × conc	1.337E-7	2.520e-08	5	5.306	<0.0001
Scale	0.001125	0.000530	.	.	.

To get an idea of how to use these adjustment statistics, let us compare three possible models that best explain the effect of the plasma dilution percentage:

$$\text{Model 1: } \eta_i = \beta_0 + \beta_1 \times \text{conc}_i$$

$$\text{Model 2: } \eta_i = \beta_0 + \beta_1 \times \text{conc}_i + \beta_2 \times \text{conc}_i^2$$

$$\text{Model 3: } \eta_i = \beta_0 + \beta_1 \times \text{conc}_i + \beta_2 \times \text{conc}_i^2 + \beta_3 \times \text{conc}_i^3$$

Since the proposed models have a gamma error structure, the commonly used fit statistic (R^2) in a simple linear regression model is not reported. Part of the results of this analysis is shown below with various metrics as goodness-of-fit measures:

With regard to the values of the goodness-of-fit metrics (Table 2.11 part (a)), the smaller they are, the better the fit. Based on the above, the accuracy of the fit of the three regression models increased as the polynomial in the linear predictor increased. That is, model three best explained the variability of the plasma clotting time. The

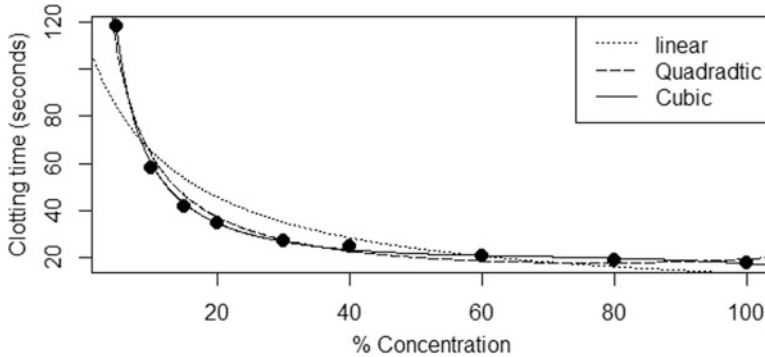


Fig. 2.12 Fitting the gamma regression model with three predictors

type III sum of squares for fixed effects and the estimated parameters under model three are tabulated in parts (b) and (c) in Table 2.11, respectively.

Parameter estimates under the linear predictor with linear, quadratic, and cubic effects are highly significant. The results suggest that a cubic effect for the percentage dilution in plasma concentration in the linear predictor is more efficient in explaining the clotting time than taking only a linear predictor with only linear or both linear and quadratic effects (Fig. 2.12).

2.5.5 *Beta Regression*

Studies in various areas of knowledge, including agriculture, often face the need to explain a variable expressed as a proportion, percentage, rate, or fraction in the continuous range (0,1). In economics, for example, the factors that influence the proportion of households that do not have a cement floor have been studied. Similarly, in plant breeding, it is desired to investigate the factors that influence the proportion of plant leaves damaged by a certain disease. In parallel, the proportion of impurities in chemical compounds is of everyday interest in physics and chemistry. While studies on electoral preferences analyze citizen participation rates and the variables that can explain them, in the field of education and academic performance, we try to explain the proportion of success in standardized tests. Moreover, it is also used to identify the factors associated with the proportion of credit used by credit card users. The public health field has also been confronted with the need to model the proportion of coverage in health programs in order to identify the sociodemographic and economic characteristics associated with whether a woman is covered. Johnson et al. (1995) presented the properties of the probability distribution of this type of variable; these researchers showed that the beta distribution can be used to model proportions, since its density can take different forms depending on the values of the two shape parameters that index the distribution. However, the beta regression that results from using the beta distribution as a

response variable in the context of generalized linear models is not very well known, but its use is increasing every day, thanks to friendly software that allow its implementation in an extremely easy manner.

Definition Let y be a continuous random variable defined in the interval $[0, 1]$ and $\alpha, \beta > 0$. Then, Y has a beta distribution with parameters of forms α and β if and only if:

$$f_Y(y) = \frac{1}{B(\alpha, \beta)} y^{\alpha-1} (1-y)^{\beta-1}, \quad 0 < y < 1$$

where $B(\alpha, \beta)$ is the beta function defined as $B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$ and Γ is the gamma function. The mean and variance of this probability density function are given by

$$E(Y) = \frac{\alpha}{\alpha + \beta} \quad \text{and} \quad \text{Var}(Y) = \frac{\alpha\beta}{(\alpha + \beta + 1)(\alpha + \beta)^2}.$$

In the context of regression analysis, the density of the beta distribution provided above is not very useful for modeling the mean of the response. Therefore, this density is reparametrized so that it contains a precision (or dispersion) parameter. This reparameterization consists of defining a $\mu = \frac{\alpha}{\alpha+\beta}$ and $\phi = \alpha + \beta$, i.e., $\alpha = \mu\phi$ and $\beta = (1 - \mu)\phi$, which means that:

$$E(y) = \mu$$

and

$$\text{Var}(y) = \frac{\mu(1-\mu)}{1 + \phi}$$

So, μ is the mean of the response variable and ϕ can be interpreted as a parameter of precision in the sense that, for a fixed μ , the higher the value of ϕ , the smaller the variance of y . The density function of y can be written as:

$$f(y; \mu, \phi) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)} y^{\mu\phi-1} (1-y)^{(1-\mu)\phi-1}, \quad 0 < y < 1$$

where $0 < \mu < 1$ and $\phi > 0$.

Let y_1, y_2, \dots, y_n be independent and identically distributed random variables, where each y_i with $i = 1, 2, \dots, n$ is modeled under the parametrized beta model with a mean μ and an unknown parameter ϕ . The model is obtained by assuming that the mean of y_i can be written as:

Table 2.12 Proportion of fruit damage (y) as a function of concentration. Percentage is equal to proportion $\times 100$

Concentration	Y
0.1	0.08
0.25	0.09
0.5	0.11
1	0.2
2	0.3
4	0.53
5	0.63
8	0.71
10	0.73
25	0.84
50	0.85
100	0.86

$$g(\mu_i) = \sum_{i=1}^k x_{ij}\beta_i = \eta_i$$

where $\beta_1, \beta_2, \dots, \beta_k$ are unknown regression parameters and x_{ij} are the k covariates ($k < n$) that are fixed and known. Finally, $g(\cdot)$ is a strictly monotone and differentiable link function that maps to the real numbers in the interval $(0, 1)$.

There are several possible options for the link function $g(\cdot)$. For example, we can use a logit link function $g(\mu) = \log\left(\frac{\mu}{1-\mu}\right)$, which is considered the most popular and asymptotically efficient, but it is also feasible to use the probit $g(\mu) = \Phi^{-1}(\mu)$ function, where $\Phi(\cdot)$ is the cumulative distribution function of a standard normal random variable, and the complementary link function $g(\mu) = \log\{-\log(1-\mu)\}$, among others (McCullagh and Nelder 1989).

Example 1 The objective of this experiment was to evaluate the effect of the concentration of a chemical compound on the proportion of damage (y) in the fruits (Table 2.12). This compound is known to inhibit the growth of an insect, but, at a certain concentration, it can cause damage to the fruits.

The proportion of damage to the fruits can be modeled under a beta distribution (μ, ϕ) . Let y_i be the proportion of damage to the fruits due to the i th concentration. The GLM components for this dataset are as follows:

$$\text{Distribution: } y_i \sim \text{beta}(\mu_i, \phi), \text{ with } E(y) = \mu \quad \text{and} \quad \text{Var}(y) = \frac{\mu(1-\mu)}{1+\phi}$$

$$\text{Linear predictor: } \eta_i = \beta_0 + \beta_1 \times \text{conc}_i$$

$$\text{Link function : } \eta_i = \text{logit}(\pi_i) = \log\left[\frac{\pi_i}{(1-\pi_i)}\right] \text{ (logit log)}$$

Table 2.13 Goodness-of-fit metrics for the linear and quadratic models and results of the quadratic model fit

(a) Fit statistics		Linear	Quadratic		
-2 Log likelihood		-6.23	-14.50		
AIC (smaller is better)		-0.23	-6.50		
AICC (smaller is better)		2.77	-0.79		
BIC (smaller is better)		1.22	-4.56		
CAIC (smaller is better)		4.22	-0.56		
HQIC (smaller is better)		-0.77	-7.22		
Pearson's chi-square		12.85	15.05		
Pearson's chi-square / DF		1.07	1.25		
(b) Type III tests of fixed effects					
Effect	Num DF	Den DF	F-value	Pr > F	
Conc	1		8.08	0.0193	
Conc × conc	1		6.11	0.0354	
(c) Parameter estimates					
Effect	Estimate	Standard error	DF	t-value	Pr > t
Intercept	-1.1425	0.2935		-3.89	0.0037
Conc	0.1572	0.05530		2.84	0.0193
Conc × conc	-0.00132	0.000534		-2.47	0.0354
Scale	9.0432	4.0045	.	.	.

Note that we are using *conc* to denote the independent variable concentration of the chemical compound. The following SAS code allows us to perform a beta regression for the dataset:

```
proc glimmix method=laplace;
model y = conc / dist=beta s;
run;
```

The “method = Laplace” statement asks SAS for the estimation method to be Laplace integration, and the “dist = beta” and “s” options invoke GLIMMIX to perform beta regression and provide fixed parameter estimation, respectively.

In order to see which type of linear, quadratic, or cubic predictor best explains the observed variability in a dataset, we make use of the fit statistics (-2 log likelihood, AIC, etc.). Part of the output is shown below in Table 2.13. According to the fit statistics in part (a), the predictor that best models this experiment is the quadratic predictor.

In Fig. 2.13, we can see that the best linear predictor to model a dataset is of the cubic order, but due to the indeterminacy (not showing here) in the *t*-value (infinity), in the hypothesis test of the estimated parameters, it was decided to take the quadratic predictor. Both predictors, quadratic and cubic, better model the proportion (percentage = proportion×100) of fruit damage caused by the concentration of the applied chemical.

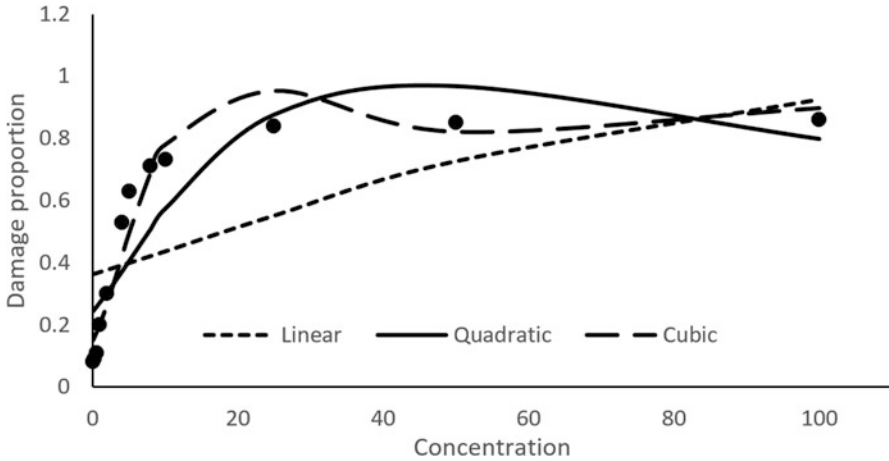


Fig. 2.13 Fitting the beta regression model

2.6 Exercises

Exercise 2.6.1 The partial dataset corresponds to an evaluation of the effects of increasing application rates of picloram (0, 1.1, 2.2, and 4.5 kg/ha) for the control of larkspur plants (data in Table 2.14). The objective of this study was to study the efficacy of picloram herbicide in controlling larkspur plants.

- List and describe the components of the GLM (distribution, systematic component (predictor), and the link function).
- Fit the model according to part (a).
- Interpret your results.

Exercise 2.6.2 Effect of pH, Brix, temperature, and nisin concentration on the growth of *Alicyclobacillus acidoterrestris* CRA7152 in apple juice. The objective of this experiment was to model the presence/absence of CRA7152 growth in apple juice as a function of pH (3.5–5.5), Brix (11–19), temperature (25–50 °C), and nisin concentration (0–70). The data are shown below (Table 2.15):

- List and describe the components of the GLM (distribution, systematic component (predictor), and the link function).
- Fit the model according to part (a).
- Interpret your findings.

Exercise 2.6.3 The objective of this experiment was to evaluate the level of toxicity of concentrations of pyrethrin and piperonyl butoxide on the mortality of beetles (*Tribolium castaneum*). Pyrethrin is a natural insecticide found in the plant *Chrysanthemum cinerariaefolium* and its flowers. The active ingredients are pyrethrins I and II, cinerins I and II, and jasmolins I and II. The dried flowers contain 0.9–1.3% pyrethrum. The crude extract contains 50–60% pyrethrum and is imported from

Table 2.14 Toxicity of picloram in controlling larkspur plants

Rep	Conc	Y	Rep	Conc	Y	Rep	Conc	Y
1	0	0		0	0		0	0
1	0	0		0	0		0	0
1	0	0		0	0		0	0
1	0	0		0	0		0	0
1	0	0		0	0		0	0
1	0	0		0	0		0	0
1	0	0		0	0		0	0
1	0	0		0	0		0	0
1	0	0		0	0		0	0
1	0	0		0	0		0	0
1	0	0		0	0		0	0
1	0	0		0	0		0	0
1	0	0		0	0		0	0
1	0	0		0	0		0	0
1	0	0		0	0		0	0
1	0	0		0	0		0	0
1	0	0		0	0		0	0
1	0	0		0	0		0	0
1	0	0		0	0		0	0
1	0	0		0	0		0	0
1	0	0		0	0		0	0
1	1.1	0		0	0		0	0
1	1.1	0		0	0		0	0
1	1.1	0		0	0		0	0
1	1.1	0		0	0		0	0
1	1.1	0		0	0		0	0
1	1.1	0		0	0		0	0
1	1.1	0		0	0		0	0
1	1.1	0		0	0		0	0
1	1.1	0		0	0		0	0
1	1.1	0		0	0		0	0
1	1.1	0		0	0		1.1	0
1	1.1	0		1.1	0		1.1	0
1	1.1	0		1.1	0		1.1	0
1	1.1	1		1.1	0		1.1	0
1	1.1	1		1.1	0		1.1	0
1	1.1	1		1.1	0		1.1	0
1	1.1	1		1.1	0		1.1	0
1	1.1	1		1.1	0		1.1	0
1	1.1	1		1.1	0		1.1	0
1	1.1	1		1.1	0		1.1	0
1	1.1	1		1.1	1		1.1	0
1	1.1	1		1.1	1		1.1	0
1	1.1	1		1.1	1		1.1	0

(continued)

Table 2.14 (continued)

Rep	Conc	Y	Rep	Conc	Y	Rep	Conc	Y
1	2.2	0		1.1	1		1.1	0
1	2.2	1		1.1	1		1.1	0
1	2.2	1		1.1	1		1.1	0
1	2.2	1		1.1	1		1.1	0
1	2.2	1		1.1	1		1.1	0
1	2.2	1		1.1	1		1.1	0
1	2.2	1		1.1	1		1.1	0
1	2.2	1		1.1	1		1.1	0
1	2.2	1		2.2	0		1.1	0
1	2.2	1		2.2	0		1.1	1
1	2.2	1		2.2	0		1.1	1
1	2.2	1		2.2	0		1.1	1
1	2.2	1		2.2	0		1.1	1
1	2.2	1		2.2	1		1.1	1
1	2.2	1		2.2	1		1.1	1
1	2.2	1		2.2	1		1.1	1
1	2.2	1		2.2	1		2.2	0
1	2.2	1		2.2	1		2.2	0
1	2.2	1		2.2	1		2.2	1
1	2.2	1		2.2	1		2.2	1
1	2.2	1		2.2	1		2.2	1
1	2.2	1		2.2	1		2.2	1
1	2.2	1		2.2	1		2.2	1
1	2.2	1		2.2	1		2.2	1
1	2.2	1		2.2	1		2.2	1
1	2.2	1		2.2	1		2.2	1
1	2.2	1		2.2	1		2.2	1
1	4.5	1		2.2	1		2.2	1
1	4.5	1		2.2	1		2.2	1
1	4.5	1		2.2	1		2.2	1
1	4.5	1		4.5	1		2.2	1
1	4.5	1		4.5	1		2.2	1
1	4.5	1		4.5	1		2.2	1
1	4.5	1		4.5	1		2.2	1
1	4.5	1		4.5	1		2.2	1
1	4.5	1		4.5	1		2.2	1
1	4.5	1		4.5	1		2.2	1
1	4.5	1		4.5	1		2.2	1
1	4.5	1		4.5	1		2.2	1
1	4.5	1		4.5	1		2.2	1
1	4.5	1		4.5	1		2.2	1
1	4.5	1		4.5	1		2.2	1
1	4.5	1		4.5	1		2.2	1
1	4.5	1		4.5	1		2.2	1
1	4.5	1		4.5	1		2.2	1

(continued)

Table 2.15 Growth of *Alicyclobacillus acidoterrestris* CRA7152

pH	Nisin	Temp (°C)	Brix	Y	pH	Nisin	Temp (°C)	Brix	Y
5.5	70	50	11	0	5.5	70	50	19	0
5.5	70	43	19	0	3.5	0	25	11	0
5.5	50	43	13	1	5.5	70	50	11	0
5.5	50	35	15	1	5.5	70	43	19	0
5.5	30	35	13	1	5.5	50	43	13	1
5.5	30	25	11	0	5.5	50	35	15	1
5.5	0	50	19	0	5.5	30	35	13	1
5.5	0	25	15	1	5.5	30	25	11	0
3.5	70	43	15	0	5.5	0	50	19	0
3.5	70	35	11	0	5.5	0	25	15	1
3.5	50	50	13	0	3.5	70	43	15	0
3.5	50	35	19	0	3.5	70	35	11	0
3.5	30	50	11	0	3.5	50	50	13	0
3.5	30	43	15	0	3.5	50	35	19	0
3.5	0	25	19	0	3.5	30	50	11	0
5	70	25	15	0	3.5	30	43	15	0
5	70	25	13	0	3.5	0	25	19	0
5	50	50	15	1	5	70	25	15	0
5	50	25	19	0	5	70	25	13	0
5	30	43	19	0	5	50	50	15	1
5	30	43	11	1	5	50	25	19	0
5	0	50	13	1	5	30	43	19	0
5	0	35	11	1	5	30	43	11	1
4	70	50	19	0	5	0	50	13	1
4	70	35	13	0	5	0	35	11	1
4	50	43	11	0	4	70	50	19	0
4	50	25	11	0	4	70	35	13	0
4	30	50	15	1	4	50	43	11	0
4	30	35	19	0	4	50	25	11	0
4	30	25	13	0	4	30	50	15	1
4	0	43	15	1	4	30	35	19	0
4	0	43	13	1	4	30	25	13	0
3.5	0	35	11	0	4	0	43	15	1
4	0	35	11	1	4	0	43	13	1
5	0	43	11	1	3.5	0	35	11	0
					4	0	35	11	1
					5	0	43	11	1
					5.5	70	50	19	0
					3.5	0	25	11	0

Table 2.16 Mixture: pyrethrin plus piperonyl; n is the number of beetles exposed and Y is number of beetles killed

Mixture	n	Y
1.5	150	138
1.06	149	75
0.75	150	32
1.35	151	129
1.03	151	65
0.8	150	19
3.3	149	143
3.07	150	112
2.9	140	37
10.65	150	141
10.46	150	117
10.32	149	56
0	200	1

Table 2.17 Results of the experiment with carbon disulfide

Dose	Number of exposed beetles	Number of dead beetles	Proportion of dead beetles
49.1	59	6	0.102
53	60	13	0.217
56.9	62	18	0.29
60.8	56	28	0.5
64.8	63	52	0.825
68.7	59	53	0.898
72.6	62	61	0.984
76.5	60	61	1

Exercise 2.6.4 The objective of this experiment was to model the probability of mortality of the toxic effect of carbon disulfide (CS_2) gas on beetles. The insects were exposed to various concentrations of this gas (in mf/L) for 5 hours (Bliss 1935), and, then the number of dead beetles (Y) was counted. The data are shown below (Table 2.17).

- List and describe the components of the GLM (distribution, systematic component (predictor), and the link function).
- Fit the model according to part (a).
- Interpret your results.

Exercise 2.6.5 A study was conducted to assess the fowlpox virus in chorioallantois by the Pock counting technique. The membrane Pock count for 50 embryos exposed to one of four dilutions of virus (multiples of $10^{(-3.86)}$). The FD column heading corresponds to the dilution factor and the number of Pocks observed (Table 2.18).

Table 2.18 Results of the fowl pox experiment

FD	Count	FD	Count	FD	Count	FD	Count
0.125	1	0.25	5	0.5	5	1	12
0.125	2	0.25	2	0.5	11	1	9
0.125	2	0.25	3	0.5	7	1	11
0.125	3	0.25	2	0.5	5	1	17
0.125	2	0.25	5	0.5	4	1	11
0.125	2	0.25	0	0.5	6	1	10
0.125	1	0.25	2	0.5	5	1	8
0.125	0	0.25	2	0.5	9	1	16
0.125	0	0.25	0	0.5	4	1	15
0.125	1	0.25	3	0.5	7	1	12
0.125	2	0.25	2	0.5	4		
0.125	1	0.25	2	0.5	8		
0.125	1			0.5	4		
0.125	2						
0.125	1						

Table 2.19 Number of reversed *Salmonella* TA98 colonies

Quinoline dosage ($\mu\text{g/placa}$)					
0	10	33	100	333	1000
15	16	16	27	33	20
21	18	26	41	38	27
19	21	33	60	41	42

- (a) List and describe the components of the GLM (distribution, systematic component (predictor), and the link function).
- (b) Fit the model according to part (a).
- (c) Interpret your findings.

Exercise 2.6.6 Data were provided by Margolin et al. (1981) from an Ames *Salmonella* reverse mutagenicity assay. The table shows the number of reversed colonies observed on each of the three plates (repeats) tested at each of the six quinoline dose levels. The focus is on testing for mutagenic effects over time in the excess variation typically observed between counts (Table 2.19).

- (a) List and describe the components of the GLM (distribution, systematic component (predictor), and the link function).
- (b) Fit the model according to part a).
- (c) Interpret your results.

Appendix

Students infected with a certain disease								
id	Day	Students	id	Day	Students	id	Day	Students
1	1	6	45	45	3	89	95	1
2	2	8	46	46	3	90	96	0
3	3	12	47	48	3	91	96	0
4	3	9	48	48	2	92	97	1
5	4	3	49	49	3	93	98	1
6	4	3	50	49	1	94	100	2
7	4	11	51	53	3	95	101	2
8	6	5	52	53	3	96	102	1
9	7	7	53	53	5	97	103	1
10	8	3	54	54	4	98	104	1
11	8	8	55	55	4	99	105	1
12	8	4	56	56	3	100	106	0
13	8	6	57	56	5	101	107	0
14	12	8	58	58	4	102	108	0
15	14	3	59	60	3	103	109	1
16	15	6	60	63	5	104	110	1
17	17	3	61	65	3	105	111	0
18	17	2	62	67	4	106	112	0
19	17	2	63	67	2	107	113	0
20	18	6	64	68	3	108	114	0
21	19	3	65	71	3	109	115	0
22	19	7	66	71	1			
23	20	7	67	72	3			
24	23	2	68	72	2			
25	23	2	69	72	5			
26	23	8	70	73	4			
27	24	3	71	74	3			
28	24	6	72	74	0			
29	25	5	73	74	3			
30	26	7	74	75	3			
31	27	6	75	75	4			
32	28	4	76	80	0			
33	29	4	77	81	3			
34	34	3	78	81	3			
35	36	3	79	81	4			
36	36	5	80	81	0			
37	42	3	81	88	2			
38	42	3	82	88	2			
39	43	3	83	90	1			
40	43	5	84	93	1			

(continued)

Students infected with a certain disease								
id	Day	Students	id	Day	Students	id	Day	Students
41	44	3	85	93	2			
42	44	5	86	94	0			
43	44	6	87	95	2			
44	44	3	88	95	1			

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



Chapter 3

Objectives of Inference for Stochastic Models



Throughout this book, we have been using the pseudonym GLMMs to denote generalized linear mixed models. The common denominator among all these models is that they all contain a linear model (LM) part, which refers to the fixed effects component of the linear predictor $X\beta$. In a GLMM, the prefix “G” indicates that the distribution of observations may not be normal, the suffix of the first M means that the linear predictor includes mixed effects and thus contains random effects, which are expressed by the term “Zb.” The fixed linear component of the predictor $X\beta$ is important because the fixed effects describe the treatment design, which, in turn, is determined by the objectives or the initial research questions that the study wishes to answer. Therefore, if the researcher proposes using a reasonable model to analyze an experiment, then he/she must be able to express each objective as a question about a model parameter or as a linear combination of model parameters.

Example Assume a factorial 2×2 model, with two levels in both factors A and B, in which all possible combinations are tested. In this case, $X\beta$ corresponds to a two-way model with interaction and a predictor given by

$$\eta_{ij} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij}; i, j = 1, 2$$

As in all the statistical models studied so far, the linear predictor is expressed in terms of the link function, and η_{ij} can estimate the mean μ_{ij} (a combination of treatments) directly if the data follow a normal distribution and indirectly if the data are not normally distributed. For this example, the inference should focus on one or more of the following options (estimable functions): a treatment combination mean; a main effect mean; the mean of factor A, which is the average of the overall levels of factor B or vice versa; the difference of the main effects or the difference of a single effect, i.e., the difference between two levels of factor A at a given level of B or the difference between two levels of factor B at a given level of factor A; and so on. Each of these options can be expressed in terms of the parameters of the linear predictor, as shown in Table 3.1.

Table 3.1 Estimable functions in a factorial 2×2 treatment structure using the identity link function

Target estimate	Parameter estimator of the linear predictor	Target estimation in terms of the expected value
Combination $A \times B$	$\eta + \alpha_i + \beta_j + (\alpha\beta)_{ij}$	μ_{ij}
Main effect of factor A	$\bar{\eta}_i = \eta + \alpha_i + \frac{1}{2} \sum_j \beta_j + \frac{1}{2} \sum_j (\alpha\beta)_{ij}$	$\bar{\mu}_i = \frac{1}{2} \sum_j \mu_{ij}$
Main effect of factor B	$\bar{\eta}_j = \eta + \frac{1}{2} \sum_i \alpha_i + \beta_j + \frac{1}{2} \sum_i (\alpha\beta)_{ij}$	$\bar{\mu}_j = \frac{1}{2} \sum_i \mu_{ij}$
Difference between level 1 and level 2 of factor A	$\bar{\eta}_1 - \bar{\eta}_2 = \alpha_1 - \alpha_2 + \frac{1}{2} \left[\sum_j (\alpha\beta)_{1j} - \sum_j (\alpha\beta)_{2j} \right]$	$\bar{\mu}_1 - \bar{\mu}_2$
Difference between level 1 and level 2 of factor B	$\bar{\eta}_{1.} - \bar{\eta}_{2.} = \frac{1}{2} \left[\sum_i (\alpha\beta)_{i1} - \sum_i (\alpha\beta)_{i2} \right] + \beta_1 - \beta_2$	$\bar{\mu}_1 - \bar{\mu}_2$
Simple effect of A given B_j (AB_j)	$\eta_{1j} - \eta_{2j} = \alpha_1 - \alpha_2 + (\alpha\beta)_{1j} - (\alpha\beta)_{2j}$	$\mu_{1j} - \mu_{2j}$
Simple effect of B given A (BA_i)	$\eta_{i1} - \eta_{i2} = (\alpha\beta)_{i1} - (\alpha\beta)_{i2} + \beta_1 - \beta_2$	$\mu_{i1} - \mu_{i2}$
Interaction between factors A and B ($A \times B$) = $AB_1 - AB_2$ = $B A_1 - B A_2$	$(\eta_{11} - \eta_{21}) - (\eta_{12} - \eta_{22}) = (\alpha\beta)_{11} - (\alpha\beta)_{21} - (\alpha\beta)_{12} + (\alpha\beta)_{22} = (\eta_{11} - \eta_{12}) - (\eta_{21} - \eta_{22})$	$\mu_{11} - \mu_{12} - \mu_{21} + \mu_{22}$

Assuming that the data have a normal distribution, which is equivalent to using an identity link function, the estimator, in terms of the linear predictor (column 2), estimates the expected values of column three. If the data do not follow a normal distribution, then column 2 indirectly estimates the expected values of column three, and, in order to estimate the expected values, link functions are required. For link functions other than identity, the estimates in column two require a more careful handling. In an experimental design with a factorial treatment structure, the analysis should focus on the interaction of the two factors. If this interaction is significant, then the simple effects are not equal; however, if the interaction is not significant, then the main effects provide useful information; otherwise, the main effects are confounded. Therefore, for this reason, in this case, it is better to focus on the simple effects.

3.1 Three Aspects to Consider for an Inference

When constructing a model, the researcher must decide whether the effects are fixed or random. This decision has important implications with respect to the estimation criteria and in the interpretation of the tests and estimates obtained. Given these implications, three important aspects, described in the following sections, must be taken into consideration in statistical modeling.

3.1.1 *Data Scale in the Modeling Process Versus Original Data*

This is a very particular issue for models with a link function other than “identity,” since the scale of the data used in the modeling process is not always the same as the scale of the original data when the assumption of normality in the response variable is no longer valid. When the data are normally distributed, the estimable function directly estimates the expected value. However, this is not true if the data follow a non-normal distribution. For example, in a logistic model for binomial data in a completely randomized design, the estimable function $\eta + \tau_i$ estimates a logit or “log” odds. In this vein, $\eta + \tau_i$ must be expressed as a probability and not as a logit, i.e., the expected value for individuals receiving the i th treatment is a probability. This requires converting the estimate to a probability, using the inverse link; that is:

$$\pi_i = \frac{1}{(1 + e^{-(\eta + \tau_i)})}$$

Thus, for functions other than “identity,” there are two ways of expressing the estimates: (1) in terms of the parameters directly estimated from the GLMM (model scale) or (2) in terms of the expected value of the response variable (data scale).

3.1.2 *Inference Space*

This problem arises only when the linear predictor contains random effects. In these models, the estimates are obtained through a linear combination (an estimable function) with fixed effects, even though the linear predictor contains random effects. $\mathbf{K}'\boldsymbol{\beta}$ denotes the estimable function, where \mathbf{K} is the matrix of order $[(p + 1) \times k]$ and $\boldsymbol{\beta}$ is the vector of fixed effects parameters of order $[(p + 1) \times 1]$. The estimable function ($\mathbf{K}'\boldsymbol{\beta}$) represents a broad inference as it generalizes results to the entire population represented by the random effects.

Although the linear combination $\mathbf{K}'\boldsymbol{\beta} + \mathbf{Z}'\mathbf{b}$ is a predictable function with \mathbf{Z}' , a matrix for random effects with nonzero coefficients, its inference is limited to only those levels defined in \mathbf{b} . Suppose that you are conducting an experiment with three treatments at different locations (L), then the estimable function $\tau_1 - \tau_2$ provides information for the inference about the difference between treatments 1 and 2 in the whole population under study. Although the predictable function $[\tau_1 - \tau_2 + (L\tau)_{1j} - (L\tau)_{2j}]$ constrains the inference space between treatments 1 and 2, it is limited to location (L_j). The type of inference produced by predictable functions is called “narrow inference” because the nonzero coefficients in matrix \mathbf{Z} reduce the scope of inference for the entire population at those levels identified in \mathbf{Z} . Thus, the predictive function $\mathbf{K}'\boldsymbol{\beta} + \mathbf{Z}'\mathbf{b}$ should be used for specific estimates, whereas the estimable function $\mathbf{K}'\boldsymbol{\beta}$ should be used for valid estimates for the entire population under study.

3.1.3 Inference Based on Marginal and Conditional Models

As mentioned in the previous chapter, the specification of a generalized linear mixed model (GLMM) is done in terms of two probability distributions: (1) the distribution of the observations, given the random effects $\mathbf{y} \mid \mathbf{b}$ and (2) the distribution of the random effects \mathbf{b} . This feature is very particular to Gaussian (and non-Gaussian) mixed models (MMs), for this reason, it is also valid for mixed models with response variables that are different from a normal distribution.

From the probability theory, the marginal probability distribution of data (\mathbf{y}) can be obtained by integrating over the random effects, \mathbf{b} , from the joint probability distribution of \mathbf{y} and \mathbf{b} . Of the two distributions, the marginal distribution of data is the only one that can be known and observable. Many non-Gaussian mixed models, which seem reasonable, do not distinguish between the distribution of $\mathbf{y} \mid \mathbf{b}$ and \mathbf{y} . Models that do not make this distinction are called marginal models. Estimates obtained by marginal models have different expected values compared to those produced by conditional models. Therefore, marginal models are not estimated in the same way as conditional models.

3.2 Illustrative Examples of the Data Scale and the Model Scale

In linear models, inference begins with the estimable function $\mathbf{K}'\boldsymbol{\beta}$, and, these models, in turn, are defined in terms of the linear function $\boldsymbol{\eta} = \mathbf{g}(\boldsymbol{\mu}) = \mathbf{X}\boldsymbol{\beta}$ (if there are no random effects) and $\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b}$ (if there are random effects in the model), whereas $\mathbf{K}'\boldsymbol{\beta}$ produces results in terms of the link function.

For linear normal response models such as LMs and LMMs, the link function is not visible because they use the “identity” function as the link. Linear combinations of model parameters directly estimate desired values such as differences between treatments and many other hypothesis tests of interest. Inference for an LM is straightforward.

For GLMs and GLMMs with a non-normal response, the estimation of $\mathbf{K}'\boldsymbol{\beta}$ yields a linear combination of elements of the linear predictor $\boldsymbol{\eta}$, which is a linear combination of $\mathbf{g}(\boldsymbol{\mu})$, typically a nonlinear function of $\boldsymbol{\mu}$. For example, with Poisson data the $\mathbf{K}'\boldsymbol{\beta}$ is a function of logarithm (log) and for binomial data, it is a function of logit or probit. However, most of the time, the researcher wants to see the binomial results expressed in terms of the probability of the outcome of interest, whereas for Poisson, the results are expressed in terms of counts. This means that since both GLMs and GLMMs carry out the estimation process on the scale of the model (depending on the link used) to report the results of interest in terms of the scale of the data, it is necessary to apply the inverse link to the predictor in terms of the model scale to express the results. To mention two examples, in the case of the logit link for binomial, the results are expressed in terms of probability and, in the case of the

Table 3.2 Percentage of germinated seeds (Y) out of total seeds (N)

Treatment (Trt)	Y (no. of germinated seeds)	N (total no. of seeds)
Trt1	54	70
Trt1	41	60
Trt1	52	70
Trt2	28	70
Trt2	22	60
Trt2	21	70
Trt3	41	70
Trt3	37	60
Trt3	47	70

Poisson model, they are expressed in terms of counts. To exemplify the model scale and the data scale, an example is shown below.

Example 3.1 Consider the following experiment in which three chemical seed coat softeners were tested for studying their effect on germination of tomato seeds in Styrofoam trays (Table 3.2).

To illustrate the above two concepts, we first analyze these data using a completely randomized design (CRD), assuming the response variable to be normal, and, then, we analyze the same experimental design but with a binomial response variable. We are interested in comparing the means of treatments using a completely randomized design. Note that for demonstrative purposes, we are assuming that Y has a normal distribution, when in fact it has a binomial distribution.

The components of this model are defined as follows:

Distribution: $y_{ij} \sim N(\mu_i, \sigma^2)$

Linear predictor: $\eta_i = \eta + \tau_i; (i = 1, 2, 3)$

Link function: $\eta_i = \mu_i$ (identity link)

The analysis of variance (ANOVA) (part (a)) and estimated parameters (part (b)) of this experimental design indicate that there is a highly significant difference between the treatments ($P = 0.0033$) for the germination of tomato seeds. Table 3.3 shows part of the results.

The estimated parameter values of the model, except for treatment three, are shown in the table above (obtained with the “solution” command) because the model is over-parameterized. The estimable functions $K'\beta$ for the treatment means are as follows:

$$K' = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{bmatrix}; \beta = \begin{bmatrix} \eta \\ \tau_1 \\ \tau_2 \\ \tau_3 \end{bmatrix}$$

Table 3.3 Results of the analysis of variance using a CRD

(a) Type III tests of fixed effects							
Effect	Num degree of freedom (DF)			Den DF	F-value	Pr > F	
Trt	2			6	17.25	0.0033	
(b) Parameter estimates							
Effect	Trt		Estimate	Standard error	DF	t-value	Pr > t
Intercept		$\hat{\eta}$	41.6667	3.1388	6	13.27	<0.0001
Trt	Trt1	$\hat{\tau}_1$	7.3333	4.4389	6	1.65	0.1496
Trt	Trt2	$\hat{\tau}_2$	-18.0000	4.4389	6	-4.06	0.0067
Trt	Trt3	$\hat{\tau}_3$	0.	..	.		
Scale			29.5556	17.0639.	.	.	

From the estimated treatment parameters $\bar{\tau}_i = \bar{\mu}_i = \hat{\eta} + \hat{\tau}_i$, we can obtain the estimated mean for each one of the treatments ($i = 1, 2, 3$) as follows: for treatment 1, $\bar{\tau}_1 = \hat{\eta} + \hat{\tau}_1 = 41.6667 + 7.3333 = 49$; for treatment 2, $\bar{\tau}_2 = \hat{\eta} + \hat{\tau}_2 = 41.6667 - 18 = 23.6667$; and for treatment 3, $\bar{\tau}_3 = \hat{\eta} + \hat{\tau}_3 = 41.6667 + 0 = 41$. The value of the mean squared error ($\hat{\sigma}^2$), which appears in the table as “Scale,” is 29.5556.

For the difference between treatments, the $\tau_i - \tau_{i'}$ values for $i \neq i'$ are as follows: $\bar{\tau}_1 - \bar{\tau}_2 = \hat{\eta} + \hat{\tau}_1 - (\hat{\eta} + \hat{\tau}_2) = \hat{\tau}_1 - \hat{\tau}_2 = 7.3333 - (-18) = 25.3333$, $\bar{\tau}_1 - \bar{\tau}_3 = \hat{\eta} + \hat{\tau}_1 - (\hat{\eta} + \hat{\tau}_3) = \hat{\tau}_1 - \hat{\tau}_3 = 7.333 - 0.0 = 7.3333$, and $\bar{\tau}_2 - \bar{\tau}_3 = \hat{\eta} + \hat{\tau}_2 - (\hat{\eta} + \hat{\tau}_3) = \hat{\tau}_2 - \hat{\tau}_3 = -18.00 - 0.0 = -18.0$. These estimates can be obtained using the Statistical Analysis Software (SAS) “estimate” and “lsmeans” commands, as shown below:

```
proc glimmix data=germi;
class trt;
model y=trt / solution;
lsmeans trt / diff e;
estimate 'lsm trt 1' intercept 1 trt 1 0;
estimate 'lsm trt 2' intercept 1 trt 0 1;
estimate 'lsm trt 3' intercept 1 trt 0 0 1;
estimate 'overall mean' intercept 1 trt 0.33333 0.33333 0.33333
0.33333;
estimate 'overall mean' intercept 3 trt 1 1 1 1 / divisor=3;
estimate 'trt diff 1&2' trt 1 -1 0;
estimate 'trt diff 2&3' trt 0 1 -1;
run;
```

The “estimate” command requires us to specify what we wish to estimate and the “intercept” command refers to the intercept (η) and “Trt” to the treatment (τ_i) effects under evaluation; the coefficients needed for the estimates are shown above. While the “lsmeans” command invokes GLIMMIX in SAS to estimate the treatment means, “diff” asks to estimate the differences between pairs of treatments, and “E”

Table 3.4 Results obtained using the “estimate” and “lsmeans” commands

(a) Differences of Trt least squares means						
Trt	Trt	Estimate	Standard error	DF	t-value	Pr > t
Trt1	Trt2	25.3333	4.4389	6	5.71	0.0013
Trt1	Trt3	7.3333	4.4389	6	1.65	0.1496
Trt2	Trt3	-18.0000	4.4389	6	-4.06	0.0067
(b) Estimates						
Label	Estimate	Standard error	DF	t-value	Pr > t	
LSM Trt 1	49.0000	3.1388	6	15.61	<0.0001	
LSM Trt 2	23.6667	3.1388	6	7.54	0.0003	
LSM Trt 3	41.6667	3.1388	6	13.27	<0.0001	
Overall mean	38.1111	1.8122	6	21.03	<0.0001	
Overall mean	38.1111	1.8122	6	21.03	<0.0001	
Trt diff 1&2	25.3333	4.4389	6	5.71	0.0013	
Trt diff 2&3	-18.0000	4.4389	6	-4.06	0.0067	

displays the coefficients of the estimable functions used in “lsmeans.” Some of the outputs of the above code are shown in Table 3.4.

Next, we analyze the same data, also using a CRD, but now assuming a binomial distribution in the response variable. N indicates the independent number of Bernoulli trials observed in the ij th observation. The components of the model are as follows:

- Distribution: $y_{ij} \sim \text{Binomial}(N_{ij}, \pi_i)$
- Linear predictor: $\eta_i = \eta + \tau_i; (i = 1, 2, 3)$
- Link function: $\eta_i = \text{logit}\left(\frac{\pi_i}{1 - \pi_i}\right)$ (logit link)

Fitting these data in a binomial model, the fixed effects solution of the parameters obtained in terms of the model scale are tabulated in Table 3.5.

The above results were obtained using the following SAS code:

```
proc glimmix data=germi;
class trt;
model y/n=trt / solution;
run;
```

Similar to the previous example, we can estimate the mean of treatments and the differences between two pairs of treatments. The linear predictors for the treatments are as follows: $\hat{\eta}_1 = \hat{\eta} + \hat{\tau}_1 = 0.5108 + 0.5093 = 1.0201$, $\hat{\eta}_2 = \hat{\eta} + \hat{\tau}_2 = 0.5108 - 1.108 = -0.5971$, and $\hat{\eta}_3 = \hat{\eta} + \hat{\tau}_3 = 0.5108 + 0.0 = 0.5108$, and, for the differences between treatments (1 and 2, 1 and 3, and 2 and 3), they are as follows: $\hat{\eta}_1 - \hat{\eta}_2 = 1.0201 - (-0.5971) = 1.6173$, $\hat{\eta}_1 - \hat{\eta}_3 = 1.0201 - 0.5108 = 0.5093$, and $\hat{\eta}_2 - \hat{\eta}_3 = -0.5971 - 0.5108 = -1.1079$, respectively

Table 3.5 Estimated parameters at the model scale

Effect	Trt	Parameter estimates					
			Estimate	Standard error	DF	t-value	Pr > t
Intercept		$\hat{\eta}$	0.5108	0.1461	6	3.50	0.0129
Trt	Trt1	$\hat{\tau}_2$	0.5093	0.2168	6	2.35	0.0571
Trt	Trt2	$\hat{\tau}_2$	-1.1080	0.2078	6	-5.33	0.0018
Trt	Trt3	$\hat{\tau}_3$	0

Using the relationship between the linear predictor and the link function $\eta_i = \text{logit}\left(\frac{\pi_i}{1-\pi_i}\right) = \log\left(\frac{\pi_i}{1-\pi_i}\right)$, we can estimate the probability of observing a favorable outcome for each of the treatments, that is, π_1, π_2 , and π_3 , respectively. Applying the inverse link, we obtain:

$$\hat{\pi}_1 = 1 / \left(1 + e^{-(\hat{\eta} + \hat{\tau}_1)}\right); \hat{\pi}_2 = 1 / \left(1 + e^{-(\hat{\eta} + \hat{\tau}_2)}\right); \text{ and } \hat{\pi}_3 = 1 / \left(1 + e^{-(\hat{\eta} + \hat{\tau}_3)}\right)$$

Substituting the corresponding values, we obtain

$$\begin{aligned} \hat{\pi}_1 &= 1 / \left(1 + e^{-(\hat{\eta} + \hat{\tau}_1)}\right) = 1 / \left(1 + e^{-1.0201}\right) = 0.735, \\ \hat{\pi}_2 &= 1 / \left(1 + e^{-(\hat{\eta} + \hat{\tau}_2)}\right) = 1 / \left(1 + e^{-(-0.5971)}\right) = 0.355, \text{ and} \\ \hat{\pi}_3 &= 1 / \left(1 + e^{-(\hat{\eta} + \hat{\tau}_3)}\right) = 1 / \left(1 + e^{-(0.5108)}\right) = 0.625 \end{aligned}$$

Here, we can see that the treatment with the highest probability of success is treatment one, followed by treatment three, whereas treatment two has the lowest probability of success. Now, for the difference between two treatments, $\tau_i - \tau_{i'}$ for $i \neq i'$, we can estimate the logarithm of the odds ratio as

$$\tau_i - \tau_{i'} = \log\left(\frac{\pi_i}{1-\pi_i}\right) - \log\left(\frac{\pi_{i'}}{1-\pi_{i'}}\right) = \log\left(\frac{\pi_i}{1-\pi_i} / \frac{\pi_{i'}}{1-\pi_{i'}}\right)$$

where, in this particular case, $\text{odds} = \left(\frac{\pi_i}{1-\pi_i}\right)$ is the odds of the treatment i and $\text{oddsratio} = \log\left(\frac{\pi_i}{1-\pi_i} / \frac{\pi_{i'}}{1-\pi_{i'}}\right)$ is the odds ratio for treatments i and i' , for $i \neq i'$.

When applying the inverse link to the above expression (odds ratio), we get

$$\text{Oddsratio} = 1 / \left(1 + e^{-(\hat{\tau}_i - \hat{\tau}_{i'})}\right)$$

The value of the odds ratios for treatments 1 and 3 is

$$\text{Oddsratio}_{1-3} = 1/(1+e^{-(\hat{\tau}_1-\hat{\tau}_3)}) = 1/(1+e^{-0.5093}) = 0.6246$$

Similarly, for the pair of treatments 1–2 and 2–3, the resulting odds ratios are $\text{Oddsratio}_{1-2} = 0.8344$ and $\text{Oddsratio}_{2-3} = 0.2483$, respectively. It is important to mention that the odds ratios are not the mean of the difference of $\pi_i - \pi_{i'}$ for $i \neq i'$.

From the previous example, it is clear that when the response variable is not normal, parameter estimation and inference occurs at two levels. The linear predictor $X\beta$ and the estimable function $K'\beta$ are expressed in terms of the link function, logit – estimates on the model scale (scale of the link function) – as in the above example. Under the logit link, the logarithm of the odds and the difference of the estimate (log odds ratio) are very common and useful terms in categorical data analysis for the estimation of treatments or treatment differences in terms of the data scale.

Commonly, estimation at the model scale in GLMs is not very easy to interpret, and, as such, the data scale plays a very important role. A data scale involves applying the inverse of the link function to the estimable function, $K'\beta$, as we did in the previous example to convert the log of the odds for each treatment to a probability. In general, we use the inverse of the link function to transform the estimates at the model scale to the data scale. The inverse of the link function is not used for estimating the differences between treatments because the link functions are generally nonlinear. This is why the inverse of the link function is not applied to the differences between treatments because it produces meaningless results.

Thus, in the logit model, we have two approximations for the difference. First, we could apply the inverse of the link function to each linear predictor for each treatment and then take the difference between probabilities: $\hat{\pi}_i - \hat{\pi}_{i'}$. That is, we can estimate the difference between $\pi_i - \pi_{i'}$ through $[1/(1 + e^{-(\eta+\tau_i)})] - [1/(1 + e^{-(\eta+\tau_{i'})})]$ and not as $[1/(1 + e^{-(\hat{\tau}_i-\hat{\tau}_{i'})})]$. Second, we know that $\tau_i - \tau_{i'}$ estimates the logarithm of the odds ratio by means of $e^{(\tau_i-\tau_{i'})}$, which produces an estimate of the odds ratio. Both approaches are valid, and the use of one approach or the other depends on the requirements of the particular study.

With the GLIMMIX procedure, we can implement the solution in terms of the data scale with the “ilink,” “exp,” and “oddsratio” commands, as shown in the following SAS code:

```
proc glimmix;
class trt;
model y/n=trt / solution oddsratio;
lsmeans trt / diff oddsratio ilink ;
estimate 'lsm trt 1' intercept 1 trt 1 0/ilink;
estimate 'lsm trt 2' intercept 1 trt 0 1/ilink;
estimate 'lsm trt 3' intercept 1 trt 0 0 1/ilink;
estimate 'overall mean' intercept 1 trt 0.33333 0.33333 0.33333
0.33333/ilink;
estimate 'overall mean' intercept 3 trt 1 1 1 1 / divisor=3 ilink;
```

```

estimate 'trt diff 1&2' trt 1 -1 0/oddsratio ilink;
estimate 'trt diff 1&3' trt 1 0 -1/oddsratio ilink;
estimate 'trt diff 2&3' trt 0 1 -1/oddsratio ilink;
estimate 'trt diff 1&2' trt 1 -1 0/exp;
estimate 'trt diff 1&3' trt 1 0 -1/exp;
estimate 'trt diff 2&3' trt 0 1 -1/exp;
run;

```

Part of the output of “proc GLIMMIX” is shown in Table 3.6. The “Odds ratio estimates” (part (a)) are the result of the “oddsratio” command in the previous program, whereas the confidence intervals are provided by default.

What appears under “Estimate” (in part (b)) is in the model scale $\hat{\eta}_i = \hat{\eta} + \hat{\tau}_i$, and what appears under “Mean” (in part (b)) is an estimate of the inverse of the link function $\hat{\pi}_i = 1/(1 + e^{-(\hat{\eta} + \hat{\tau}_i)})$ and, in this case, is a probability that corresponds to the data scale. Similarly, what appears under “Estimate” is in model scale $\hat{\tau}_i - \hat{\tau}_j$, whereas the “Odds ratio” values were estimated using $e^{(\tau_i - \tau_j)}$ and are in data scale.

Under “Estimates” column in Table 3.7, the log odds ratio appears as an “Exponentiated estimate” regardless of whether we use the “oddsratio” or “exp” option in the “estimate” command. For the overall mean, the inverse of the link function applied to $\hat{\eta} + \frac{1}{3}(\hat{\tau}_1 + \hat{\tau}_2 + \hat{\tau}_3)$ is 0.5772, which is totally different from the average of $\hat{\pi}_{is}$; that is, $\frac{1}{3}(\hat{\pi}_1 + \hat{\pi}_2 + \hat{\pi}_3) = \frac{1}{3}(0.735 + 0.355 + 0.655) = 0.5816$. This illustrates that we have to be extremely careful when using the output of proc GLIMMIX, as it can produce outputs in terms of both the model scale and the data scale through the application of the inverse of the link function; however, this has to be applied appropriately, otherwise, we will get meaningless results.

Example 3.2: Randomized complete block design (RCBD) with normal and binomial responses

Now, assume that we have the same example but in an RCBD. The three treatments were tested in each of the blocks, as shown in Table 3.8.

In this example, first, the data are analyzed assuming a normal response and assuming that the block effect is fixed; then, they are analyzed assuming a binomial response.

The model components under a Gaussian response variable are as follows:

Distribution: $y_{ij} \sim N(\mu_{ij}, \sigma^2)$

Linear predictor: $\eta_{ij} = \eta + \tau_i + \text{block}_j$; ($i, j = 1, 2, 3$)

Link function: $\eta_{ij} = \mu_{ij}$; (identity link)

From the theory of linear models, we know that we can estimate the i th treatment mean through

$$\bar{\eta}_{i\cdot} = \frac{1}{3} \sum_{j=1}^3 y_{ij} = \eta + \tau_i + \frac{1}{3} \sum_{j=1}^3 \text{block}_j = \eta + \tau_i + \overline{\text{bloq}}.$$

Table 3.6 Results of the “ilink,” “exp,” and “oddsratio” commands

(a) Odds ratio estimates							
Trt	Trt	Estimate	DF	95% confidence limits			
Trt1	Trt3	1.664	6	0.979	2.829		
Trt2	Trt3	0.330	6	0.199	0.549		
(b) Trt least squares means							
Trt	Estimate	Standard error	DF	t-value	Pr > t	Mean	Standard error Mean
Trt1	1.0201	0.1602	6	6.37	0.0007	0.7350	0.03121
Trt2	-0.5971	0.1478	6	-4.04	0.0068	0.3550	0.03384
Trt3	0.5108	0.1461	6	3.50	0.0129	0.6250	0.03423
(c) Differences of Trt least squares means							
Trt	Trt	Estimate	Standard error	DF	t-value	Pr > t	Odds ratio
Trt1	Trt2	1.6173	0.2180	6	7.42	0.0003	5.039
Trt1	Trt3	0.5093	0.2168	6	2.35	0.0571	1.664
Trt2	Trt3	-1.1080	0.2078	6	-5.33	0.0018	0.330

Table 3.7 Estimates at the model scale and at the data scale

Label	Estimates							
	Estimate	Standard error	DF	t Value	Pr > t	Mean	Standard error Mean	Exponentiated estimate
LSM Trt 1	1.0201	0.1602	6	6.37	0.0007	0.7350	0.03121	
LSM Trt 2	-0.5971	0.1478	6	-4.04	0.0068	0.3550	0.03384	
LSM Trt 3	0.5108	0.1461	6	3.50	0.0129	0.6250	0.03423	
Overall Mean	0.3113	0.08746	6	3.56	0.0119	0.5772	0.02134	
Trt diff 1&2	1.6173	0.2180	6	7.42	0.0003	0.8344	0.03011	5.0393
Trt diff 1&3	0.5093	0.2168	6	2.35	0.0571	0.6246	0.05083	1.6642
Trt diff 2&3	-1.1080	0.2078	6	-5.33	0.0018	0.2483	0.03878	0.3302
Trt diff 1&2	1.6173	0.2180	6	7.42	0.0003			5.0393
Trt diff 1&3	0.5093	0.2168	6	2.35	0.0571			1.6642
Trt diff 2&3	-1.1080	0.2078	6	-5.33	0.0018			0.3302

Table 3.8 Percentage of germinated seeds (Y) out of total seeds (N) in a randomized complete block design

Treatment	Block	Y (no. of germinated seeds)	N (total no. of seeds)
Trt1	Block1	54	70
Trt1	Block2	41	60
Trt1	Block3	52	70
Trt2	Block1	28	70
Trt2	Block2	22	60
Trt2	Block3	21	70
Trt3	Block1	41	70
Trt3	Block2	37	60
Trt3	Block3	47	70

where $\overline{\text{bloq.}} = 1/3 \sum_{j=1}^3 \text{block}_j$.

For the mean difference of two treatments i and i' , this is estimated as

$$\bar{\eta}_{i.} - \bar{\eta}_{i'.} = \eta + \tau_i + \overline{\text{bloq.}} - (\eta + \tau_{i'} + \overline{\text{bloq.}}) = \tau_i - \tau_{i'}$$

The goal of this experiment could be to compare the treatment means, that is, $\bar{\eta}_{1.} = \bar{\eta}_{2.} = \bar{\eta}_{3.}$, equivalently – this can be expressed as $\tau_1 = \tau_2 = \tau_3$ – or to compare one treatment with the average of the other treatments: for example, to compare treatment 1 with the averages of treatments 2 and 3 (Trt1.vs.average.Trt2.and.Trt3).

For the hypothesis test of the equality of treatments ($\tau_1 = \tau_2 = \tau_3$), the estimable function $K'\beta$ is given by:

$$K' = \begin{bmatrix} 0 & 1 & 0 & -1 & 0 & 0 & 0 \\ 0 & 0 & 1 & -1 & 0 & 0 & 0 \end{bmatrix}; \beta = \begin{bmatrix} \eta \\ \tau_1 \\ \tau_2 \\ \tau_3 \\ \text{bloq}_1 \\ \text{bloq}_2 \\ \text{bloq}_3 \end{bmatrix}$$

While for contrasts Trt1.vs.average.Trt2.and.Trt3 and Trt2.vs.average.Trt1.and.Trt3, $K'\beta$ is given by

$$\text{Trt1.vs.average.Trt2.and.Trt3} = \bar{\eta}_{1.} - 1/2(\bar{\eta}_{2.} + \bar{\eta}_{3.}) = \tau_1 - \left(\frac{\tau_2 - \tau_3}{2}\right)$$

$$\text{Trt2.vs.average.Trt1.and.Trt3} = \bar{\eta}_{2.} - 1/2(\bar{\eta}_{1.} + \bar{\eta}_{3.}) = \tau_2 - \left(\frac{\tau_1 - \tau_3}{2}\right)$$

$$K' = \begin{bmatrix} 0 & 2 & -1 & -1 & 0 & 0 & 0 \\ 0 & 1 & -2 & 1 & 0 & 0 & 0 \end{bmatrix}; \beta = \begin{bmatrix} \eta \\ \tau_1 \\ \tau_2 \\ \tau_3 \\ \text{bloq}_1 \\ \text{bloq}_2 \\ \text{bloq}_3 \end{bmatrix}$$

The following GLIMMIX procedure allows us to implement the above example.

```
proc glimmix;
class trt block;
model y = trt block/solution;
lsmeans trt / diff e;
estimate 'lsm trt1' intercept 3 trt 3 0 0 0 block 1 1 1 / divisor=3;
estimate 'overall mean' intercept 3 trt 1 1 1 1 block 1 1 1 1 / divisor=3;
estimate 'average trt1&trt2' intercept 6 trt 3 3 0 block 2 2 2 /
divisor=6;
estimate 'average trt1&trt2&trt3' intercept 9 trt 3 3 3 3 block 3 3 3 3
3/divider=9;
estimate 'trt1 vs trt2' trt 1 -1 0 ;
estimate 'trt1 vs trt3' trt 1 0 -1;
estimate 'trt2 vs trt3' trt 0 1 -1;
estimate 'trt1 vs trt2' trt 1 -1 0, 'trt1 vs trt3' trt 1 0 -1, 'trt2 vs
trt3' trt 0 1 -1/divisor=1,1,1 adjust=sidak;
contrast 'trt1 vs trt2' trt 1 -1 0 ;
contrast 'trt1 vs trt3' trt 1 0 -1;
contrast 'trt2 vs trt3' trt 0 1 -1;
contrast 'trt1 vs average trt1, trt2' trt 2 -1 -1;
contrast 'trt2 vs average trt1, trt3' trt -1 2 -1;
contrast 'type 3 trt ss' trt 1 0 -1 0, trt 0 1 -1;
contrast 'type 3 trt test' trt 2 -1 -1, trt -1 2 -1;
run;
```

Part of the GLIMMIX output is shown below in Table 3.9. Parameter estimation for treatments 1–2 and blocks 1–2 are shown below, except for treatment and block 3. This is because it is an incomplete rank model. The generalized inverse is used in the estimation through the SWEEP operator of SAS. In this case, it sets the last class effect equal to zero (Table 3.9).

“Coefficients” (part (a) of Table 3.10) obtained with option E for the least squares means of treatments in “lsmeans” shows how SAS uses this information in the parameter solution to calculate the treatment means (part (b)). In part (c), we can see the difference of means obtained with the “diff” option in “lsmeans.”

The estimates obtained from the “estimate” command with multiple estimable functions and in the “Sidak” adjustment and contrasts are shown in Table 3.11. This adjustment allows us to control for type I errors. The “adjust” option in “estimate” in the Sidak adjustment (part (b)) allows us to obtain the adjusted P -values denoted as $AdjP$ in addition to $Pr > |t|$.

Table 3.9 Estimation of treatment and block parameters

Parameter estimates							
Effect	Trt	BLOCK	Estimate	Standard error	DF	<i>t</i> -value	Pr > <i>t</i>
Intercept ($\hat{\eta}$)			43.5556	3.1866	4	13.67	0.0002
Trt ($\hat{\tau}_1$)	1		7.3333	3.4907	4	2.10	0.1036
Trt ($\hat{\tau}_2$)	2		-18.0000	3.4907	4	-5.16	0.0067
Trt ($\hat{\tau}_3$)	3		0				
BLOCK ($\hat{\text{block}}_1$)		1	1.0000	3.4907	4	0.29	0.7887
BLOCK ($\hat{\text{block}}_2$)		2	-6.6667	3.4907	4	-1.91	0.1288
BLOCK ($\hat{\text{block}}_3$)		3	0				
Scale ($\hat{\sigma}^2$)			18.2778	12.9243			

Table 3.10 Coefficients for treatment and block used in least squares

(a) Coefficients for Trt least squares means					
Effect	Trt	BLOCK	Row1	Row2	Row3
Intercept			1	1	1
Trt	1		1		
Trt	2			1	
Trt	3				1
BLOCK		1	0.3333	0.3333	0.3333
BLOCK		2	0.3333	0.3333	0.3333
BLOCK		3	0.3333	0.3333	0.3333

(b) Trt least squares means					
Trt	Estimate	Standard error	DF	<i>t</i> -value	Pr > <i>t</i>
1	49.0000	2.4683	4	19.85	<0.0001
2	23.6667	2.4683	4	9.59	0.0007
3	41.6667	2.4683	4	16.88	<0.0001

(c) Differences of Trt least squares means						
Trt	_Trt	Estimate	Standard error	DF	<i>t</i> -value	Pr > <i>t</i>
1	2	25.3333	3.4907	4	7.26	0.0019
1	3	7.3333	3.4907	4	2.10	0.1036
3	3	-18.0000	3.4907	4	-5.16	0.0067

The planned contrasts in matrix K' and with the *F*-values obtained with the “contrast command” produce the same results (part (c)).

Now, the same dataset is fitted using the same predictor but assuming that the response variable is binomial. This analysis intends to show the options available in the SAS commands when you want to fit non-normal responses; in this case, it is binomial. Practically, the same commands used in the previous program with normal data are used, but, now, some other options (“ilink,” “oddsratio,” or “exp”) are exemplified with details under what circumstances they should be used. This is because all estimable functions produce estimates at the model scale, and we must

Table 3.11 Multiple estimates and contrasts

(a) Estimates						
Label	Estimate	Standard error	DF	<i>t</i> -value	Pr > <i>t</i>	
LSM Trt1	49.0000	2.4683	4	19.85	<0.0001	
Overall mean	38.1111	1.4251	4	26.74	<0.0001	
Average Trt1&Trt2	36.3333	1.7454	4	20.82	<0.0001	
Average Trt1&Trt2&Trt3	38.1111	1.4251	4	26.74	<0.0001	
Trt1 vs Trt2	25.3333	3.4907	4	7.26	0.0019	
Trt1 vs Trt3	7.3333	3.4907	4	2.10	0.1036	
Trt2 vs Trt3	-18.0000	3.4907	4	-5.16	0.0067	
(b) Estimate adjustment for multiplicity: Sidak						
Label	Estimate	Standard error	DF	<i>t</i> -value	Pr > <i>t</i>	Adj <i>P</i>
Trt1 vs Trt2	25.3333	3.4907	4	7.26	0.0019	0.0057
Trt1 vs Trt3	7.3333	3.4907	4	2.10	0.1036	0.2796
Trt1 vs Trt3	-18.0000	3.4907	4	-5.16	0.0067	0.0200
(c) Contrasts						
Label	Num DF	Den DF	<i>F</i> -value	Pr > <i>F</i>		
Trt1 vs Trt2	1	4	52.67	0.0019		
Trt1 vs Trt3	1	4	4.41	0.1036		
Trt2 vs Trt3	1	4	26.59	0.0067		
Trt1 vs average Trt1,Trt2	1	4	29.19	0.0057		
Trt2 vs average Trt1,Trt3	1	4	51.37	0.0020		
Type 3 Trt SS	2	4	27.89	0.0045		
Type 3 Trt Test	2	4	27.89	0.0045		

decide what conversions are necessary to obtain the results at the data scale. Below, the estimable functions and the appropriate conversion required to produce the results on the data scale are listed.

- (a) Least squares means (“lsmeans”) for normal data and an inverse link (“ilink”) for non-normal data
- (b) Difference between pairs of treatment means of “lsmeans” for normal data and “odds ratio” for non-normal data
- (c) Estimation of the mean of a treatment (“estimate”) for normal data and an inverse link (“ilink”) for non-normal data
- (d) Estimation of a treatment *i* vs treatment *i*’: exponentiation (“exp”) (or odds ratio)
- (e) Multiple estimates of treatment differences as “exp” (or odds ratio) for non-normal data
- (f) In “contrast estimation,” conversion to the data scale is not necessary, since it is only an *F*-statistic test.

The following GLIMMIX program shows how to implement this model with a binomial response.

```

proc glimmix;
class trt block;
model y/n = trt block/solution oddsratio;
lsmeans trt / diff e oddsratio;
estimate 'lsm trt1' intercept 3 trt 3 0 0 0 block 1 1 1 1/divider=3 ilink;
estimate 'difference trt1 vs trt2' trt 1 -1 0/exp;
estimate 'avg trt1&trt2&trt3' intercept 9 trt 3 3 3 3 block 3 3 3 3
3/divider=9;
estimate 'trt1 vs trt2' trt 1 -1 0/exp;
estimate 'trt1 vs trt3' trt 1 0 -1/exp;
estimate 'trt2 vs trt3' trt 0 1 -1/exp;
estimate 'trt1 vs trt2' trt 1 -1 0, 'trt1 vs trt3' trt 1 0 -1, 'trt1 vs
trt3' trt 0 1 -1/exp adjust=sidak;
contrast 'trt1 vs trt2' trt 1 -1 0;
contrast 'trt1 vs trt3' trt 1 0 -1;
contrast 'trt2 vs trt3' trt 0 1 -1;
contrast 'trt1 vs average trt1, trt2' trt 2 -1 -1;
contrast 'trt2 vs average trt1, trt3' trt -1 2 -1;
contrast 'type 3 trt ss' trt 1 0 -1 0, trt 0 1 -1;
contrast 'type 3 trt test' trt 2 -1 -1, trt -1 2 -1;
run;

```

Part of the output is shown in Table 3.12. The estimated treatment and block parameters of the model are given in part (a) of Table 3.12; the last two effects of both classes were restricted to zero because they are incomplete rank design matrices. In part (b), the type III tests of fixed effects and in part (c) the odds ratio estimates are provided. Note that $\hat{\sigma}^2$ does not appear in the output because the variance of the binomial distribution is not an independent parameter.

In Table 3.12 (parts (b) and (c)), which shows the sum of the squares of fixed effects type III as well as the odds ratio, it can be seen that only the effect of treatments is significant but not the effect of blocks, which indicates that it is valid to analyze these data using a completely randomized design. Two sets of odds ratios were estimated (part (c)): one for the treatment effects and the other for the block effects in the model. In the calculation of odds ratios, generally, the last level of the factor is compared with the rest of the levels of that same factor.

The estimates obtained with “estimate”, in Table 3.13 (parts (a) and (b)), are results in terms of the model scale, whereas the last column is obtained by applying EXP ($e^{\tau_i - \tau_i'}$).

The least squares means for treatment and the linear predictors of treatment differences (parts (a) and (b) of Table 3.14, respectively) obtained with “lsmeans” are the values under the “Estimate” column, and, these, together with their corresponding standard errors, were obtained using the linear predictor $\hat{\eta}_{ij} = \hat{\eta} + \hat{\tau}_i + \widehat{\text{block}}_{.}$.

These estimates are on the model scale, whereas the values under the “Mean” column and their respective standard errors were obtained by applying the inverse link to obtain the probabilities of success of each treatment ($\hat{\pi}_i$). While the estimated linear predictors for the mean differences were obtained with the “oddsratio” option, the mean difference in the data scale is obtained by taking the inverse of these predictors.

Table 3.12 Results of the analysis of variance in a binomial model

(a) Parameter estimates							
Effect	Trt	BLOCK	Estimate	Standard error	DF	<i>t</i> -value	Pr > <i>t</i>
Intercept ($\hat{\eta}$)			0.5099	0.1883	4	2.71	0.0536
Trt ($\hat{\tau}_1$)	1		0.5097	0.2169	4	2.35	0.0785
Trt ($\hat{\tau}_2$)	2		-1.1088	0.2079	4	-5.33	0.0059
Trt ($\hat{\tau}_3$)	3		0				
BLOCK ($\hat{\text{block}}_1$)		1	0.06541	0.2088	4	0.31	0.7698
BLOCK ($\hat{\text{block}}_2$)		2	-0.07205	0.2164	4	-0.33	0.7559
BLOCK ($\hat{\text{block}}_3$)		3	0				
(b) Type III tests of fixed effects							
Effect	Num DF		Den DF		<i>F</i> -value		Pr > <i>F</i>
Trt	2		4		29.55		0.0040
BLOCK	2		4		0.20		0.8258
(c) Odds ratio estimates							
Trt	BLOCK	_Trt	_BLOCK	estimate	DF	95% confidence limits	
1		3		1.665	4	0.912	3.040
2		3		0.330	4	0.185	0.588
	1		3	1.068	4	0.598	1.906
	2		3	0.930	4	0.510	1.697

3.3 Fixed and Random Effects in the Inference Space

In an analysis, inference can be directed solely at fixed effects (population inference) or at a combination of fixed and random effects (specific inference). To illustrate these two levels of inferences, we will consider two examples:

3.3.1 A Broad Inference Space or a Population Inference

In practice, the random effects in a linear mixed model (LMM) should represent the population from which the data were collected and should be included in studies as if they came from a well-planned sample. In a model, random effects can be locations, regions, states, blocks, and so on, and they have two very particular characteristics.

- Random effects represent the target population.
- Random effects have a probability distribution.

These two characteristics allow us to have a broad inference space where we can calculate point estimates, estimate intervals, and perform hypothesis testing applicable to the entire population represented by the random effects. Formally, an estimate or hypothesis test based on an LMM indicates that we have a broad

Table 3.13 Different estimates obtained with “estimate”

(a) Estimates										
Label	Estimate	Standard error	DF	t-value	Pr > t	Mean	Standard error Mean	Exponentiated estimate		
LSM Trt1	1.0174	0.1603	4	6.35	0.0032	0.7345	0.03127	.		
Average Trt1&Trt2&Trt3	0.3080	0.08768	4	3.51	0.0246	Non-est		5.0452		
Trt1 vs Trt2	1.6184	0.2181	4	7.42	0.0018	Non-est		1.6647		
Trt1 vs Trt3	0.5097	0.2169	4	2.35	0.0785	Non-est		0.3300		
Trt2 vs Trt3	-1.1088	0.2079	4	-5.33	0.0059	Non-est				
(b) Estimate adjustment for multiplicity: Sidak										
Label	Estimate	Standard error	DF	t-value	Pr > t	Adj P	Exponentiated estimate			
Trt1 vs Trt2	1.6184	0.2181	4	7.42	0.0018	0.0053	5.0452			
Trt1 vs Trt3	0.5097	0.2169	4	2.35	0.0785	0.2175	1.6647			
Trt1 vs Trt3	-1.1088	0.2079	4	-5.33	0.0059	0.0177	0.3300			

Table 3.14 Estimated linear predictors for treatments and treatment differences with their respective inverse values

(a) Trt least squares means							
Trt	Estimate	Standard error	DF	t-value	Pr > t	Mean	Standard error Mean
1	1.0174	0.1603	4	6.35	0.0032	0.7345	0.03127
	-0.6011	0.1480	4	-4.06	0.0153	0.3541	0.03385
	0.5077	0.1462	4	3.47	0.0255	0.6243	0.03429
(b) Differences of Trt least squares means							
Trt	_Trt	Estimate	Standard error	DF	t-value	Pr > t	Odds ratio
1	2	1.6184	0.2181	4	7.42	0.0018	5.045
1	3	0.5097	0.2169	4	2.35	0.0785	1.665
2	3	-1.1088	0.2079	4	-5.33	0.0059	0.330

inference space defined by the estimable function $K'\beta$ if Z is a matrix with coefficients equal to zero; otherwise, the estimation or hypothesis test is defined by the prediction function $K'\beta + Z'\beta$, which is a specific inference.

3.3.2 Mixed Models with a Normal Response

In Example 3.2, the response variable was assumed as a function of fixed effects due to treatments and blocks, since block effects were also assumed to be fixed effects. Now, suppose that applications of treatments were done by three different people (blocks); then, assuming that the block effects are fixed, this would be questionable since each person does their job according to their experience, skill, and so forth. Clearly, there is some variability between blocks that is not due to the experiment and this has to be removed, so the effects due to blocks must be considered random. In this example, let us assume that the three blocks (persons) were randomly selected from a population. Thus, the components of the model are defined as follows:

Distribution:

- (a) $y_{ij} | \text{block}_j \sim N(\mu_{ij}, \sigma^2)$
- (b) $\text{bloque}_j \sim N(0, \sigma_{\text{block}}^2)$

Linear predictor: $\eta_{ij} = \eta + \tau_i + \text{block}_j$ ($i, j = 1, 2, 3$)

Link function: $\eta_{ij} = \mu_{ij}$ (identity link)

Note the impact of changing the estimable function for the mean of treatments. In Example 3.2, the estimable function was defined by $E(\bar{y}_{i.}) = \eta + \tau_i + \overline{\text{block}}..$ Now with the mixed model (fixed effects and random blocks), the estimable function is

defined by $E(\bar{y}_i \cdot) = \eta + \tau_i$ because $E(\text{block}) = 0$. Therefore, the estimable function for the mean in each of the treatments is $\eta + \tau_i$. In this situation, two questions arise:

- How much do the results obtained from a fixed effects model differ from those obtained from a mixed model?
- How can we compare the two results?

The following program allows us to estimate a mixed model with a normal response.

```
proc glimmix;
class trt block;
model y = trt /solution;
random block/solution;
lsmeans trt /diff e;
estimate 'lsm trt1' intercept 1 trt 1 0 0 0 |block 0 0 0 0;
estimate 'lsm trt2' intercept 1 trt 0 1 0 |block 0 0 0 0;
estimate 'lsm trt3' intercept 1 trt 0 0 0 1 |block 0 0 0 0;
estimate 'blup trt1' intercept 3 trt 3 0 0 0 |block 1 1 1 /divisor=3;
estimate 'blup trt2' intercept 3 trt 0 3 0 |block 1 1 1 /divisor=3;
estimate 'blup trt3' intercept 3 trt 0 0 3 |block 1 1 1 /divisor=3;
run;
```

In the previous SAS GLIMMIX code, the “estimate” command shows the coefficients associated with the fixed effects before the vertical bar (|) and after the vertical bar, are provided the coefficients for the random effects associated with the model, that is:

$$K' \beta + Z' b = \left[\begin{array}{c} \text{efectosfijos} \\ \begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \eta \\ \tau_1 \\ \tau_2 \\ \tau_3 \end{pmatrix} + \begin{array}{c} \text{efectosaleatorios} \\ \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} \text{block}_1 \\ \text{block}_2 \\ \text{block}_3 \end{pmatrix} \end{array} \right]$$

Part of the output is shown in Table 3.15. Subsection (a) shows the estimated variance components due to blocks, and for conditional observations, the effect of the blocks is $\hat{\sigma}_{\text{block}}^2 = 11.2778$, whereas the mean squared error (MSE) is $\hat{\sigma}^2 = 18.2778$. On the other hand, the fixed effects solution obtained with the “solution” option of the parameters is provided in part by (b). The analysis of variance (part (c)) indicates that there is a significant difference between treatments ($P = 0.0045$), and so the null hypothesis must be rejected ($H_0 : \mu_1 = \mu_2 = \mu_3$).

The means estimated with the “estimate” statement, given that the mean block effect is zero, the mean and the best linear unbiased predictor for each treatment are similar, as shown in Table 3.16 (part (a)). Subsections (b) and (c) show the means and differences between two means estimated with the “lsmeans” statement and the “diff” option.

Table 3.15 Variance components, fixed effects, and fixed effects test

(a) Covariance parameter estimates						
Cov Parm		Estimate	Standard error			
<i>BLOCK</i>		11.2778	17.8966			
<i>Residual</i>		18.2778	12.9243			
(b) Solutions for fixed effects						
Effect	Trt	Estimate	Standard error	DF	t-value	Pr > t
<i>Intercept</i>		41.6667	3.1388	2	13.27	0.0056
<i>Trt</i>	1	7.3333	3.4907	4	2.10	0.1036
<i>Trt</i>	2	-18.0000	3.4907	4	-5.16	0.0067
<i>Trt</i>	3	0.	..	.		
(c) Type III tests of fixed effects						
Effect	Num DF	Den DF	F-value	Pr > F		
<i>Trt</i>	2	4	27.89	0.0045		

3.4 Marginal and Conditional Models

The process of analyzing a dataset has two main objectives: the first is model selection, which aims to find well-fitting parsimonious models for the responses being measured, and the second is model prediction, where estimates from the selected models are used to predict quantities of interest and their uncertainties.

The differences that may arise in this analysis process are mainly due to the choice of unidentifiable constraints on random effects. To compare two different models, we must compare analogous quantities. Different constraints can lead to apparently extremely different but inferentially identical models. The conditional model is believed to be the basic model, and any conditional model leads to a specific marginal model. Lee and Nelder (2004) proposed and worked on conditional models derived from generalized hierarchical linear models (GHLMs) and marginal models derived from these conditional models. Marginal models have often been fitted using generalized estimating equations (GEEs), the drawbacks of which are also discussed.

3.4.1 Marginal Versus Conditional Models

Consider two models with a normal distribution: one is a random effects model (a mixed model)

$$y_{ij} = \mu + \tau_i + b_j + \varepsilon_{ij}$$

where $b_j \sim N(0, \sigma_b^2)$ and $\varepsilon_{ij} \sim N(0, \sigma^2)$. The other is a marginal model

Table 3.16 Estimated means, best linear unbiased estimates (BLUEs), and BLUPs for treatment and the difference between two means

(a) Estimates						
Label	Estimate	Standard error	DF	<i>t</i> -value	Pr > <i>t</i>	
lsm trt1	49.0000	3.1388	4	15.61	<0.0001	
lsm trt2	23.6667	3.1388	4	7.54	0.0017	
lsm trt3	41.6667	3.1388	4	13.27	0.0002	
blup trt1	49.0000	2.4683	4	19.85	<0.0001	
blup trt2	23.6667	2.4683	4	9.59	0.0007	
blup trt3	41.6667	2.4683	4	16.88	<0.0001	
(b) Trt least squares means						
Trt	Estimate	Standard error	DF	<i>t</i> -value	Pr > <i>t</i>	
1	49.0000	3.1388	4	15.61	<0.0001	
1	23.6667	3.1388	4	7.54	0.0017	
2	41.6667	3.1388	4	13.27	0.0002	
(c) Differences of Trt least squares means						
Trt	_Trt	Estimate	Standard error	DF	<i>t</i> -value	Pr > <i>t</i>
1	2	25.3333	3.4907	4	7.26	0.0019
1	3	7.3333	3.4907	4	2.10	0.1036
2	3	-18.0000	3.4907	4	-5.16	0.0067

$$E(y_{ij}) = \mu + \tau_i$$

where the elements in $V(y) = \Sigma$ are variances and covariances that have an arbitrary correlation structure. Zeger et al. (1988) pointed out that given a marginal model, the generalized estimating equations are consistent. An obvious advantage of using random effects models is that they allow conditional inferences in addition to marginal inferences (Robinson 1991). Using the model with random effects, we can obtain not only the conditional mean

$$\mu_{ij}^C = E(y_{ij}|b_j) = \mu + \tau_i + b_j$$

but also the marginal mean

$$\mu_{ij} = E(\mu_{ij}^C) = E(y_{ij}|b_j) = \mu + \tau_i$$

whereas with the marginal model, we can obtain only the marginal mean μ_{ij} .

It may be reasonable to assume that the unobservable characteristic of the random effects of blocks (b_j) follows a certain distribution. However, the center of this distribution cannot be identified because it is confounded with the intercept. Therefore, in the random effects model, we put the unidentifiable constraints $E(b_i) = 0$ and $E(\varepsilon_{ij}) = 0$ as we do for error terms in linear models. In the mixed model, these restrictions are $\sum_j \hat{b}_j = 0$ and $\sum_j \hat{\varepsilon}_{ij} = 0$ in any estimation procedure. First, we

Table 3.17 Mortality of coffee seedling clones (C) in different substrates (S)

Block	S	C	Mortality	Pct	Block	S	C	Mortality	Pct
1	3	1	3.33	0.0333	3	3	1	6.6	0.066
1	3	3	16	0.16	3	3	2	10	0.1
1	3	2	16	0.16	3	3	3	56.6	0.566
1	1	1	3.33	0.0333	3	2	1	3.3	0.033
1	1	3	6.6	0.066	3	2	2	26.6	0.266
1	1	2	3.3	0.033	3	2	3	40	0.4
1	2	1	10	0.1	3	4	1	3.3	0.033
1	2	3	3.33	0.0333	3	4	2	46	0.46
1	2	2	3.33	0.0333	3	4	3	33.3	0.333
1	4	1	3.33	0.0333	3	1	1	6.6	0.066
1	4	3	16	0.16	3	1	2	43.3	0.433
1	4	2	13.3	0.133	3	1	3	50	0.5
2	4	3	3.3	0.033	4	4	1	33	0.33
2	4	1	3.3	0.033	4	4	2	10	0.1
2	4	2	20	0.2	4	4	3	23.3	0.233
2	1	3	10	0.1	4	2	3	50	0.5
2	1	1	3.33	0.0333	4	1	2	23.3	0.233
2	1	2	6.6	0.066	4	1	3	6.6	0.066
2	2	3	36.6	0.366	4	2	1	16	0.16
2	2	1	26.6	0.266	4	2	2	10	0.1
2	2	2	43.3	0.433	4	2	3	16	0.16
2	3	3	3.3	0.033					

consider the case in which the data follow a normal distribution. We then briefly discuss how the results differ for data with a non-normal distribution.

3.4.2 Normal Distribution

Example The effect of different substrates (factor A), i.e., three substrates made from vermicompost and one from compost, on the development of physiological variables and mortality of cuttings of three clones (factor B) of robusta coffee (*Coffea canephora* p.) was evaluated. The levels of factor A are randomly assigned to rows in each block, with the following restriction: each block receives levels A1, A2, A3, and A4 and each level of factor B (B1, B2, and B3) is randomly assigned to each level of factor A in each block. The data for this experiment are tabulated in Table 3.17.

Note that while there are two randomization processes, there are effectively three sizes of experimental units: rows for A levels, columns for B levels, and row–column intersections for A × B combinations. Thus, the experimental design used was a complete randomized design with a strip-plot treatment arrangement.

The model, for these data, is given below:

$$y_{ijk} = \mu + b_k + \alpha_i + (ab)_{ik} + \beta_j + (\beta b)_{jk} + (\alpha\beta)_{ij} + \varepsilon_{ijk}$$

where y_{ijk} is the k th response observed at the i th level of factor A and at the j th level of factor B, μ is the overall mean, b_k is the random effect due to blocks assuming $b_k \sim N(0, \sigma_b^2)$, α_i is the fixed effect due to substrate type (S), $(ab)_{ik}$ is the random effect due to the interaction of a substrate with blocks assuming $(ab)_{ik} \sim N(0, \sigma_{ab}^2)$, β_j is the fixed effect due to the coffee clone type (C), $(\beta b)_{jk}$ is the random effect due to the interaction of a coffee clone with blocks assuming $(\beta b)_{jk} \sim N(0, \sigma_{\beta b}^2)$, $(\alpha\beta)_{ij}$ is the interaction fixed effect between a substrate and a coffee clone, and ε_{ijk} is the normal random error $\varepsilon_{ijk} \sim N(0, \sigma^2)$. The components of the model for this dataset are as follows:

Linear predictor: $\eta_{ijk} = \mu + b_k + \alpha_i + (ab)_{ik} + \beta_j + (\beta b)_{jk} + (\alpha\beta)_{ij}$

Distributions: $y_{ijk} | b_k, (ab)_{ik}, (\beta b)_{jk} \sim N(\mu_{ijk}, \sigma^2)$

$$b_k \sim N(0, \sigma_b^2); (ab)_{ik} \sim N(0, \sigma_{ab}^2); (\beta b)_{jk} \sim N(0, \sigma_{\beta b}^2)$$

Link function: $\eta_{ijk} = \mu_{ijk}$

The following GLIMMIX syntax sets a GLMM with a normal response.

```
proc glimmix;
class block s c;
model y=s|c;
random intercept s w/subject=block;
lsmeans s*c / slicediff=s;
run;
```

Part of the results of this analysis is shown below. The estimated variance components for blocks, block \times substrate, blocks \times clone, and the MSE are $\hat{\sigma}_b^2 = 23.4714$, $\hat{\sigma}_{ab}^2 = 35.4995$, $\hat{\sigma}_{\beta b}^2 = 67.0160$ and $\hat{\sigma}^2 = \text{CME} = 139.58$, respectively, which are listed in part (a) of Table 3.18. However, the fixed effects tests for both factors and the interaction (part (b)) are not statistically significant.

According to the “slicediff = s” option in the “lsmeans” statement, Table 3.19 shows the simple effects of each substrate level at varying clone levels.

3.4.3 Non-normal Distribution

Example Using the data in Table 3.17 but under a beta distribution, the components of the GLMM change slightly:

Table 3.18 Estimated variance components and type III tests of fixed effects

(a) Covariance parameter estimates				
Cov Parm	Subject	Estimate	Standard error	
Intercept	Block	23.4714	58.9336	
S	Block	35.4995	44.9134	
C	Block	67.0160	64.0909	
Residual		139.58	49.9056	
(b) Type III tests of fixed effects				
Effect	Num DF	Den DF	F-value	Pr > F
S	3	8.318	0.48	0.7076
C	2	5.935	1.68	0.2650
S*C	6	16.65	0.44	0.8441

Table 3.19 Simple effect comparisons across substrate levels

Simple effect comparisons of S*C least squares means by S							
Simple effect level	C	_C	Estimate	Standard error	DF	t- value	Pr > t
S 1	1	2	-12.9814	10.9667	15	-1.18	0.2549
S 1	1	3	-12.1564	10.9667	15	-1.11	0.2851
S 1	2	3	0.8250	10.1636	15	0.08	0.9364
S 2	1	2	-6.8325	10.1636	15	-0.67	0.5116
S 2	1	3	-15.2779	9.8708	15	-1.55	0.1425
S 2	2	3	-8.4454	9.8708	15	-0.86	0.4057
S 3	1	2	-4.4620	12.8219	15	-0.35	0.7327
S 3	1	3	-19.0350	12.8124	15	-1.49	0.1581
S 3	2	3	-14.5730	11.4417	15	-1.27	0.2222
S 4	1	2	-11.5925	10.1636	15	-1.14	0.2719
S 4	1	3	-8.2425	10.1636	15	-0.81	0.4301
S 4	2	3	3.3500	10.1636	15	0.33	0.7463

Distributions: $y_{ijk} | b_k, (\alpha b)_{ik}, (\beta b)_{jk} \sim \text{Beta}(\mu_{ijk}, \phi)$, where ϕ is the scale parameter.

$$b_k \sim N(0, \sigma_b^2); (\alpha b)_{ik} \sim N(0, \sigma_{\alpha b}^2); (\beta b)_{jk} \sim N(0, \sigma_{\beta b}^2)$$

Linear predictor: $\eta_{ijk} = \mu + b_k + \alpha_i + (\alpha b)_{ik} + \beta_j + (\beta b)_{jk} + (\alpha\beta)_{ij}$

Link function: $\eta_{ijk} = \text{logit}(\mu_{ijk})$

The following GLIMMIX syntax sets a beta response variable.

```
proc glimmix method=laplace;
class block s c;
model pct=s|c/dist=beta;
random intercept s w/subject=block;
lsmeans s*c/plot=meanplot (sliceby=s join) slicediff=s ilink;
run;
```

Table 3.20 Variance components and the fixed effects test

(a) Covariance parameter estimates				
Cov Parm	Subject	Estimate	Standard error	
Intercept	block	0.06723	0.1617	
S	block	0.1594	0.1420	
C	block	0.1932	0.1687	
Scale ($\hat{\phi}$)		16.6041	5.6153	
(b) Type III tests of fixed effects				
Effect	Num DF	Den DF	F-value	Pr > F
S	3	8	0.74	0.5584
C	2	6	2.60	0.1540
S*C	6	15	0.59	0.7303

Table 3.21 Simple effect comparisons across substrate levels

Simple effect comparisons of S*C least squares means by S							
Simple effect level	C	_C	Estimate	Standard error	DF	t-value	Pr > t
S 1	1	2	-0.9698	0.6578	15	-1.47	0.1611
S 1	1	3	-0.9298	0.6696	15	-1.39	0.1852
S 1	2	3	0.03999	0.5457	15	0.07	0.9426
S 2	1	2	-0.4407	0.5432	15	-0.81	0.4299
S 2	1	3	-0.8555	0.5237	15	-1.63	0.1231
S 2	2	3	-0.4149	0.4911	15	-0.84	0.4115
S 3	1	2	-0.4588	0.8285	15	-0.55	0.5879
S 3	1	3	-1.3224	0.7773	15	-1.70	0.1095
S 3	2	3	-0.8636	0.6375	15	-1.35	0.1955
S 4	1	2	-0.9880	0.5619	15	-1.76	0.0991
S 4	1	3	-0.7138	0.5739	15	-1.24	0.2326
S 4	2	3	0.2741	0.5261	15	0.52	0.6099

Some of the SAS output from this analysis is shown below. The variance components estimated for blocks, block \times substrate, blocks \times clone, and the scale parameter are $\hat{\sigma}_b^2 = 0.06723$, $\hat{\sigma}_{ab}^2 = 0.1594$, $\hat{\sigma}_{bb}^2 = 0.1932$, and with scale parameter $\hat{\phi} = 16.6041$, respectively, which are listed in part (a) of Table 3.20. However, the fixed effects tests for both factors and the interaction (part (b)) are not statistically significant. Unlike a normal distribution (the previous example), the variance components (multiplied by 100) under the beta distribution are smaller, and the type III fixed effects test is closer to be significant.

Table 3.21 shows, for each substrate level at varying clone levels, the estimates (linear predictors) of the simple effects. These effects differ from the previous results, but this is mainly because in a GLMM, these values correspond to the linear predictors estimated at the model scale and not to the estimated means at the data

Table 3.22 Total yields (grams) of barley varieties in 12 independent trials

Location	Variety				
	Manchuria	Svansota	Velvet	Trebi	Peatland
1	81.0	105.4	119.7	109.7	98.3
1	80.7	85.3	80.4	87.2	84.2
2	146.6	142.0	150.7	191.5	145.7
2	100.4	115.5	112.2	147.4	108.1
3	82.3	77.3	78.4	131.3	896
3	103.1	105.1	116.5	139.9	129.6
4	119.8	121.4	124.0	140.8	124.8
4	98.9	61.9	96.2	125.5	75.7
5	98.9	89.0	69.1	89.3	104.1
5	66.4	49.9	96.7	61.9	80.3
6	86.9	77.1	78.9	101.8	96.0
6	67.7	66.7	67.4	91.8	94.1

scale (Example 3.4.2). It is also important to note that the degrees-of-freedom correction in the estimation of means cannot yet be used in the estimation of a GLMM.

3.5 Exercises

Exercise 3.5.1 The data in the Table 3.22 below show the yield of five barley varieties in a randomized complete block experiment conducted in Minnesota (Immer et al. 1934).

- Write a complete description of the statistical model associated with this study and the assumptions of this model.
- Compute the ANOVA for the design model according to part (a) and determine whether there is a significant difference in the varieties.
- Use the least significance difference (LSD) method to make pairwise comparisons of variety mean yields.

Exercise 3.5.2 Lew (2007) conducted an experiment to determine whether cultured cells respond to two drugs (chemical formulations). The experiment was conducted using a cell culture line placed in Petri dishes. Each experimental trial consisted of three Petri dishes: one treated with drug 1, one treated with drug 2, and one untreated as a control. The data are shown in the following Table 3.23:

- Write a complete description of the statistical model associated with this study and the assumptions of this model.
- Analyze the data using a completely randomized design. Is there a significant difference between the treatment groups?

Table 3.23 Number of cells cultured in different drugs

	Control	Drug 1	Drug 2
Ensayo 1	1147	1169	1009
Ensayo 2	1273	1323	1260
Ensayo 3	1216	1276	1143
Ensayo 4	1046	1240	1099
Ensayo 5	1108	1432	1385
Ensayo 6	1265	1562	1164

- (c) Analyze the data as a randomized complete block design, where the number of trials represents a blocking factor.
- (d) Is there any difference in the results obtained in (a) and (b)? If so, explain what might be the cause of the difference in results and what method would you recommend?

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



Chapter 4

Generalized Linear Mixed Models for Non-normal Responses



4.1 Introduction

Generalized linear mixed models (GLMMs) have been recognized as one of the major methodological developments in recent years, which is evidenced by the increased use of such sophisticated statistical tools with broader applicability and flexibility. This family of models can be applied to a wide range of different data types (continuous, categorical (nominal or ordinal), percentages, and counts), and each is appropriate for a specific type of data. This modern methodology allows data to be described through a distribution of the exponential family that best fits the response variable. These complex models were not computationally possible up until recently when advances in statistical software have allowed users to apply GLMMs (Zuur et al. 2009; Stroup 2012; Zuur et al. 2013). Researchers in fields other than statistical science are also interested in modeling the structure of data. For example, in the social sciences there have been applications in the field of education when several tests are applied to students; in longitudinal personality studies when the occurrence of an emotion is repeatedly observed over time over a set of people; and in surveys to investigate the political preference of a population, among others.

Likewise, agriculture and life sciences are other major areas, where the measurement of response variables depart from the conventionally used classical methodology based on “normality” to model or describe the data set, i.e., data that generally fall within the nominal, ordinal or interval (continuous) scales of measurement. In a GLMM, the data response does not undergo any transformation, but, instead, the response is modeled as a function of the expected value through a linear relationship with the explanatory variables. GLMMs, a powerful tool, allow proper modeling of variations between groups and between space and time, leading to accuracy in the modeling of the observed data as well as in the estimation of variance components.

4.2 A Brief Description of Linear Mixed Models (LMMs)

Before addressing GLMMs, we present a brief overview of linear mixed models (LMMs). An LMM is a model whose response variable is normal and assumes: (1) that the relationship between the mean of the dependent variable (y) and fixed and random effects can be modeled as a linear function; (2) that the variance is not a function of the mean; and (3) that random effects follow a normal distribution.

The classic representation of an LMM in the matrix form, is

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b} + \boldsymbol{\varepsilon} \quad (4.1)$$

where \mathbf{y} is the vector ($n \times 1$) of the response variable; \mathbf{X} is the design matrix ($n \times (p + 1)$) of fixed effects with rank k ; $\boldsymbol{\beta}$ is the vector of unknown parameters ($(p + 1) \times 1$); \mathbf{Z} is the design matrix ($n \times q$) of random effects; and \mathbf{b} is the vector of unknown parameters of random effects ($q \times 1$), assuming that the vector of random effects \mathbf{b} follows a normal distribution with mean $\mathbf{0}$ and variance matrix \mathbf{G} , that is, $\mathbf{b} \sim N(\mathbf{0}, \mathbf{G})$. Finally, $\boldsymbol{\varepsilon}$ is the error vector with a normal distribution with mean $\mathbf{0}$ and a variance–covariance matrix ($\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \mathbf{R})$); both vectors \mathbf{b} and $\boldsymbol{\varepsilon}$ are assumed to be independent of each other.

Model 4.1, as previously mentioned, can be described in terms of a probability distribution in two ways: the first is the marginal model $\mathbf{y} \sim N(E[\mathbf{y}] = \mathbf{X}\boldsymbol{\beta}, \text{Var}[\mathbf{y}] = \mathbf{V} = \mathbf{Z}\mathbf{G}\mathbf{Z}' + \mathbf{R})$, where the mean is based solely on the fixed effects, and the parameters describing the random effects are contained in the variance and covariance matrix \mathbf{V} (Littell et al. 2006), while the second form is the conditional model $\mathbf{y} \mid \mathbf{b} \sim N(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b}, \mathbf{R})$. Under normality assumptions, both models are exactly the same and hence produce the same solution, whereas when normality is not satisfied, the models produce different solutions (Stroup 2012).

4.3 Generalized Linear Mixed Models

Most datasets in agricultural, biological, and social sciences often fall outside the scope of the traditional methods taught in introductory statistics and statistical methods. Often, these data (response variables) are: (a) binary (the presence or absence of a trait of interest, success or failure, the infection status of an individual, or the expression of a genetic disorder); (b) proportional (the ratio of females to males, infection or mortality rates within a group of individuals); or (c) counts (the number of emerging seedlings, the number of sprouts, etc.), where basic statistical methods attempt to quantify the effects of each predictor variable. However, often, studies of these experiments involve random effects, the purpose of which is to quantify variation among individuals or units. The most common random effects are blocks in experimental or observational studies that are replicated across sites (locations or environments) or over time. Random effects also encompass variations

among individuals (when measuring multiple responses per individual such as survival of multiple offspring or sex ratios of multiple offspring), genotypes, species, and regions or periods over time.

GLMMs are a powerful class of statistical tools that combine the concepts and ideas of generalized linear models (GLMs) with linear mixed models (LMMs). That is, a GLMM is an extension of the GLM, in which the linear predictor contains random effects in addition to fixed effects. These models handle a wide range of both response distributions and scenarios in which observations are sampled. GLMMs extend the theory of LMMs to response variables that have a non-normal distribution. In GLMMs, the response data are not transformed; instead, the explanatory variables are expressed as a linear relationship through a function g of the expectation of $\mathbf{y} \mid \mathbf{b}$; that is, the response is conditional on random effects. This performs the link function that relates the response to the explanatory variables in a linear manner, thus allowing the use of standard LMM techniques for estimation and hypothesis testing.

A conditional model is used to describe a GLMM with non-Gaussian errors (Model 4.1), given a link function (g), as shown below:

$$g(\boldsymbol{\eta}) = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b},$$

which is a function of the conditional expectation given by

$$E[\mathbf{y} \mid \mathbf{b}] = g^{-1}(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b}) = g^{-1}(\boldsymbol{\eta}) = \boldsymbol{\mu} \quad (4.2)$$

where $g^{-1}(\cdot)$ is the inverse link and the other terms have already been mentioned earlier. The fixed and random effects are combined to form the conditional linear predictor

$$\boldsymbol{\eta} = g(E[\mathbf{y} \mid \mathbf{b}]) = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b} \quad (4.3)$$

The relationship between the linear predictor and the vector of observations is modeled as follows:

$$\mathbf{y} \mid \mathbf{b} \sim (g^{-1}(\boldsymbol{\eta}), \mathbf{R}) \quad (4.4)$$

The above notation (4.4) expresses the conditional distribution of \mathbf{y} , given \mathbf{b} has a mean $g^{-1}(\boldsymbol{\eta})$ and variance \mathbf{R} . Note that instead of specifying the distribution for \mathbf{y} as in the case of a GLM, we specify a distribution for the conditional response $\mathbf{y} \mid \mathbf{b}$.

The variance and covariance matrix for the observations is given by:

$$V(\mathbf{y}) = E[V(\mathbf{y} \mid \mathbf{b})] + V(E[\mathbf{y} \mid \mathbf{b}]) = \mathbf{A}^{1/2} \mathbf{R} \mathbf{A}^{1/2} + \mathbf{Z} \mathbf{G} \mathbf{Z}' \quad (4.5)$$

where matrix \mathbf{A} is a diagonal matrix containing the variance functions of the model. GLMMs cover an important group of statistical models, such as:

- (a) Linear models (LMs): absence of random effects, identity link function and the assumption of a normal distribution.
- (b) Generalized linear models (GLMs): random effects are absent, link function is different from the identity function, and the response variables are non-normally distributed.
- (c) Linear mixed models (LMMs): presence of random effects, identity link function and normal distribution assumed for the response variable.

GLMMs have been formulated to correct the shortcomings of LMMs, as there are many cases where the assumptions made in linear mixed models are inadequate. First, an LMM assumes that the relationship between the mean of the dependent variable (y) and fixed and random effects (β, b) can be modeled through a linear function. This assumption is questionable, like when a researcher wishes to model the incidence of a disease or the success or failure of an event.

The second assumption of an LMM is that variance is not a function of the mean and that the random effects follow a normal distribution. The assumption of constant variance is not met when the response variable is binary (1, 0). In this case, the variance is $\pi(1 - \pi)$, which is a function of the mean. The result is a random variable, which can take two values (0, 1); in contrast, the normal distribution can take any real number. Finally, the predictions for an LMM can take any real value, whereas the predictions for a binary variable are bounded in the interval (0, 1), since it is a probability and this prediction cannot support negative values.

Historically, a number of options have been used to address and solve some LMM problems, even though their use is not the most appropriate. These include applying logarithmic transformations ($\log(y)$), transformations using the square root (\sqrt{y}), arcsine transformations ($\text{seno}^{-1}(y)$), and so on. However, many of these transformations use linear mixed models by ignoring the fact that these models are not the most accurate, despite being aware that the response variable does not satisfy the assumption of normality. These options are attractive because they are relatively simple and easy to implement using the LMM machinery. However, they circumvent the problem that a linear mixed model is not the best model for analyzing data.

4.4 The Inverse Link Function

In a GLMM, the canonical link function maps the original data to the linear predictor of the model $g(\eta) = X\beta + Zb$. This linear predictor can be transformed to an observed data scale through an inverse link function. In other words, the inverse link function is used to map the value of the linear predictor for the i th observation to the conditional mean at the data scale η_i . For example, suppose that we are conducting an experiment in which we are assessing the number of undesirable weeds observed in a crop of interest after the application of a certain number of treatments; the response variable is assumed to have a Poisson distribution with a mean λ_{ij} , the linear predictor of which is given by

$$\eta_{ij} = \eta + \tau_i + b_j$$

where η is the intercept, τ_i is the fixed effect due to treatments, and b_j is the random effect assuming $b_j \sim N(0, \sigma_b^2)$.

To obtain the inverse function of the following predictor

$$\log(\lambda_{ij}) = g(\eta_{ij}) = \eta + \tau_i + b_j,$$

we proceed by exponentiating both sides of the previous equation, with which we obtain the inverse function of the link shown below:

$$\lambda_{ij} = e^{\eta + \tau_i + b_j},$$

which is denoted as $g^{-1}(g(\eta_{ij})) = g^{-1}(\eta + \tau_i + b_j)$.

Therefore, for this example, λ_{ij} depends on the linear predictor through the inverse link function and the variance σ_{ij}^2 depends on λ_{ij} through the variance function.

4.5 The Variance Function

The variance function is used to model the inconsistent variability of the phenomenon under study. With GLMMs, the residual variability arises from two sources, namely, the variability of the distribution of sampling units in an experimental arrangement (blocks, plots, locations, etc.) and the variability due to overdispersion. Overdispersion can be modeled in several ways. When dealing with a GLMM, the scale parameter or the dispersion parameter ϕ is extremely important since it can either increase or decrease the variance in the model for each observation.

$$\text{Var}(y_{ij}|b_j) = \phi \text{Var}(\eta_{ij})$$

If overdispersion exists, one way to remove it is to add the random effects (in SAS `_residual_`) of each observation to the linear predictor. Another alternative is to use another distribution to model the dataset; for example, the two-parameter negative binomial (NB) distribution (η_{ij}, ϕ) instead of the single-parameter Poisson distribution (λ_{ij}) in the case of count data.

4.6 Specification of a GLMM

A GLMM is composed of three parts: (1) fixed effects that convey systematic and structural differences in responses; (2) random effects that convey stochastic differences between blocks or other random factors, as these effects allow generalizations

Table 4.1 Common distributions with their respective link functions

Distribution	Link function	Distribution syntax	Syntax of the link function
Binomial	Logit or probit	dist = binomiallbinlb	link = logit or probit
Poisson	Log	dist = PoissonlPoi	link = log
Beta	Logit	dist = beta	link = logit
Normal	Identity	dist = normallgaussian	link = identitylid
Negative binomial	Log	dist = negbinomiallnegbinl nb	link = log
Multinomial	Cumulative logit	dist = multinomiallmulti	link = cumlogitlclogit

of the population from which the sampling units have been (randomly) sampled; and (3) distribution of errors. Thus, a complete definition of a GLMM is as follows:

$$\mathbf{y} \mid \boldsymbol{\mu} \sim f(\boldsymbol{\mu}, \boldsymbol{\phi}) \text{ (conditional distribution)}$$

$$\boldsymbol{b} \sim N(0, \boldsymbol{G}) \text{ (random effects)}$$

$$g(\boldsymbol{\mu}) = \boldsymbol{\eta} \text{ (link function)}$$

$$\boldsymbol{\eta} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{b} \text{ (linear predictor)}$$

where the distribution function $f(\cdot)$ is a member of the exponential family, $g(\boldsymbol{\mu})$ is the linear function, \boldsymbol{X} and \boldsymbol{Z} are the design matrices, and $\boldsymbol{\beta}$ and \boldsymbol{b} are the unknown parameters for fixed and random effects, respectively.

When fitting a GLMM, the data remain on the original measurement scale (data scale). However, when means are estimated from a linear function of the explanatory variables (the predictor), these means are on the model scale. A link function is used to link the model scale back to the original data scale. This is not the same as transforming the original measurements to a different measurement scale. For example, applying the log transformation for counts followed by an analysis of variance (ANOVA) under a normal distribution is not the same as fitting a generalized linear model, assuming a Poisson distribution and using a log link (Gbur et al. 2012). In the first case, the least squares means would normally be equal to the arithmetic means, whereas in the second case, the means are inversely linked to the data scale, which may not be equal to the arithmetic means of the original sample.

The distribution specifications in “proc GLIMMIX” have default link functions, but it is always highly recommended to explicitly code the link function, since for some type of response variable, more than one alternative exists. This way, there is no doubt that an appropriate function was used. Using the wrong link function will lead to totally meaningless and incorrect results. Table 4.1 shows some common distributions, the appropriate link function, and the proper syntax for each.

For a complete list, see the online Statistical Analysis Software (SAS/STAT) documentation for PROC GLIMMIX.

4.7 Estimation of the Dispersion Parameter

The overall measures of fit compare the observed values of the response variable with fitted (predicted) values. The dispersion parameter is unknown and therefore must be estimated. There are two methods for estimating the overdispersion parameter. McCullagh (1983) proposed estimating overdispersion as follows:

$$\phi = \frac{(\mathbf{y} - \boldsymbol{\mu})' \mathbf{V}_{\boldsymbol{\mu}}^{-1} (\mathbf{y} - \boldsymbol{\mu})}{N - p} = \frac{\text{Pearson's } \chi^2}{N - p}$$

where $\mathbf{V}_{\boldsymbol{\mu}}^{-1}$ is the diagonal matrix of the variance functions and $N - p$ is the degree of freedom for lack of fit. Later, McCullagh and Nelder (1989) suggested using deviance

$$\hat{\phi} = \frac{\text{Deviance}}{N - p} = \frac{-2[\ln(LM_1) - \ln(L(M_2))]}{N - p}$$

Deviance is a global fit statistic that also compares fitted and observed values; however, its exact function depends on the likelihood function of the random component of the model. Deviance compares the maximum value of the likelihood function of a model, like M_1 , with the maximum possible value of the likelihood function that is calculated using data. When data are used in the likelihood function, the model is saturated and has as many parameters as possible. Thus, M_2 is saturated and has as many parameters as the data. Model M_2 tries to fit the data and gives the highest possible value for the likelihood.

If the overdispersion parameter is significantly greater than one, this indicates that overdispersion exists; in other words, it indicates that the variance is greater than the mean. Therefore, the parameter should be used to adjust the variance. If overdispersion is not taken into account, inflated test statistics may be generated. However, when the dispersion parameter is less than 1, the test statistics are more conservative, which is not considered a big problem.

The following example is intended to show how GLIMMIX in SAS estimates the dispersion parameter in a GLMM.

Example An agronomist wants to test the effectiveness of a new herbicide offered on the market (we will denote this as herb_N) and compare it with the herbicide that has been used for several cycles (herb_C). The experimental arrangement used was a randomized complete block design as shown below (Table 4.2).

The components of a GLMM with a Poisson response variable are listed below:

$$\begin{aligned} \text{Distribution: } y_{ij} \mid b_j &\sim \text{Poisson}(\lambda_{ij}) \\ b_j &\sim N\left(0, \sigma_{\text{bloque}}^2\right) \end{aligned}$$

Table 4.2 Number of undesirable weeds per plot

Block	Herb_C	Herb_N
1	1	36
2	5	109
3	21	30
4	7	48
5	2	3
6	6	0
7	0	5
8	19	26

Linear predictor: $\eta_{ij} = \eta + \text{herbicide}_i + b_j$

Link function: $\log(\lambda_{ij}) = \eta_{ij}$

This model assumes that the slopes are the same for each herbicide. The following SAS code is used for the proposed model:

```
proc glimmix nobound method=laplace;
class block trt;
model count = trt/dist=poisson link=log;
random block;
lsmeans trt/ilink lines;
run;
```

Explanation The “method = ”option is used to specify the method used to optimize the logarithm of the likelihood function. In “proc GLIMMIX,” there are two popular methods: adaptive quadrature (quad) or Laplace (laplace), which are the preferred methods for categorical response variables. Both of these methods fit a conditional model. When the quadrature method is used (method = quad), subjects (individuals) must be declared in the random effects (e.g., for the above program, “random intercept/subject=block”). In addition, processing random effects by subject is more efficient than using the syntax “random block” random effects in blocks. The “dist” option is where you specify the probability distribution that is appropriate for the type of response; in this case, it is the Poisson distribution. The “link” option is for specifying the link function of the distribution. The “ddfm” option is omitted so that GLIMMIX uses – by default – the method for calculating the denominator degrees of freedom for the fixed effects tests that result from the model. The “ilink” option converts the estimates of the treatment means (lsmeans) on the model scale to the data scale. Finally, “proc GLIMMIX” supports the “lines” option, which adds letter groups to the mean differences resulting from using “lsmeans.”

The most relevant parts of the SAS output, for the purposes of what we want to show, are shown in Tables 4.3 and 4.4. The fit statistics of the fitted model are shown in part (a) and part (b) of Table 4.3. The $-2 \log$ likelihood statistic is extremely

Table 4.3 Fit statistics and variance components

(a) Fit statistics (Akaike’s information criterion (AIC), a small sample bias corrected Akaike’s information criterion (AICC), Bozdogan Akaike’s information criterion (CAIC), Schwarz’s Bayesian information criterion (BIC), Hannan and Quinn information criterion (HQIC))		
–2 Log likelihood	175.35	
AIC (smaller is better)	181.35	
AICC (smaller is better)	183.35	
BIC (smaller is better)	181.59	
CAIC (smaller is better)	184.59	
HQIC (smaller is better)	179.74	
(b) Fit statistics for conditional distribution		
–2 Log L (count r. effects)	139.03	
Pearson’s chi-square	77.56	
Pearson’s chi-square/degree of freedom (DF)	4.85	
(c) Covariance parameter estimates		
Cov Parm	Estimate	Standard error
Block	1.5590	0.8690

Table 4.4 Type III fixed effects tests and estimated least squares means

(a) Type III tests of fixed effects							
Effect	Num DF	Den DF	F-value	Pr > F			
Herbicide	1	7	101.34	<0.0001			
(b) Trts least squares means							
Trts	Estimate	Standard error	DF	t-value	Pr > t	Mean	Standard error mean
Herb_C	1.4604	0.4696	7	3.11	0.0171	4.3076	2.0227
Herb_N	2.8947	0.4561	7	6.35	0.0004	18.0778	8.2447

useful for comparing nested models, whereas the different versions of information criteria that exist, such as Akaike information criterion (AIC), Akaike’s information criteria with small sample bias correction (AICC), Bayesian information criterion (BIC), Bozdogan Akaike’s information criteria (CAIC), and Hannan and Quinn information criteria (HQIC), are useful when comparing models that are not necessarily nested (subsection (a)). The table of fit statistics for the conditional distribution shows the sum of the independent contributions to the conditional (part (b)) –2 log likelihood, the value of which is 139.03, whereas the value of Pearson’s statistic divided by the degrees of freedom for the conditional distribution (Pearson’s chi – square/DF) is 4.85.

The estimated dispersion parameter ($\phi = \text{Pearson’s chi-square/DF}$) has a value far from 1; in this case, it is $\hat{\phi} = 4.85$, which indicates that there is a strong overdispersion. This may be because the specified distribution of the data is not appropriate, the counts are too small, or the variance function was not correctly

specified. The estimate of the variance component due to a block is tabulated in part (c) of Table 4.2, the estimated value of which is $\hat{\sigma}_{\text{block}}^2 = 1.559$.

The fixed effects test and least squares means are shown in Table 4.4. The type III fixed effects tests indicate that there is a highly significant difference (part (a)) in the effectiveness of herbicides in weed suppression; the estimated means with their respective standard errors are tabulated under the “Mean” column (part (b)). The “Estimate” column containing the estimates of the means of lsmeans is on the model scale. They are derived from the log likelihood function. SAS always lists the means obtained with lsmeans from the model scale when creating least squares means test tables. The “Mean” column has been converted back to the data scale using the “ilink” inverse link function. These values are estimates of the average counts for each treatment level (in this case, the herbicide type on the data scale). When we report the results, we must replace the corresponding model’s least squares values in the test tables with these estimates (means on the data scale corresponding to the values in the “Mean” column).

Since there is a strong overdispersion ($\hat{\phi} > 1$), assuming that the data have a Poisson distribution is risky because this implies that the mean and variance are equal, which is an assumption implying that the data have a Poisson distribution, i.e., that the mean and variance are the same. A useful alternative distribution might be a negative binomial distribution; this distribution has a mean λ and variance $\lambda + \lambda\phi^2$ with $\phi > 0$ commonly known as the scale parameter.

The following is the specification of the components of a GLMM with a negative binomial (NB) response variable:

$$\text{Distribution : } y_{ij} \mid b_j \sim \text{Negative binomial}(\lambda_{ij}, \phi)$$

$$b_j \sim N(0, \sigma_{\text{block}}^2)$$

$$\text{Linear predictor: } \eta_{ij} = \eta + \text{herbicide}_i + b_j$$

$$\text{Link function: } \log(\lambda_{ij}) = \eta_{ij}$$

The GLIMMIX procedure also allows modeling a GLMM with a negative binomial response variable:

```
proc glimmix data=itam nobound method=laplace;
class block trts;
model count = trts/dist=negbin;
random block;
lsmeans trts/ilink;
run;
```

Part of the output is shown in Table 4.5. The fit statistics for the model comparison (part (a)) and that for the conditional distribution (part (b)) are both provided by the GLIMMIX procedure when a conditional distribution is specified. Since in the previous analysis, it was observed that overdispersion exists when assuming a Poisson distribution, the results – under a negative binomial distribution – indicate that this overdispersion problem no longer exists; i.e., the binomial distribution is no

Table 4.5 Fit statistics under a negative binomial distribution

(a) Fit statistics		
-2 Log likelihood		120.50
AIC (smaller is better)		128.50
AICC (smaller is better)		132.13
BIC (smaller is better)		128.81
CAIC (smaller is better)		132.81
HQIC (smaller is better)		126.35
(b) Fit statistics for conditional distribution		
-2 Log L (count r. effects)		120.50
Pearson's chi-square		9.21
Pearson's chi-square/DF		0.58
(c) Fit statistics and Pearson's chi-square/DF		
	Poisson	Negative binomial
-2 Log likelihood	175.35	120.50
AIC (smaller is better)	181.35	128.50
AICC (smaller is better)	183.35	132.13
BIC (smaller is better)	181.59	128.81
CAIC (smaller is better)	184.59	132.81
HQIC (smaller is better)	179.74	126.35
$\hat{\phi}$ (Pearson's chi-square/DF)	4.85	0.58

longer overdispersed ($\hat{\phi} = 0.58$). In other words, the negative binomial distribution does a better job than the Poisson distribution in fitting these data, since it effectively controls the overdispersion.

Comparing the fit statistics tabulated in Table 4.3 subsection (c) under both distributions, we can observe that when the data are modeled under a negative binomial distribution, the values of the fit statistics are lower than those under a Poisson distribution, since the dispersion parameter $\hat{\phi} < 1$. This indicates that the negative binomial models this dataset better.

4.8 Estimation and Inference in Generalized Linear Mixed Models

4.8.1 Estimation

In GLMMs, inference involves the estimation and testing of the hypotheses of unknown parameters in β , G , and R as well as the best linear unbiased predictions (BLUPs) of random effects, b . In most modern statistical tools, including GLMMs, parameter fitting is performed via maximum likelihood (ML) or methods derived from this method. For simple analyses, in which the response variables are normal, classical ANOVA methods are based on calculating the differences of the sums of

the squares that produce the same results as an ML estimation. However, this equivalence is not obtained in models with more complex structures such as LMMs or GLMMs. To find the ML estimators, in GLMMs, one must integrate over all possible values of the random effects. For GLMMs, this computation is at best slow and at worst (a large number of random effects) computationally infeasible.

Statisticians have proposed several ways to approximate the parameter estimates of a GLMM, including penalized quasi-likelihood (PQL) and pseudo-likelihood methods (Schall 1991; Wolfinger and O'Connell 1993; Breslow and Clayton 1993), Laplace approximations (Raudenbush et al. 2000) and Gauss–Hermite quadrature (Pinheiro and Chao (2006), and Bayesian methods based on Markov chain Monte Carlo (Gilks et al. 1996). In all these approaches, researchers must distinguish between a standard ML estimation, which estimates the standard deviations of the random effects assuming that the fixed effects estimates are precisely correct, and restricted maximum likelihood (REML), a variant that averages over the uncertainty in the fixed effects parameters (Pinheiro and Bates 2000; Littell et al. 2006).

The ML method underestimates the standard deviations of random effects, except in extremely large datasets, but it is most useful for comparing models with different fixed effects. Pseudo- and quasi-likelihood methods are the simplest and the most widely used in approximating a GLMM. They are widely implemented in statistical packages that promote the use of GLMMs in many areas of ecology, biology, and quantitative and evolutionary genetics (Breslow 2004). Unfortunately, pseudo- and quasi-likelihood methods produce biases in parameter estimation if the standard deviations of the random effects are large, especially when using binary data (Rodriguez and Goldman 2001; Goldstein and Rasbash 1996). Lee and Nelder (2001) have implemented several improvements to the PQL version, but these are not available in most common statistical software packages. As a rule of thumb, PQL performs poorly for Poisson data when the average number of counts per treatment combination is less than five or for binomial data when the expected numbers of successes and failures for each observation are less than five (Breslow 2004). Another disadvantage of PQL is that it calculates a quasi-likelihood rather than the true likelihood. Because of this, many statisticians believe that PQL-based methods should not be used for inference.

There are two more accurate approximations available, which also reduce bias. One is the Laplace approximation (Raudenbush et al. 2000), which approximates the true likelihood of a GLMM instead of a quasi-likelihood, allowing the maximum likelihood method in the GLMM inference process. The other approach is called Gauss–Hermite quadrature (Pinheiro and Chao 2006), which is more accurate than the Laplace approximation but is slower (requires more computational resources). Therefore, the procedures for parameter estimation of a GLMM that are approximations are as follows:

The penalized quasi-likelihood method performs the estimation process by alternating between (1) estimating the fixed parameters by fitting a GLM with a variance–covariance matrix based on an LMM fit and (2) estimating the variances and

covariances by fitting a GLM with unequal variances calculated from the previous GLM fit. Pseudo-likelihood, a close cousin of the ML method, estimates variances differently and estimates a scale parameter to account for overdispersion (some authors use these terms interchangeably). In summary, GLMMs require an iterative process in parameter estimation. Two categories of iterative procedures are used by SAS: linearization and integral approximation. The GLIMMIX procedure uses the pseudo-likelihood method in linearization, and integral approximation uses the Laplace approximation or adaptive methods such as Gauss–Hermite quadrature. These methods maximize the log likelihood of the exponential distribution family, i.e., non-normal distributions. The pseudo-likelihood method is the default procedure in the GLIMMIX procedure (Proc GLIMMIX). The Laplace method and quadrature are an approximation for maximum likelihood, but the Laplace method is computationally simpler than quadrature and also provides excellent estimates.

4.8.2 Inference

After estimating the parameter values in a GLMM, the next step is to extract information and draw statistical conclusions from a given dataset through careful analysis of the parameter estimates (confidence intervals, hypothesis testing) and select a model that best describes or explains the most variability in the dataset. Inference can generally be based on three types: (a) hypothesis testing, (b) model comparison, and (c) Bayesian approaches. Hypothesis testing compares test statistics (F -test in ANOVA) to verify their expected distributions under the null hypothesis (H_0), estimating the value of P (P -value) to determine whether H_0 can be rejected. On the other hand, model selection compares candidate model fits. These can be selected using hypothesis testing; that is, testing nested versus more complex models (Stephens et al. 2005) or using information theory approaches such as Wald tests (Z , χ^2 , t , and F). In model selection, likelihood ratio (LR) tests can ensure the significance of factors or choose the best of a pair of candidate models. On the other hand, information criteria allow multiple comparisons and selections of non-nested models. Among these criteria are the Akaike information criterion (AIC) and related information criteria that use deviance as a measure of fit, adding a term to penalize more complex models. Information criteria can provide better estimates. Variations of AIC are highly common when sample sizes are not large (AICC), when there is overdispersion in the data (quasi-AIC, QAIC), or when one wishes to identify/determine the number of parameters in a model (Bayesian information criterion, BIC).

4.9 Fitting the Model

The mathematics behind a GLMM is quite complex. It is difficult to conceptualize the use of constructs such as distributions, link functions, log likelihood, and quasi-likelihood when fitting a model. Perhaps the following points will help explain the modeling process.

- (a) An analysis of variance model is a vector of linear predictors (equation) with unknown parameter estimates.
- (b) Each distribution has a corresponding probability function.
- (c) The vector of linear predictors is substituted into the likelihood function.
- (d) Solutions to the parameter estimates are found by minimizing the negative of the log likelihood function ($-\log$ likelihood).
- (e) The means (least squares means – lsmeans) are derived from the parameter estimates and are on the model scale.
- (f) The link function converts the mean estimates at the model scale to the original data scale.

The key concepts of proc GLIMMIX are (1) it uses a distribution to estimate the model parameters; it does not fit the data to a distribution, and (2) the data values are not transformed by the link function; the link function converts the means (least squares means) to the data scale after estimation at the model scale.

4.10 Exercises

1. As a simple example of these types of data, consider the following results of an experiment on wheat germination, carried out in pots under glass. The experiment consisted of four blocks of six treatments (Table 4.6).
 - (a) According to the response variable, what type(s) of probability distribution do you suggest for the variable?
 - (b) Construct a GLMM to study the effect of treatments on seed germination.
 - (c) Analyze the dataset according to the model proposed in (a). Is the probability distribution proposed in (a) adequate?
 - (d) Is there a significant difference in the proportion of germinated seeds between treatments?

Table 4.6 Number of seeds not germinating (out of 50)

	Trt1	Trt2	Trt3	Trt4	Trt5	Trt6
A	10	11	8	9	7	6
B	8	10	3	7	9	3
C	5	11	2	8	10	7
D	1	6	4	13	7	10

Table 4.7 Control of cockchafer larvae

	A		B		C		D		E		F		G		H	
	a	B	a	b	a	b	a	b	a	b	a	b	a	b	a	b
Trt1	3	7	7	15	5	7	5	14	0	3	1	7	1	10	4	13
Trt2	4	3	3	12	4	2	12	5	2	3	1	6	3	5	4	11
Trt3	3	10	6	12	4	4	1	14	2	2	1	7	1	8	7	10
Trt4	5	8	4	11	1	5	5	9	2	7	3	7	0	3	3	12
Trt5	4	6	4	11	2	2	3	8	0	1	5	4	1	6	1	8

2. Table 4.7 shows the counts per sample area of a variety type of cockchafer larva (two age groups a and b). The experiment consisted of five treatments in eight randomized blocks and two age groups to study the differential effects of treatments on insect age.
- (a) Considering the type of answer of this exercise; what type(s) of probability distribution(s) do you suggest for this type of response?
 - (b) Construct a GLMM to study the effect of treatments and the age of Cockchafer larvae.
 - (c) Analyze the dataset according to the model proposed in (a).
 - (d) Is the model used in (a) sufficient? If so, discuss your findings.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



Chapter 5

Generalized Linear Mixed Models for Counts



5.1 Introduction

Data in the form of counts regularly appear in studies in which the number of occurrences is investigated, such as the number of insects, birds, or weeds in agricultural or agroecological studies; the number of plants transformed or regenerated using modern breeding techniques; the number of individuals with a certain disease in a medical study; and the number of defective products in a quality improvement study, among others. These counts can be counted per unit of time, area, or volume. When using a generalized linear model (GLM) with a Poisson distribution, it is often found that there is excessive dispersion (extra variation) that is no longer captured by the Poisson model. In these cases, the data must be modeled with a negative binomial distribution that has the same mean as the Poisson distribution but with a variance greater than the mean. Most experiments have some form of structure due to the experimental design (completely randomized design (CRD), randomized complete block design (RCBD), incomplete block, or split-plot design) or the sampling design, which must be incorporated into the predictor to adequately model the data.

5.2 The Poisson Model

A Poisson distribution with parameter λ belongs to the exponential family and is a discrete random variable, whose probability function is equal to

$$f(y) = \frac{e^{-\lambda} \lambda^y}{y!}; \lambda > 0, y = 0, 1, 2, \dots$$

The mean and variance of a Poisson random variable are equal, i.e., $E(y) = \text{Var}(y) = \lambda$. A Poisson distribution is often used to model responses that are “counts.” As λ increases, the Poisson distribution becomes more symmetric and eventually it can be reasonably approximated by a normal distribution.

Let y_{ij} be the value of the count variable associated with unit i at level one and with unit j at level two, given a set of explanatory variables. Therefore, we can express this as

$$f(y_{ij}) = \frac{e^{-\lambda_{ij}} \lambda_{ij}^{y_{ij}}}{y_{ij}!}, y_{ij} = 0, 1, 2, \dots$$

and the logarithm of the likelihood is given by:

$$\log f(y_{ij}) = \log \left(\frac{e^{-\lambda_{ij}} \lambda_{ij}^{y_{ij}}}{y_{ij}!} \right) = -\lambda_{ij} + y_{ij} \log(\lambda_{ij}) - \log(y_{ij}!).$$

A Poisson distribution has very particular mathematical properties that are used when we model “counts.” For example, the expected value of y is equal to the variance of y , such that

$$E(y_{ik}) = \text{Var}(y_{ik}) = \lambda_{ij}$$

Then, λ_{ij} is necessarily a nonnegative number, which could lead to difficulties if we consider using the identity bound function in this context. The natural logarithm is mainly used as a link function for expected “counts.” For single-level (factor) data, Poisson regression model is considered, where we work with the natural logarithm of the counts, $\log(\lambda_i)$, whereas for multilevel data (more than two factors), mixed models with Poisson data are considered a better choice for the logarithm of the counts λ_{ij} .

Suppose that given the random effects of b , the counts y_1, y_2, \dots, y_n are conditionally independent such that $y_{ij} | b_j \sim \text{Poisson}(\lambda_{ij})$, where

$$\log(\lambda_{ij}) = \eta + \tau_i + b_j.$$

This is a special case of a generalized linear mixed model (GLMM) in which the link function of this family of distributions is $g(\lambda_{ij}) = \log(\lambda_{ij})$. The dispersion parameter ϕ , in this case, is equal to 1.

Sometimes, if the data counts are extremely large, their distribution can be approximated to a continuous distribution. Whereas, if all the counts are large enough, then the square root of the counts is viable for fitting the model as it allows the variance to be stabilized. However, as mentioned in previous chapters, the estimation process under normality can be problematic, as it can provide negative fitted values and predictions, which is illogical.

5.2.1 CRD with a Poisson Response

An CRD is a design in which a fixed number of t treatments is randomly assigned to r experimental units. The linear predictor describing the mean structure of this GLM is

$$\eta_{ij} = \eta + \tau_i$$

where η_{ij} denotes the ij th link function of the i th treatment in the j th observation, η is the intercept, and τ_i is the fixed effect due to treatment i ($i = 1, 2, \dots, t; j = 1, 2, \dots, r_i$), with t treatments and r_i replicates in each treatment i .

Example *Effect of a subculture on the number of shoots during micropropagation of sugarcane.*

The objective of micropropagation in sugarcane is to produce vegetative material identical to the donor so that its genetic integrity is preserved. Despite this, somaclonal variation has been observed in plants derived from in vitro culture regardless of explant, variety, ploidy level, number of subcultures, and generation route used, among others. A total of 8 explants were planted in temporary immersion bioreactors (explant/bioreactor) to determine whether the number of subcultures (10 subcultures) influences the number of shoots observed per explant. In this example, we have r_i observations ($j = 1, 2, \dots, r_i$) on each of the 10 subcultures ($i = 1, 2, \dots, 10$) in a completely randomized design (Appendix 1: Data: Subcultures). The analysis of variance (ANOVA) table (Table 5.1) for this model is given below:

The components of the GLM are set out below:

Distribution: $y_{ij} \sim \text{Poisson}(\lambda_{ij})$

Linear predictor: $\eta_{ij} = \eta + \tau_i$

Link function: $\log(\lambda_{ij}) = \eta_{ij}$

where y_{ij} denotes the number of sprouts observed in subculture i explant j ($i = 1, 2, \dots, 10; j = 1, 2, \dots, 8$), η_{ij} is the ij th link function, η is the intercept, and τ_i is the fixed effect of subculture i .

Table 5.1 Analysis of variance

Sources of variation	Degrees of freedom	
	Unbalanced design	Balanced design
Subculture	$t - 1 = 10 - 1 = 9$	$t - 1$
Error	$\sum_{i=1}^{10} r_i - t = 164$	$t(r - 1)$
Total	$\sum_{i=1}^{10} r_i - 1 = 173$	$tr - 1$

Table 5.2 Model information and estimation methods

(a) Model information	
Dataset	WORK.SUGAR
Response variable	NB
Response distribution	Poisson
Link function	Log
Variance function	Default
Variance matrix	Diagonal
Estimation technique	Maximum likelihood
Degrees of freedom method	Residual
(b) Dimensions	
Columns in X	
Columns in Z	0
Subjects (blocks in V)	1
Max Obs per subject	

The following Statistical Analysis Software (SAS) code allows analyzing an CRD with a Poisson response.

```
proc glimmix data=sugar method=laplace;
class repl sub1 ;
model nb=sub/dist=poisson s link=log;
lsmeans sub/lines ilink;
run;quit;
```

While most of the commands used have been explained before, the options in the model statement “dist,” “s,” and “link” communicate to the SAS the type of data distribution, the fixed effects solution, and the link to use, respectively. In addition, the “lines” option asks the GLIMMIX procedure in the “lsmeans” (least squares means) command for mean comparisons, and the “ilink” option provides the inverse link function.

Part of the output is shown in Table 5.2, where part (a) shows the model and the methods used to fit the statistical model, whereas part (b) lists the dimensions of the relevant matrices in the model specification.

Due to the absence of random effects in this model, there are no columns in matrix Z . The 11 columns in matrix X comprise an intercept and 10 columns for the effect of subcultures.

The goodness-of-fit statistics of the model are shown in part (a) of Table 5.3. The value of the generalized chi-squared statistic over its degrees of freedom (DFs) is less than 1. (Pearson’s chi – square/DF = 0.79). This indicates that there is no overdispersion and that the variability in the data has been adequately modeled with the Poisson distribution.

Subsection (b) of Table 5.3 shows the maximum likelihood (ML) (“Estimate”), parameter estimates, standard errors, and t -tests for the hypothesis of the parameters.

Table 5.3 Fit statistics and estimated parameters

(a) Fit statistics (Akaike’s information criterion (AIC), a small sample bias Corrected Akaike’s information criterion (AICC), Bozdogan Akaike’s information criteria (CAIC), Schwarz’s Bayesian information criterion (BIC), Hannan and Quinn information criterion (HQIC))							
–2 Log likelihood						1062.11	
Akaike information criterion (AIC) (smaller is better)						1082.11	
AICC (smaller is better)						1083.46	
Bayesian information criterion (BIC) (smaller is better)						1113.70	
CAIC (smaller is better)						1123.70	
HQIC (smaller is better)						1094.93	
Pearson’s chi-square						137.70	
Pearson’s chi-square/DF						0.79	
(b) Parameter estimates							
Effect	sub1		Estimate	Standard error	DF	t-value	Pr > t
Intercept		$\hat{\eta}$	3.6687	0.04124	164	88.96	<0.0001
sub1	1	$\hat{\tau}_1$	–1.0809	0.07389	164	–14.63	<0.0001
sub1	2	$\hat{\tau}_2$	–0.9043	0.06664	164	–13.57	<0.0001
sub1	3	$\hat{\tau}_3$	–0.5596	0.06839	164	–8.18	<0.0001
sub1	4	$\hat{\tau}_4$	–0.3412	0.06398	164	–5.33	<0.0001
sub1	5	$\hat{\tau}_5$	0.2177	0.05540	164	3.93	0.0001
sub1	6	$\hat{\tau}_6$	0.2257	0.05452	164	4.14	<0.0001
sub1	7	$\hat{\tau}_7$	0.2631	0.05178	164	5.08	<0.0001
sub1	8	$\hat{\tau}_8$	0.3387	0.05109	164	6.63	<0.0001
sub1	9	$\hat{\tau}_9$	0.2684	0.05478	164	4.90	<0.0001
sub1	10	$\hat{\tau}_{10}$	0

Table 5.4 (part (a)) shows significance tests for the fixed effects in the model “Type III fixed effects tests.” These tests are Wald tests and not likelihood ratio tests. The effect of a subculture on the number of shoots is highly significant in this model with a value of $P < 0.0001$, indicating that the 10 subcultures do not produce the same number of shoots, that is, the number of subcultures affects the average shoot production in the explant.

The least squares means obtained with “lsmeans” (part (b) in Table 5.4) are the values under the column “Estimate,” which along with the standard errors, were calculated with the linear predictor $\hat{\eta}_i = \hat{\eta} + \hat{\tau}_i$. These estimates are on the model scale, whereas the “Mean” column values and their respective standard errors are on the data scale, which were obtained by applying the inverse link to obtain the $\hat{\lambda}_i$ values, i.e., $\hat{\lambda}_i = \exp(\hat{\eta}_i)$ with their respective standard errors.

A comparison of means, using the option “lines,” is presented in Fig. 5.1. In this figure, we can see that in the first subcultures, the average production is minimal but it increases as subcultures increase from 5 to 8, and, in subculture 9, the average number of shoots per explant begins to decrease.

Table 5.4 Type III tests of fixed effects and least squares means (means)

(a) Type III tests of fixed effects							
Effect	Num DF	Den DF	<i>F</i> -value		Pr > <i>F</i>		
sub1	9	164	120.14		<0.0001		
(b) sub1 least squares means							
sub1	Estimate	Standard error	DF	<i>t</i> -value	Pr > <i>t</i>	Mean	Standard error mean
1	2.5878	0.06131	164	42.21	<0.0001	13.3000	0.8155
2	2.7644	0.05234	164	52.81	<0.0001	15.8696	0.8307
3	3.1091	0.05455	164	56.99	<0.0001	22.4000	1.2220
4	3.3274	0.04891	164	68.03	<0.0001	27.8667	1.3630
5	3.8864	0.03699	164	105.08	<0.0001	48.7333	1.8025
6	3.8944	0.03567	164	109.18	<0.0001	49.1250	1.7522
7	3.9318	0.03131	164	125.57	<0.0001	51.0000	1.5969
8	4.0073	0.03015	164	132.91	<0.0001	55.0000	1.6583
9	3.9370	0.03606	164	109.18	<0.0001	51.2667	1.8487
10	3.6687	0.04124	164	88.96	<0.0001	39.2000	1.6166
	$\hat{\eta}_i$	error _{std} ($\hat{\eta}_i$)				$\hat{\lambda}_i$	error _{std} ($\hat{\lambda}_i$)

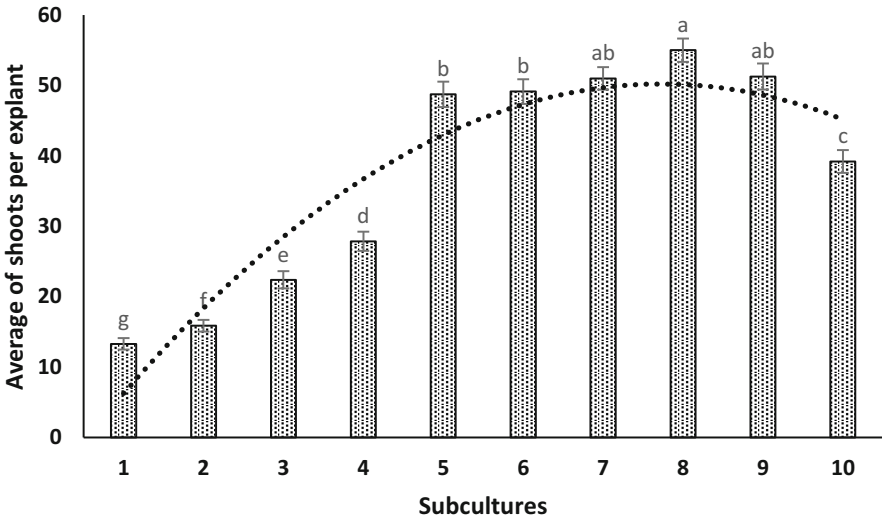


Fig. 5.1 Average number of shoots per subculture. Bars with different letters are statistically different using $\alpha = 0.05$

5.2.2 Example 2: CRDs with Poisson Response

Researchers want to determine whether the application of a new growth compound to walnut trees changes the amount of nuts produced per tree. They were applied at three different times (pre-flowering = 1, flowering = 2, and post-flowering = 3) and

Table 5.5 Number of nuts per tree (y_{ij}) in each of the combinations of the two factors

Trt	y_{ij}	Trt	y_{ij}	Trt	y_{ij}	Trt	y_{ij}	Trt	y_{ij}
C	1	A3	79	A3	89	B1	50	B2	138
A2	118	B3	99	C	21	A2	69	B1	69
A1	69	A1	50	A2	79	A3	69	A3	138
B1	89	B3	118	B3	99	A1	21	C	11
B1	99	A3	99	A1	79	B2	118	B2	89
B2	158	C	50	B1	118	C	30	B3	158
A1	89	A2	127	A2	89	B2	99	B3	118

in two formulations (A and B) plus a control (C). In addition to the treatments (Trt) there was a control, where no compound was applied. In total, 7 treatments were randomly applied to the experimental units (trees), i.e., 35 trees, in a rectangular arrangement (as shown below). The average number of nuts y_{ij} observed in the formulation and the time of application are provided in Table 5.5.

The components of the GLMM are listed below:

$$\text{Distribution : } y_{ij} \mid r_j \sim \text{Poisson}(\lambda_{ij})$$

$$r_j \sim N(0, \sigma_{\text{tree}}^2)$$

$$\text{Linear predictor: } \eta_{ij} = \eta + \tau_i + r_j$$

$$\text{Link function: } \log(\lambda_{ij}) = \eta_{ij}$$

where y_{ij} denotes the number of nuts in treatment i on tree j ($i = 1, 2, \dots, 7; j = 1, 2, \dots, 5$), η_{ij} is the linear predictor, η is the intercept, τ_i is the fixed effect due to treatment i , and r_j is the random effect due to tree j .

The following SAS statements allows a GLMM to be fitted in a completely randomized design with a Poisson response variable.

```
proc glimmix data=crd_nuez nobound method=laplace;
class trt rep;
model count = trt/dist=Poi link=log;
random rep;
lsmeans trt/lines ilink;
run;
```

The options in the model statement, dist, s and ilink communicates to SAS the type of data distribution, the fixed effects solution and to compute the inverse link, respectively. In addition, the option “lines” requests the GLIMMIX procedure in the “lsmeans” (least squares means) command, and the mean comparisons and the “ilink” option provide the inverse of the link function.

Part of the results is presented in Table 5.6. The value of the statistic for conditional distribution (part (a)) indicates that there is a strong overdispersion ($\chi^2/df = 3.62$), and the variance component estimates due to sampling in the experimental units (trees) is $\hat{\sigma}_{\text{tree}}^2 = 0.035$ (part (b)).

Table 5.6 Results of the analysis of variance

(a) Fit statistics for conditional distribution				
-2 Log L (count r. effects)				354.60
Pearson's chi-square				126.54
Pearson's chi-square/DF				3.62
(b) Covariance parameter estimates				
Cov Parm		Estimate	Standard error	
rep		0.03573	0.02362	
(c) Type III tests of fixed effects				
Effect	Num DF	Den DF	F-value	Pr > F
Trt	6	24	59.55	<0.0001

In addition, Table 5.6 (part (c)) shows the type III tests of fixed effects, indicating that there is a significant difference between treatments on the average number of nuts per tree ($P = 0.0001$). However, it is not recommended to continue with the inference and analysis of the experiment due to the presence of extra-variance (commonly known as overdispersion; Pearson's chi - square/DF = 3.62) in the data that strongly affects the F -test and the standard errors of the means.

A highly effective alternative to deal with the inconvenience of overdispersion in the data is to use a different distribution to the Poisson distribution. A negative binomial distribution is an excellent option for count data with overdispersion. Assuming that the conditional distribution of the observations is given by:

$$y_{ij} | r_j \sim \text{Poisson}(\lambda_{ij}),$$

where $\lambda_{ij} \sim \text{Gamma}(1/\phi, \phi)$, ϕ as the scale parameter and $r_j \sim N(0, \sigma_{\text{tree}}^2)$. The resulting new GLMM is:

$$\text{Distribution : } y_{ij} | r_j \sim \text{Negative Binomial}(\lambda_{ij}, \phi),$$

$$r_j \sim N(0, \sigma_{\text{tree}}^2)$$

$$\text{Linear predictor : } \eta_{ij} = \eta + \tau_i + r_j$$

$$\text{Link function: } \log(\lambda_{ij}) = \eta_{ij}$$

The following GLIMMIX statements for fitting this model under a negative binomial distribution in a CRD manner is provided next.

```
proc glimmix data=crd_nuez nobound method=laplace;
class trt rep;
model count = trt/dist=Negbin link=log;
random rep;
lsmeans trt/lines ilink;
run;
```

Table 5.7 Poisson and negative binomial model fit statistics

(a) Fit statistics		
Poisson	Negative	Binomial
-2 Log likelihood	374.83	328.03
AIC (smaller is better)	390.83	346.03
AICC (smaller is better)	396.37	353.23
BIC (smaller is better)	387.71	342.51
CAIC (smaller is better)	395.71	351.51
HQIC (smaller is better)	382.45	336.60
(b) Fit statistics for conditional distribution		
Poisson	Negative	Binomial
-2 Log L (count r. effects)	354.60	316.06
Pearson's chi-square	126.54	32.02
Pearson's chi-square/DF	3.62	0.91

Table 5.8 Variance component estimates and fixed effects tests

(a) Covariance parameter estimates				
Cov Parm	Estimate	Standard error		
Rep	0.04288	0.03398		
Scale	0.06141	0.02428		
(b) Type III tests of fixed effects				
Effect	Num DF	Den DF	F-value	Pr > F
Trt	6	24	18.75	<0.0001

Part of the results is listed below. The information criteria in Table 5.7 part (a) are helpful in choosing which model best fits the dataset. Clearly, the negative binomial distribution provides the best fit to these data. On the other hand, in the conditional fit statistics (part (b)), we observed that the Poisson model had a strong overdispersion (Pearson's chi - square/DF = 3.62) and that by fitting the data under a negative binomial distribution, the overdispersion of the dataset was removed (Pearson's chi - Square/DF = 0.91).

Table 5.8 shows the variance component estimates (part (a)) and the type III tests of fixed effects (part (b)). The estimated variance parameter, due to trees, is $\hat{\sigma}_{tree}^2 = 0.04288$, and the estimated scale parameter (Scale) is $\hat{\phi} = 0.06141$. The type III tests of fixed effects (part (b)) show that there is a highly significant effect of treatments on the average number of nuts ($P < 0.0001$).

The values under the column "Estimates" are the estimates of the linear predictor $\hat{\eta}_i$ (the model scale), and the values under "Mean" are the means $\hat{\lambda}_i$ (the data scale) with their respective standard errors obtained with the command "lsmeans" and "ilink" (Table 5.9). The results show that the treatments implemented in this experiment showed a higher average number of walnuts than did the "control" treatment C. In general, formula B applied to the walnut trees at the full-flowering stage showed a higher nut production.

Table 5.9 Estimates on the model scale (“Estimate”) and means on the data scale (“Mean”)

Trt least squares means							
Trt	Estimate	Standard error	DF	<i>t</i> -value	Pr > <i>t</i>	Mean	Standard error mean
A1	4.0865	0.1560	24	26.19	<0.0001	59.5307	9.2890
A2	4.5624	0.1519	24	30.04	<0.0001	95.8162	14.5529
A3	4.5293	0.1519	24	29.82	<0.0001	92.6956	14.0783
B1	4.4349	0.1529	24	29.01	<0.0001	84.3417	12.8958
B2	4.7863	0.1504	24	31.82	<0.0001	119.86	18.0304
B3	4.7641	0.1504	24	31.67	<0.0001	117.23	17.6335
C	3.0499	0.1742	24	17.51	<0.0001	21.1140	3.6785

Interest often arises in areas of agricultural and biological sciences to conduct experiments that involve random effects (blocks, locations, etc.) and response variables different from the normal distribution. For example, suppose that a certain number of treatments are being tested at different randomly selected locations, out of a sufficiently large number of locations. At each location, the experimental units are randomly assigned to each of the treatments. Let y_{ij} be the number of (observed) individuals possessing the characteristic of interest in the i th treatment in the j th block. The model for the mean structure of this experiment is

$$\eta_{ij} = \eta + \tau_i + b_j$$

where η is the intercept, τ_i is the fixed effect due to the i th treatment i , and b_j is the random effect of the block j with $b_j \sim N(0, \sigma_{\text{block}}^2)$.

5.2.3 Example 3: Control of Weeds in Cereal Crops in an RCBD

One of the main problems when growing cereal crops is the competition that exists between the weeds and seedlings. If a field supervisor is interested in testing five designed treatments plus a control for weed control in cereal crops, then a randomized complete block design (four blocks) should be used. Table 5.10 shows the number of weed plants observed in each of the treatments (y_{ij}) in parentheses.

Table 5.11 shows the sources of variation and the degrees of freedom of a randomized complete block design used in this experiment.

Since the response is count, it will be modeled using a GLMM with a Poisson response variable, which is stated below:

$$\begin{aligned} \text{Distribution : } y_{ij} \mid b_j &\sim \text{Poisson}(\lambda_{ij}) \\ b_j &\sim N(0, \sigma_{\text{block}}^2) \end{aligned}$$

Table 5.10 Number of weeds in each treatment (the number in parentheses corresponds to the treatment number)

Block						
A	(1) 438	(4) 17	(2) 538	(5) 18	(3) 77	(6) 115
B	(3) 61	(2) 422	(6) 57	(1) 442	(5) 26	(4) 31
C	(5) 77	(3) 157	(4) 87	(6) 100	(2) 377	(1) 319
D	(2) 315	(1) 380	(5) 20	(3) 52	(4) 16	(6) 45

Table 5.11 Analysis of variance

Sources of variation	Degrees of freedom
Block	$b - 1 = 4 - 1 = 3$
Treatment	$t - 1 = 6 - 1 = 5$
Error	$(t - 1)(b - 1) = 15$
Total	$tb - 1 = 23$

$$\text{Linear predictor: } \eta_{ij} = \eta + \tau_i + b_j$$

$$\text{Link function: } \log(\lambda_{ij}) = \eta_{ij}$$

where y_{ij} denotes the number of weed plants observed in treatment i and block j ($i = 1, 2, \dots, 6; j = 1, 2, 3, 4$), η_{ij} is the linear predictor, η is the intercept, τ_i is the fixed effect due to treatment i , and b_j is the random block effect ($b_j \sim N(0, \sigma_{\text{block}}^2)$).

Using the GLIMMIX procedure, the following syntax specifies the analysis of a GLMM with a Poisson response.

```
proc glimmix nobound method=laplace;
class Block Trt;
model Count = Trt/dist=Poisson s;
random block;
lsmeans Trt/diff lines ilink;
run; quit;
```

Note that in the above syntax, we use “method = laplace” (or we can also use “method = quadrature”) to fit the mixed model and obtain the chi-squared/DF fit statistic. If the method of integration is not specified, then a generalized chi-squared/DF statistic is obtained. The auxiliary options after the “lsmeans” command are described below: “diff” provides paired comparisons between treatments, “lines” provides the pair comparison of means using letters, and “ilink” provides the value of the inverse of the link function. Some of the outputs are listed below.

Table 5.12 (a) presents the basic information about the model and estimation procedure used.

Subsection (b) of Table 5.12 shows/ lists the “Dimensions” of the relevant matrices used in the model. The random effects matrix Z indicates that there are four columns due to blocks, and the fixed effects matrix X indicates that there is one column for the intercept plus six columns due to treatments.

Table 5.12 Basic model information

(a) Model information	
Dataset	WORK.DBCA
Response variable	Counting
Response distribution	Poisson
Link function	Log
Variance function	Default
Variance matrix	Not blocked
Estimation technique	Maximum likelihood
Likelihood approximation	Laplace
Degrees of freedom method	Containment
(b) Dimensions	
G-side Cov. parameters	1
Columns in X	7
Columns in Z	4
Subjects (blocks in V)	1
Max Obs per subject	24

Table 5.13 Model fit statistics

(a) Fit statistics	
-2 Log likelihood	434.46
AIC (smaller is better)	448.46
AICC (smaller is better)	455.46
BIC (smaller is better)	444.16
CAIC (smaller is better)	451.16
HQIC (smaller is better)	439.03
(b) Fit statistics for conditional distribution	
-2 Log L (Count r. effects)	418.66
Pearson's chi-square	283.09
Pearson's chi-square/DF ($\hat{\phi}$)	11.80

The “Fit statistics” and “Fit statistics for conditional distribution” (parts (a) and (b) of Table 5.13, respectively) show information about the fit of the GLMM. The generalized chi-squared statistic measures the sum of the residual squares in the final model and the relationship with its degrees of freedom; this is a measure of the variability of the observations about the model around the mean.

The value of Pearson's chi-square/DF for the conditional distribution is 11.8, well above up 1. This value gives strong evidence of overdispersion in the dataset. In other words, this value is calling our distribution and linear predictor assumption into question, which means that the variance function was not adequately specified.

Table 5.14 Variance component estimates, parameter estimates, and type III tests of fixed effects

Cov Parm		Estimate	Standard error				
Block ($\hat{\sigma}_{\text{block}}^2$)		0.01840	0.01377				
(a) Solutions for fixed effects							
Effect	Trt		Estimate	Standard error	DF	t-value	Pr > t
Intercept		$\hat{\eta}$	4.3637	0.08808	3	49.54	<0.0001
Trt	1	$\hat{\tau}_1$	1.6056	0.06155	15	26.09	<0.0001
Trt	2	$\hat{\tau}_2$	1.6508	0.06132	15	26.92	<0.0001
Trt	3	$\hat{\tau}_3$	0.09042	0.07769	15	1.16	0.2627
Trt	4	$\hat{\tau}_4$	-0.7416	0.09888	15	-7.50	<0.0001
Trt	5	$\hat{\tau}_5$	-0.8101	0.1012	15	-8.00	<0.0001
Trt	6	$\hat{\tau}_6$	0
(b) Type III tests of fixed effects							
Effect	Num DF	Den DF	F-value	Pr > F			
Trt	5	15	523.57	<0.0001			

Table 5.15 Estimated least squares means (“Mean”)

Trt least squares means							
Trt	Estimate	Standard error	DF	t-value	Pr > t	Mean	Standard error mean
1	5.9693	0.07237	15	82.49	<0.0001	391.25	28.3139
2	6.0145	0.07217	15	83.33	<0.0001	409.34	29.5437
3	4.4541	0.08652	15	51.48	<0.0001	85.9802	7.4390
4	3.6221	0.1060	15	34.19	<0.0001	37.4150	3.9643
5	3.5536	0.1081	15	32.86	<0.0001	34.9372	3.7784
6	4.3637	0.08808	15	49.54	<0.0001	78.5467	6.9186
	$\hat{\eta}_i$	$\text{error}_{\text{std}}(\hat{\eta}_i)$				$\hat{\lambda}_i$	$\text{error}_{\text{std}}(\hat{\lambda}_i)$

The F -test for testing $H_0 (\tau_1 = \tau_2 = \dots = \tau_6)$ or equivalent ($\mu_1 = \mu_2 = \dots = \mu_6$) indicates that there is a highly significant difference ($P < 0.0001$) in the average number of weeds in at least one treatment (part (c)) (Table 5.14).

The estimates of the linear predictor on the model scale for each of the treatments ($\hat{\eta}_i$) and the inverse of the linear predictor ($\hat{\lambda}_i$) on the data scale (with their respective standard errors) are calculated as follows $\hat{\eta}_i = \hat{\eta} + \hat{\tau}_i$ and $\hat{\lambda}_i = \exp(\hat{\eta}_i)$, respectively. These values are listed in Table 5.15.

The “plots” option in the “proc GLIMMIX” statement creates a set of plots for the raw residuals, Pearson residuals, and studentized residuals.

The panel consists of a plot of studentized residuals versus the linear predictor ($\hat{\eta}_i$), a histogram of the residuals with a normal density superimposed, a plot of residual versus quantiles, and a box plot for the residuals. The panel of studentized residuals indicates the possibility of a slightly skewed distribution (Fig. 5.2). In this figure, we can see that the range of values of the residuals changes, as do the values

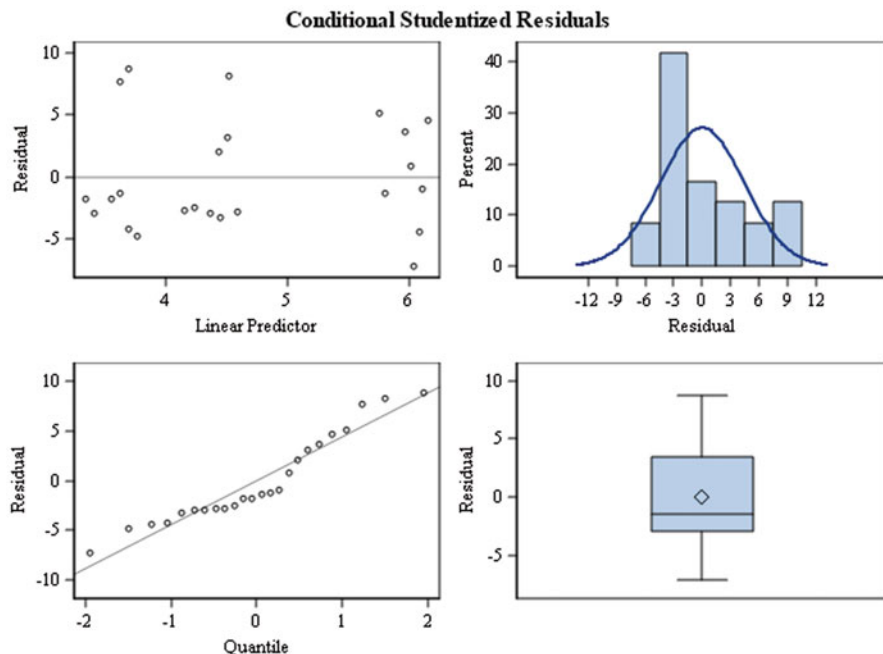


Fig. 5.2 Studentized conditional residuals

of the linear predictor, indicating that the assumption of constant variance is no longer met. The residuals–quantiles plot confirms the constant variance violation. A nonconstant variance may also suggest an incorrect selection of the response distribution or variance function.

5.2.4 Overdispersion in Poisson Data

Linear mixed models assume that the observations have a normal distribution conditional to the fixed effects of parameters. In addition, the mean μ is independent of the variance σ^2 , whereas, in most GLMMs that assume a binomial or Poisson distribution, the variance “dispersion” is set to 1. That is, if the mean is known, then we assume that the variance is also known. The extra variability not predicted by a generalized linear model’s random component reflects overdispersion. Overdispersion occurs because the mean and variance components of a GLM are related and depend on the same parameter that is being predicted through the predictor set. However, if overdispersion is present in a dataset, then the estimated standard errors and test statistics of the overall goodness of fit will be distorted and

adjustments must be made. In other words, when there is overdispersion in a dataset, the standard errors of the estimated parameters are too small, which leads to test statistics for the model parameters that are too large (i.e., type I error increases).

Overdispersion can be caused by several factors: omission of predictor variables in the model, high correlation in the observations due to nested effects, misspecification of the systematic component, or incorrect distribution of the data. Systematic or overdispersion deviations may be the result of incorrect assumptions about the stochastic and/or systematic component of the model. The model may also not fit the dataset well because of an incorrect choice of the link function. Systematic deviations may also result from lack of either random effects or independence of observations. These random factors should generally address deviance violations and problems associated with the systematic component.

According to Stroup (2013), overdispersion occurs when the variance exceeds the theoretical variance under the distribution model of the data. For any distribution with a nontrivial variance function, overdispersion is theoretically possible for distributions belonging to the one-parameter exponential family because they lack a scale parameter to mitigate the mean–variance relationship; therefore, models such as Poisson distribution are vulnerable to overdispersion. In summary, overdispersion occurs when:

- (a) The variance is larger than expected, which leads to standard errors that are not correct.
- (b) The mean structure is not well specified.
- (c) The linear predictor η is not well specified.
- (d) The chosen distribution of the data is not appropriate.
- (e) Predictor variables are omitted.
- (f) Observations are significantly correlated.

If we do not account for overdispersion, we underestimate the standard errors (for a large variance, the standard errors are not correct) and inflate the statistical tests causing the type I error to inflate and the confidence intervals to be unreliable. Fig. 5.3 shows that as the predicted mean $\hat{\mu}$ increases, the residuals have a larger spread in the plot, indicating that the variance may increase as a function of the mean, whereas Fig. 5.4 shows a nonconstant variance.

In the fit statistics obtained under the GLMM with the Poisson distribution (part (b), Table 5.13), the value of the statistic of Pearson's chi – square/DF = 11.8) indicates that there is a strong overdispersion in the dataset. Another aspect provided by the output is the value of the test statistic F ($F = 523.57$) tabulated in (part (c)) of Table 5.14. A value too large may indicate that the fit is incorrect. Once the researcher has detected overdispersion, he/she must consider the strategy that will take to remedy it. There are three possible alternatives to evaluate (test) and eliminate overdispersion. Below, we will review the three aforementioned alternatives.

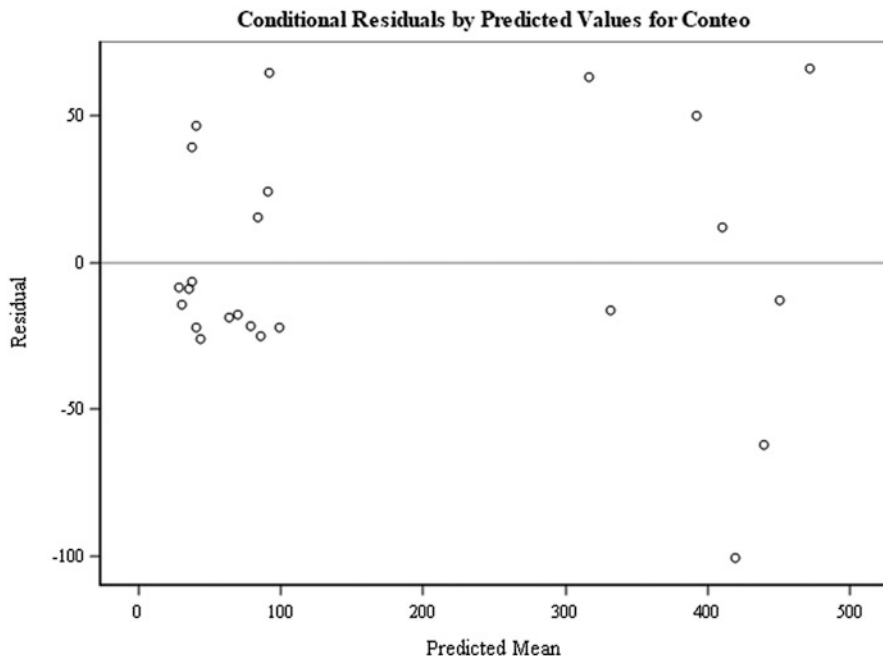


Fig. 5.3 Conditional residuals versus predicted values on the data scale

5.2.4.1 Using the Scale Parameter

The first alternative is to add a scale parameter and replace $\text{Var}(y_{ij}|b_j) = \lambda_{ij}$ by $\text{Var}(y_{ij}|b_j) = \phi\lambda_{ij}$. This consists of replacing the logarithm of the conditional likelihood $y_{ij} \log(\lambda_{ij}) - \lambda_{ij} - \log(y_{ij})$ by the quasi-likelihood $y_{ij} \log(\lambda_{ij}) - \lambda_{ij}/\phi$, assuming that $\phi > 1$ could adequately model the observed variance.

The following GLIMMIX syntax invokes this alternative of adding a scale parameter under a Poisson response variable.

```
proc glimmix;
class Block Trt;
model Count = Trt/dist=Poisson;
random intercept/subject=block;
random _residual_;
lsmeans Trt/ilink;
run;
```

The SAS code is highly similar to that previously used with the addition of the “random _residual_” command to the program. Note that the Laplace integration method (“method = laplace”) has been removed, which causes the estimation to be performed using the pseudo-likelihood (PL) method; the scale parameter is estimated and used in the adjustment of the standard errors and test statistics. The

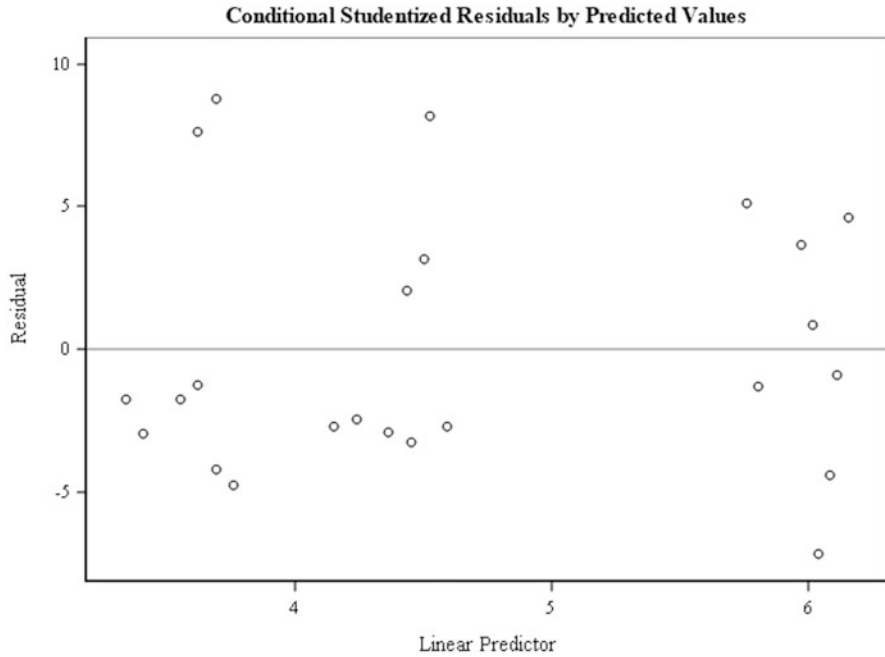


Fig. 5.4 Residuals on the model scale

GLIMMIX procedure uses the generalized chi-square divided by its degrees of freedom ($\text{Gener.chi - square}/\text{DF} = \hat{\phi}$) as the estimate of the scale parameter. All standard errors are multiplied by $\sqrt{\hat{\phi}}$, and all F -test values are divided by $\hat{\phi}$. Table 5.16 shows part of the results.

In Table 5.16, we observe the fit statistics (part (a)), covariance parameter estimates (part (b)), and the value of the scale parameter, which is equal to $\hat{\phi} = 19.4848$ (Residual(VC)). The value of the F -statistic under the Poisson distribution in the analysis is 26.87 (part (c)); this value is obtained by dividing the F -value from the previous analysis by $(523.57/\hat{\phi})$. The results indicate that even under this adjustment, overdispersion exists and that this value increases from 11.8 to 19.4848 (part (a)). The inclusion of the scale parameter affects the variance estimate due to blocks σ_{block}^2 as well as the estimates of treatment means (part (d)), but the main impact is on the standard errors.

The inclusion of the scale parameter implies that there is a quasi-likelihood, meaning that there is no true likelihood of the model and, therefore, there is no true likelihood process that provides a true expected value of λ and a variance of $\phi\lambda$.

Table 5.16 Results of the adjustment by adding the scale parameter

(a) Fit statistics							
-2 Res log pseudo-likelihood							29.48
Generalized chi-square							350.73
Gener. chi-square/DF							19.48
(b) Covariance parameter estimates							
Cov Parm		Subject	Estimate	Standard error			
Intercept		block	0.004981	0.02077			
Residual Variance component (VC)			19.4848	7.1346			
(c) Type III tests of fixed effects							
Effect	Num DF	Den DF	F-value	Pr > F			
Trt	5	15	26.87	<0.0001			
(d) Trt least squares means							
Trt	Estimate	Standard error	DF	t-value	Pr > t	Mean	Standard error mean
1	5.9779	0.1166	15	51.29	<0.0001	394.60	45.9935
2	6.0231	0.1142	15	52.74	<0.0001	412.84	47.1443
3	4.4626	0.2396	15	18.63	<0.0001	86.7160	20.7753
4	3.6306	0.3609	15	10.06	<0.0001	37.7352	13.6205
5	3.5621	0.3734	15	9.54	<0.0001	35.2362	13.1576
6	4.3722	0.2504	15	17.46	<0.0001	79.2189	19.8383

5.2.4.2 Linear Predictor Review

In count and binomial response variables, it is important to check whether the linear predictor is correctly specified, that is, whether it is being randomly affected by the experimental units within blocks. If λ_{ij} is being randomly affected by the experimental units within blocks, which is important in count and binomial response variables, then, the ANOVA table should include the effect of the block \times treatment source of variation; this must be specified in the linear predictor in a GLMM. Thus, the linear predictor is specified as

$$\eta_{ij} = \eta + \tau_i + b_j + (b\tau)_{ij}$$

$$\text{Distribution : } y_{ij} \mid b_j, b\tau_{ij} \sim \text{Poisson}(\lambda_{ij})$$

$$b_j \sim N(0, \sigma_{block}^2)$$

$$b\tau_{ij} \sim N(0, \sigma_{block \times \tau}^2)$$

$$\text{Linear predictor: } \eta_{ij} = \eta + \tau_i + b_j + (b\tau)_{ij}$$

$$\text{Link function: } \log(\lambda_{ij}) = \eta_{ij}.$$

The following GLIMMIX program allows the above model to be adjusted:

Table 5.17 Results of the fit by redefining the predictor of the model

(a) Fit statistics for conditional distribution							
-2 Log L (Count r. effects)				156.38			
Pearson's chi-square				2.58			
Pearson's chi-square/DF				0.11			
(b) Covariance parameter estimates							
Cov Parm	Subject	Estimate	Standard error				
Intercept	block	0.05969	0.05758				
Trt	block	0.1152	0.04115				
(c) Type III tests of fixed effects							
Effect	Num DF	Den DF	F-value	Pr > F			
Trt	5	15	41.48	<0.0001			
(d) Trt least squares means							
Trt	Estimate	Standard error	DF	t-value	Pr > t	Mean	Standard error mean
1	5.9692	0.2106	15	28.34	<0.0001	391.20	82.3947
2	6.0037	0.2106	15	28.51	<0.0001	404.92	85.2707
3	4.3674	0.2170	15	20.12	<0.0001	78.8402	17.1100
4	3.4255	0.2301	15	14.89	<0.0001	30.7370	7.0714
5	3.4005	0.2298	15	14.80	<0.0001	29.9786	6.8891
6	4.3027	0.2175	15	19.79	<0.0001	73.8997	16.0707

```
proc glimmix method=laplace;
class Block Trt;
model Count = Trt/dist=Poisson;
random intercept Trt/subject=block;
lsmeans Trt/ilink ;
run;
```

Part of the output is shown in Table 5.17. The results tabulated in part (a) indicate that the overdispersion has been eliminated ($\hat{\phi} = 0.11$), but there is a risk of underestimating the variance. For this reason, it is highly recommended that the value of $\hat{\phi}$ should be close to 1. The estimated variance components (part (b)) for blocks and block \times treatments are $\sigma_{\text{block}}^2 = 0.05969$ and $\sigma_{\text{block} \times \text{Trt}}^2 = 0.1152$, respectively.

The type III tests of fixed effects are highly significant ($P = 0.0001$), indicating that the six treatments are not equally effective in weed control (part (c)). The values in part (d) under the “Mean” column are the means on the original scale of the data for each of the treatments with their respective standard errors. The values of the means – compared with the previous ones – (using the scale parameter) do not vary much, but the standard errors have a more marked variation.

5.2.4.3 Using a Different Distribution

Another way to account for the problem of overdispersion when using a Poisson distribution is to change the assumed distribution of the response variable. Poisson variables have the same mean and variance, but, in biological sciences, with variables such as counts, this assumption is not always true. A negative binomial distribution is a good alternative (see Example 5.2), as previously discussed. A negative binomial variable's mean is denoted by the parameter $\lambda > 0$ and variance $\lambda + \phi\lambda^2$ by $\phi > 0$. That is, the expected value $E(y) = \lambda$ and variance $\text{Var}(y) = \lambda + \phi\lambda^2$, where ϕ is the scale parameter. The components of this model are shown below:

Given that $y_{ij} \mid b_j \sim \text{Poisson}(\lambda_{ij})$, it is assumed that $\lambda_{ij} \sim \text{Gamma}(1/\phi, \phi)$, with ϕ as the scale parameter and $b_j \sim N(0, \sigma_{\text{block}}^2)$. The new specification of the resulting GLMM is as follows:

$$\text{Distribution : } y_{ij} \mid b_j \sim \text{Negative Binomial}(\lambda_{ij}, \phi)$$

$$b_j \sim N(0, \sigma_{\text{block}}^2)$$

$$\text{Linear predictor: } \eta_{ij} = \eta + \tau_i + b_j$$

$$\text{Link function: } \log(\lambda_{ij}) = \eta_{ij}.$$

The following GLIMMIX statements fit the model with a negative binomial distribution.

```
proc glimmix method=laplace;
class block Trt;
model count = Trt/dist=NegBin;
random block;
lsmeans Trt/ilink;
run;
```

Some of the most relevant outputs from GLIMMIX are presented in Table 5.18. Pearson's chi-squared (Pearson's chi – square/DF) value of 0.88 (part (a)) shows that overdispersion in the dataset has been removed. The estimated scale parameter tabulated in part (b) (Scale) is $\hat{\phi} = 0.1080$. This value is not the same scale parameter estimated using the Poisson model with the “random _residual_” command, since the methodology for calculating them in these models is different. However, as mentioned above, both scale parameters affect the relationship between the mean and variance in the Poisson and negative binomial distributions.

The value of the test statistic shown in part (c) of Table 5.18, under the negative binomial distribution for the effect of treatments, is highly similar to the value obtained with the Poisson distribution when the effect of the block \times treatment interaction was added to the linear predictor. The values under “Estimate” are estimates of the linear predictor on the model scale (part (d)), whereas those under the “Mean” column are the treatment means on the data scale, using the negative binomial distribution. Of the three proposed alternatives to fit these data, the last two (including in the predictor the block–treatment interaction and assuming a negative binomial distribution) provides a better fit.

Table 5.18 Fitting results by redefining the model structure

(a) Fit statistics for conditional distribution							
-2 Log L (Count r. effects)				235.11			
Pearson's chi-square				21.13			
Pearson's chi-square/DF				0.88			
(b) Covariance parameter estimates							
Cov Parm	Estimate			Standard error			
Block	0.07713			0.06955			
Scale	0.1080			0.03768			
(c) Type III tests of fixed effects							
Effect	Num DF	Den DF	F-value	Pr > F			
Trt	5	15	41.11	<0.0001			
(d) Trt least squares means							
Trt	Estimate	Standard error	DF	t-value	Pr > t	Mean	Standard error mean
1	6.0280	0.2179	15	27.66	<0.0001	414.90	90.4085
2	6.0465	0.2174	15	27.81	<0.0001	422.64	91.8789
3	4.3941	0.2228	15	19.72	<0.0001	80.9704	18.0426
4	3.5190	0.2335	15	15.07	<0.0001	33.7516	7.8815
5	3.4684	0.2338	15	14.83	<0.0001	32.0863	7.5030
6	4.3439	0.2235	15	19.44	<0.0001	77.0085	17.2111

5.2.5 Factorial Designs

Many experiments involve studying the effects of two or more factors. Factorial designs are the most efficient for these types of experiments. In a factorial design, all possible combinations of factor levels are investigated in each replicate. If there are *a* levels of factor A and *b* levels of factor B, then each replicate contains all *ab* treatment combinations.

5.2.5.1 Example: A 2 × 4 Factorial with a Poisson Response

This application refers to a factorial experiment involving explants from cotyledons of cucumber (*Cucumis sativus* L.) with two factors, i.e., genotype (two levels) and culture medium (four levels). Each of the eight combinations of the genotype and culture levels were applied to four Petri dishes, each containing six leaf explants. The response variable was the number of buds in each of the leaf explants, i.e., the response variable was a count. There are two sources of variation in this application, namely, variation between Petri dishes and variation between the explants within the Petri dishes (Table 5.19).

The sources of variation and degrees of freedom for this experiment are shown in Table 5.20.

The components that define this model are shown below:

$$\begin{aligned} \text{Distribution : } & y_{ijk} \mid \text{petri.dish}_k, \\ & \text{explante}(\text{petri.dish})_{l(k)} \sim \text{Poisson}(\lambda_{ijk}) \text{petri.dish}_j \sim N\left(0, \sigma_{\text{petri.dish}}^2\right), \\ & \text{explant}(\text{petri.dish})_{l(k)} \sim N\left(0, \sigma_{\text{explant}(\text{petri.dish})}^2\right) \end{aligned}$$

$$\text{Linear predictor : } \eta_{ijkl} = \eta + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \text{petri.dish}_k + \text{explant}(\text{petri.dish})_{l(k)}$$

$$\text{Link function: } \log(\lambda_{ijkl}) = \eta_{ijkl}$$

where η_{ijkl} is the linear predictor in genotype i ($i = 1, 2$), culture medium j ($j = 1, 2, 3, 4$), Petri.dish k ($k = 1, 2, 3, 4$), and explant l ($l = 1, 2, 3, 4, 5, 6$), η is the intercept, α_i is fixed effect due to genotype i , β_j is the fixed effect due to culture medium j , $(\alpha\beta)_{ij}$ is the effect of the interaction between genotype i and culture medium j , Petri.dish_k is the random effect of the Petri.dish, and $\text{explant}(\text{Petri.dish})_{l(k)}$ is the random effect of the explant within the Petri.dish, assuming $\text{Petri.dish}_j \sim N(0, \sigma_{\text{Petri.dish}}^2)$ and $\text{explant}(\text{Petri.dish})_{l(k)} \sim N(0, \sigma_{\text{explant}(\text{Petri.dish})}^2)$.

The following GLIMMIX procedure fits a factorial experiment with a Poisson response.

```
proc glimmix method=laplace ;
class genotype culture petri.dish explant ;
model y = genotype | culture / dist=Poisson ;
random petri.dish explant (petri.dish) ;
lsmeans genotype | culture / ilink lines ;
run ;
```

Some of the SAS output is shown in Table 5.21. The fit statistics in part (a) for this dataset are shown below. Note that “method = laplace” was used for the estimation process and to obtain Pearson’s fit statistic χ^2/DF . The result indicates that there is evidence of overdispersion (Pearson’s chi – square/DF = 1.84).

Overdispersion, as discussed before, implies more variability in the data than would be expected, potentially explaining the lack of fit in a Poisson model. Part (b) shows the variance component estimates due to Petri_dish, which is equal to $\hat{\sigma}_{\text{Petri.dish}}^2 = 0.003616$, and, for the explants within Petri.dish, it is $\hat{\sigma}_{\text{explant}(\text{Petri.dish})}^2 = 0.01462$. However, the type III test of fixed effects indicates that there is a statistically significant effect of genotype, culture medium, and the interaction of both factors (part c).

The plot of residuals against the linear predictor in Fig. 5.5 provides further evidence of possible overdispersion.

The least squares means on the model scale for the genotype (part (a)), the culture medium (part (b)), and the interaction between both factors (part (c)) are listed under the “Estimate” column of Table 5.22, whereas under the “Mean” column are the means of these factors but in terms of the data.

Table 5.19 Number of buds counted in the cucumber experiment

Genotype	Culture	Petridish	Explant					
			1	2	3	4	5	6
1	1	1	10	7	5	7	12	1
	1	2	6	16	20	5	16	13
	1	3	10	12	13	0	12	15
	1	4	10	12	2	8	15	2
	2	1	20	16	14	18	17	20
	2	2	12	8	18	20	20	17
	2	3	22	13	24	15	10	14
	2	4	11	12	18	19	14	18
	3	1	20	18	15	18	20	18
	3	2	5	20	0	18	5	0
	3	3	17	10	20	12	14	21
	3	4	4	8	5	12	10	15
	4	1	10	4	0	5	4	8
	4	2	5	5	6	2	3	1
2	4	3	7	5	10	2	10	0
	4	4	9	5	8	4	4	7
	1	1	16	9	10	11	9	12
	1	2	13	7	3	2	3	12
	1	3	14	6	9	9	15	18
	1	4	13	0	3	7	5	2
	2	1	8	9	10	9	15	9
	2	2	11	12	8	9	12	10
	2	3	15	6	9	9	16	16
	2	4	18	12	6	0	13	14
	3	1	8	12	6	4	6	11
	3	2	10	10	12	12	15	10
	3	3	10	10	17	10	14	12
	3	4	10	14	14	14	9	14
4	1	9	5	1	9	15	9	
4	2	2	6	2	3	7	0	
4	3	4	12	2	0	4	3	
4	4	6	1	9	3	5	8	

Table 5.20 Sources of variation and degrees of freedom

Sources of variation	Degrees of freedom
Genotype	$a - 1 = 2 - 1 = 1$
Culture	$b - 1 = 4 - 1 = 3$
Genotype \times culture	$(a - 1)(b - 1) = 3$
Petri.dish \times Explant	$c(r - 1) = 4 \times 6 - 1 = 23$
Error	(by difference) = 161
Total	$abc - 1 = 191$

Table 5.21 Conditional fit statistics, variance component estimates, and type III tests of fixed effects under the Poisson distribution

(a) Fit statistics for conditional distribution				
-2 Log L (y r. effects)				1168.16
Pearson's chi-square				354.01
Pearson's chi-square/DF				1.84
(b) Covariance parameter estimates				
Cov Parm	Estimate		Standard error	
Petri.dish	0.003616		0.006014	
Explant (Petri.dish)	0.01462		0.008798	
(c) Type III tests of fixed effects				
Effect	Num DF	Den DF	F-value	Pr > F
Genotype	1	161	11.01	0.0011
Culture	3	161	57.30	<0.0001
Genotype*culture	3	161	3.95	0.0095

Since there is overdispersion in the data, we will fit the GLMM again using the negative binomial distribution. That is, under the following GLMM:

Distribution : $y_{ij} \mid \text{Petri.dish}_k, \text{explant}(\text{Petri.dish})_{l(k)} \sim \text{Negative Binomial}(\lambda_{ij}, \phi)$,

$$\text{Petri.dish}_j \sim N(0, \sigma_{\text{Petri.dish}}^2),$$

$$\text{explant}(\text{Petri.dish})_{l(k)} \sim N\left(0, \sigma_{\text{explant}(\text{Petri.dish})}^2\right),$$

Linear predictor : $\eta_{ijkl} = \eta + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \text{Petri.dish}_k + \text{explant}(\text{Petri.dish})_{l(k)}$

$$\text{Link function: } \log(\lambda_{ijkl}) = \eta_{ijkl}$$

and the scale parameter ϕ .

The following GLIMMIX program allows us to fit a GLMM with a negative binomial response variable.

```
proc glimmix ;
class genotype culture petri.dish explant ;
model y = cultivar | culture / dist = NegBin link = log ;
random petri.dish explant (petri.dish) ;
lsmeans cultivar | culture ;
run ;
```

It should be noted that this program is very similar to the previous one, and the only difference is that now a negative binomial distribution is used (“dist = negbin”). Part of the results is presented in Table 5.23. As we have already mentioned, a negative binomial distribution is another model for count variables when there is overdispersion in the dataset. If Pearson's chi-squared value divided over the degrees of freedom is less than or equal to 1, then the overdispersion is 0 or close to 0, which means that the model is able to efficiently capture the degree of overdispersion.

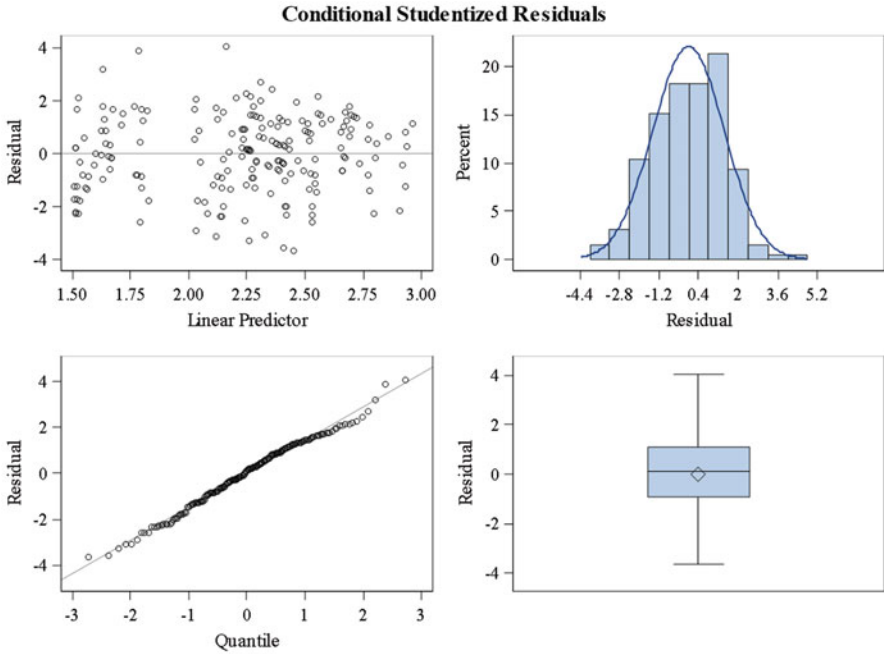


Fig. 5.5 Studentized conditional residuals

Based on the conditional distribution, Pearson’s chi-squared ($\chi^2/DF = 0.83$) fit statistic indicates that we have no evidence of overdispersion, so we can justify the negative binomial distribution, which is better than the Poisson distribution implemented above. In part (b), we show that the estimated scale parameter is $\hat{\phi} = 0.1712$. This value is not the same as the parameter for the quasi-Poisson model obtained with the “random_residual_” command. Note that the variance components were slightly affected. Additionally, in Table 5.23, we can see the type III tests for the fixed effects of the model in part (c), where a significant effect of genotype, culture, and the interaction between both factors (genotype*culture) can be observed on the number of buds in the leaf explant.

The “lines” option in the “lsmeans” command is used to obtain Fisher’s least significant difference (LSD) means for both factors and their interaction. The means and their respective standard errors, on the model scale (“Estimate” column) and on the data scale (“Mean” column), are tabulated in Table 5.24, the genotype and culture medium are in Table 5.25, and the interaction between both factors is in Table 5.26. The estimated values in this mean comparison for cultivar (Table 5.24) correspond to the values of the linear predictor $\hat{\eta}_i$ on the model scale, whereas the means on the data scale is $\hat{\lambda}_i$ (part (a)) and the comparison of means (on the model scale) are tabulated in part (b).

Table 5.22 Estimates on the model scale and means on the data scale under the Poisson distribution

(a) Genotype least squares means								
Genotype	Estimate	Standard error	DF	<i>t</i> -value	Pr > <i>t</i>	Mean	Standard error mean	
1	2.2979	0.05165	161	44.49	<0.0001	9.9533	0.5141	
2	2.1345	0.05298	161	40.29	<0.0001	8.4531	0.4479	
(b) Culture least squares means								
Culture	Estimate	Standard error	DF	<i>t</i> -value	Pr > <i>t</i>	Mean	Standard error mean	
1	2.1984	0.06180	161	35.57	<0.0001	9.0107	0.5569	
2	2.5684	0.05607	161	45.81	<0.0001	13.0456	0.7314	
3	2.4609	0.05738	161	42.89	<0.0001	11.7156	0.6723	
4	1.6371	0.07445	161	21.99	<0.0001	5.1402	0.3827	
(c) Genotype*culture least squares means								
Genotype	Culture	Estimate	Standard error	DF	<i>t</i> -value	Pr > <i>t</i>	Mean	Standard error mean
1	1	2.2465	0.07676	161	29.26	<0.0001	9.4547	0.7258
1	2	2.7789	0.06395	161	43.45	<0.0001	16.1018	1.0298
1	3	2.5331	0.06932	161	36.54	<0.0001	12.5925	0.8729
1	4	1.6331	0.09793	161	16.68	<0.0001	5.1196	0.5014
2	1	2.1503	0.07958	161	27.02	<0.0001	8.5877	0.6834
2	2	2.3580	0.07370	161	31.99	<0.0001	10.5694	0.7790
2	3	2.3887	0.07290	161	32.77	<0.0001	10.8997	0.7945
2	4	1.6411	0.09760	161	16.81	<0.0001	5.1609	0.5037

Table 5.23 Conditional fit statistics, variance component estimates, and type III tests of fixed effects under the negative binomial distribution

(a) Fit statistics for conditional distribution				
-2 Log L (y r. effects)	1143.90			
Pearson's chi-square	159.95			
Pearson's chi-square/DF	0.83			
(b) Covariance parameter estimates				
Cov Parm	Estimate	Standard error		
Petri.dish	-0.02717	.		
Explant (Petri.dish)	-0.04323	.		
Scale	0.1712	0.03514		
(c) Type III tests of fixed effects				
Effect	Num DF	Den DF	<i>F</i> -value	Pr > <i>F</i>
Genotype	1	161	4.43	0.0369
Culture	3	161	25.91	<0.0001
Genotype*culture	3	161	1.44	0.0322

Table 5.24 Estimates on the model scale and means on the data scale under the negative binomial distribution

(a) Cultivar least squares means							
Genotype	Estimate	Standard error	DF	<i>t</i> -value	Pr > <i>t</i>	Mean	Standard error mean
1	2.3054	0.05407	161	42.64	<0.0001	10.0287	0.5423
2	2.1426	0.05535	161	38.71	<0.0001	8.5219	0.4717
	$\hat{\eta}_i$					$\hat{\lambda}_i$	
(b) T grouping of genotype least squares means ($\alpha=0.05$)							
LS means with the same letter are not significantly different							
	Genotype						Estimate
1	2.3054						A
2	2.1426						B

Table 5.25 Means estimates on the model scale and data scale for the culture medium

(a) Culture least squares means							
Culture	Estimate	Standard error	DF	<i>t</i> -value	Pr > <i>t</i>	Mean	Standard error mean
1	2.2061	0.07653	161	28.82	<0.0001	9.0802	0.6950
2	2.5766	0.07198	161	35.80	<0.0001	13.1527	0.9468
3	2.4684	0.07300	161	33.81	<0.0001	11.8031	0.8617
4	1.6451	0.08708	161	18.89	<0.0001	5.1815	0.4512
	$\hat{\eta}_j$					$\hat{\lambda}_j$	
(b) T grouping of culture least squares means ($\alpha=0.05$)							
LS means with the same letter are not significantly different							
	Culture						Estimate
2	2.5766						A
3	2.4684						A
1	2.2061						B
4	1.6451						C

For the culture medium (Table 5.25), the estimated values in this comparison of means correspond to the values of the linear predictor $\hat{\eta}_j$ (on the model scale), but, by applying the inverse link to $\hat{\eta}_j$, we obtain the values under the “Mean” column that provide the means on the data scale (part (a)). The mean comparisons on the model scale are shown in part (b).

The results indicate that the means in culture media 2 and 3 provided a statistically similar average number of buds compared to the means in culture media 1 and 4 (see Fig. 5.6).

The interaction between both factors (Table 5.26), the average number of buds, and the mean comparisons are shown in Table 5.26.

Table 5.26 Estimates on the model scale and means on the data scale for the interaction between genotype and culture medium

Genotype*culture least squares means								
Genotype	Culture	Estimate	Standard error	DF	t-value	Pr > t	Mean	Standard error mean
1	1	2.2540	0.1072	161	21.02	<0.0001	9.5255	1.0212
1	2	2.7869	0.09844	161	28.31	<0.0001	16.2310	1.5978
1	3	2.5401	0.1020	161	24.91	<0.0001	12.6805	1.2933
1	4	1.6408	0.1233	161	13.31	<0.0001	5.1595	0.6360
2	1	2.1582	0.1093	161	19.75	<0.0001	8.6558	0.9457
2	2	2.3663	0.1050	161	22.53	<0.0001	10.6582	1.1196
2	3	2.3967	0.1045	161	22.94	<0.0001	10.9865	1.1478
2	4	1.6493	0.1230	161	13.41	<0.0001	5.2036	0.6401
		$\hat{\eta}_{ij}$					$\hat{\lambda}_{ij}$	

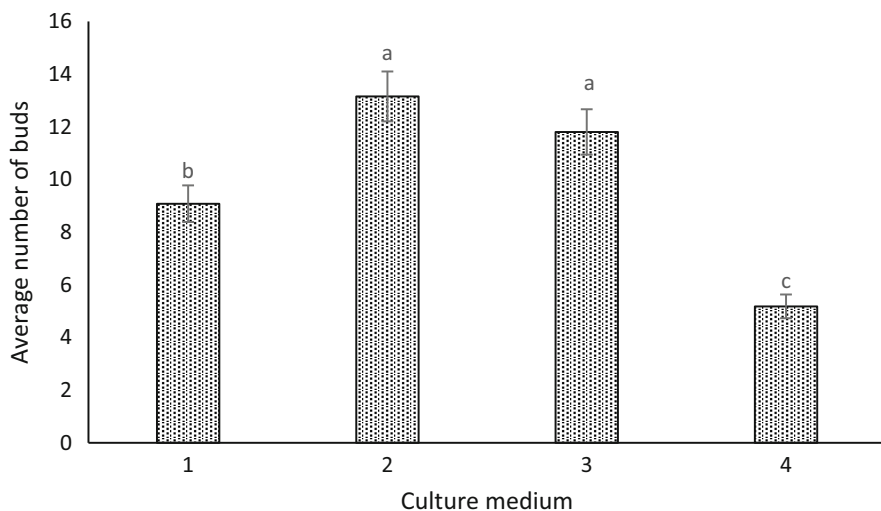


Fig. 5.6 Comparison of the average number of buds as a function of the type of culture medium (LSD, $\alpha = 0.05$)

The values under “Estimates” (Table 5.26) correspond to those of the linear predictor $\hat{\eta}_{ij}$ (model scale), but the values under “Mean” correspond to the means $\hat{\lambda}_{ij}$ on the data scale.

Graphically, Fig. 5.7 shows that genotype 1 in culture medium 2 provides the highest number of buds, whereas the lowest number of buds was observed in culture medium 4. For genotype 2, the highest number of buds was observed in culture media 2 and 3. Finally, culture medium 4 is less suitable for both genotypes.

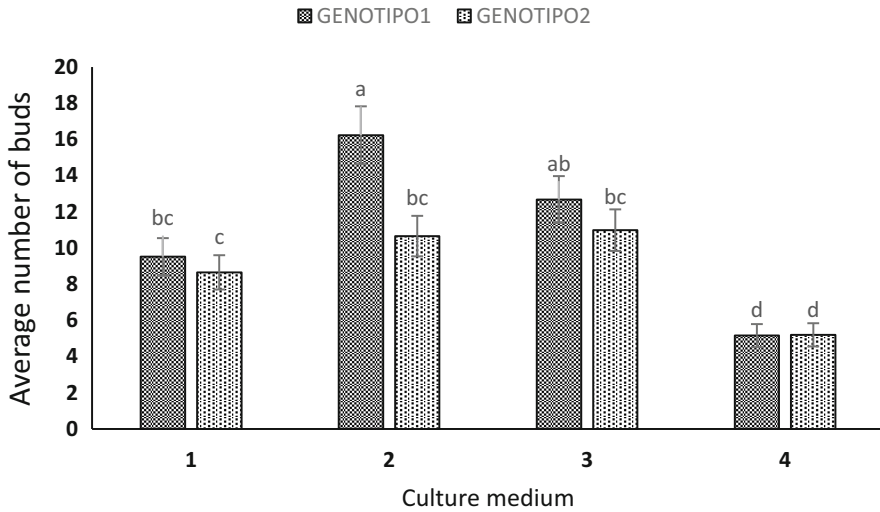


Fig. 5.7 Effect of the cultivar \times culture medium interaction on the average number of buds (LSD, $\alpha = 0.05$)

5.2.6 Latin Square (LS) Design

A Latin square (LS) is used where heterogeneity is associated with the crossing of two factors, generally, both with the same number of levels. This design was originally used in agricultural experimentation with plots placed in a square arrangement, with expected heterogeneity along the rows and columns of the square. Blocking in both directions across rows and columns is done in this experimental design. Sometimes in experimentation, blocking in two directions may be appropriate, i.e., the use of an LS design is a good option. Some examples are provided below to illustrate the use of this experimental design:

- Field experiments on plots set in a square arrangement with rows and columns that contribute to the heterogeneity between plots. For example, gradients of fertility, moisture, management practices, and so on.
- Experiments in greenhouses, rooms with a controlled environment, or growth chambers where the placement of shelves, trays, etc. with respect to walls or light sources can introduce systematic variability related to temperature, humidity, or light in different directions (e.g., left to right, back to front, or top to bottom).
- Laboratory experiments in which there are two potential sources of variability (e.g., technicians, machines, etc.) and researchers are aware of the possible impact of variation from both sources.

For an LS layout, the number of rows (r) and columns (c) should be equal to the number of treatments (t) and the number of replicates of each treatment. The assignment of treatments is such that each treatment appears exactly once in each

Table 5.27 Sources of variation and degrees of freedom of a Latin square design

Sources of variation	Degrees of freedom
Rows	$t - 1$
Columns	$t - 1$
Treatments	$t - 1$
Error	$(t - 1)(t - 2)$
Total	$t \times t - 1$

row and column, with each row and column containing a full set of treatments. Thus, the treatment effect estimates are independent of the differences between rows or columns, and the rows, columns, and treatments are orthogonal to each other.

The analysis of variance for this experimental design, assuming that there are r rows, c columns, and t treatments, with $r = c = t$, contains the following sources of variability (Table 5.27).

From the analysis of variance table, the linear model for an LS design with t treatments is as follows:

$$y_{ijk} = \mu + f_j + c_k + \tau_i + \varepsilon_{ijk}$$

where y_{ijk} is the response observed in treatment i in row f and column c , μ is the overall mean, f_j is the random effect of row j assuming $f_j \sim N(0, \sigma_f^2)$, c_k is the random effect of column k with $c_k \sim N(0, \sigma_c^2)$, τ_i is the fixed effect of treatment i , and ε_{ijk} is the distributed random error term $N(0, \sigma^2)$. Note that the treatments are allocated in the jk th quadrant (in row j and column k).

5.2.6.1 Latin Square Design with a Poisson Response

In a series of field experiments, several “inducer-attractant” strategies were tested to control insect pests in oilseed rape. In one experiment, the use of wild turnip rape (turnip rape) as an earlier flowering trap crop (TR) (the “attractor”) was tested together with the use of a repellent (an antifeedant) applied to oilseed rape in spring (S, the “inducer”). Untreated oilseed rape (U) was included as a control. The experiment was set up as a 6×6 Latin square with two replicates of each of the three treatments per row and column. An assessment of the number of mature pollen beetles was made on 10 plants per plot in early April, 1 day after spraying the repellent (antifeedant). The average number of adult beetles sampled on 10 plants per plot was recorded (Appendix 1: Data: Beetles). The question is: Is there evidence that the attractor or inducer works? That is, are fewer beetles present in the proposed treatments compared to the control?

The model components that define this GLMM are as described below:

Distribution: $y_{ijkl} | f_j, c_k \sim \text{Poisson}(\lambda_{ijkl})$

$$f_j \sim N(0, \sigma_f^2), c_k \sim N(0, \sigma_c^2)$$

Linear predictor: $\eta_{ijkl} = \eta + f_j + c_k + \tau_i$

$$\text{Link function: } \log(\lambda_{ijkl}) = \eta_{ijkl}$$

where η_{ijkl} is the linear predictor that relates the effect of the repetition l ($l = 1, 2$) in row j ($j = 1, 2, \dots, 6$) and column k ($k = 1, 2, \dots, 6$) when treatment i is applied ($i = 1, 2, 3, \dots$), η is the intercept, τ_i is the fixed effect of treatment i , f_j is random effect of row j , and c_k is the random effect due to column k , assuming that there is no interaction between the rows and columns as well as between the treatments and rows or the treatments and columns. The assumed distributions for rows and columns are $f \sim N(0, \sigma_f^2)$ and $c_k \sim N(0, \sigma_c^2)$, respectively. The model uses the linear predictor (η_{ijkl}) to estimate the means ($\lambda_{ijkl} = \mu_{ijkl}$) of the treatments.

The following GLIMMIX program fits a Latin square design with a Poisson response:

```
Proc glimmix nobound method=laplace;
class Row Column Treatment;
model count = treatment/dist=Poi link=log;
random row column;
lsmeans treatment/lines ilink;
run;
```

Part of the output is shown in Table 5.28. In the values of the fit statistics (part (a)), we observe that the value of Pearson's chi-square divided by the degrees of freedom is less than 1 ($\frac{\chi^2}{DF} = 0.55$), indicating that there is no overdispersion in the data and that the Poisson distribution adequately models the dataset.

The type III tests of fixed effects in part (b) indicate that there is no significant evidence of differences between the treatments ($P = 0.0621$).

Part (c) of Table 5.28 shows the estimates of treatments on the model scale ("Estimate") and on the data scale ("Mean") with their respective standard errors. The values 4.6191, 6.9396, and 5.1561 (under the "Mean" column) correspond to the treatment means for S, TR, and U, respectively.

5.2.6.2 Randomized Complete Block Design in a Split Plot

Sometimes the researcher is interested in testing multiple factors using different experimental units, and, in most cases, the experimenter cannot randomly accommodate the treatment combinations. Suppose that one wishes to test two factors, A and B with a and b levels each, respectively. The levels of the first factor (A) are randomly applied to the primary experimental units. Then, the levels of the second

Table 5.28 Results of the analysis of variance

(a) Fit statistics for conditional distribution							
-2 Log L (Conteo r. effects)							147.03
Pearson's chi-square							19.78
Pearson's chi-square/DF							0.55
(b) Type III tests of fixed effects							
Effect	Num DF	Den DF	F-value		Pr > F		
Treatment	3	23	3.14		0.0621		
(c) Treatment least squares means							
Treatment	Estimate	Standard error	DF	t-value	Pr > t	Mean	Standard error mean
S	1.5302	0.1343	23	11.39	<0.0001	4.6191	0.6204
TR	1.9372	0.1096	23	17.68	<0.0001	6.9396	0.7605
U	1.6402	0.1271	23	12.90	<0.0001	5.1561	0.6555
	$\hat{\tau}_i$					$\hat{\lambda}_i$	

factor (B) are applied to the secondary subunits formed within the primary unit in which the first factor was applied. In other words, the primary experimental unit (whole plot) was used for the application of the first factor; then, after this, it was divided to form the secondary experimental units (subplots) for the application of the levels of the second factor. Since the split-plot design has two levels of experimental units, the whole plot portions (primary units) and subplots (secondary units) have different experimental errors. Split-plot experiments were invented in agriculture by Fisher (1925), and their importance in industrial experimentation has been widely recognized (Yates 1935).

As a simple illustration, consider a study of three pulp preparation methods (factor A) and four temperature levels (factor B) on the effect of paper tensile strength (paper quality). A batch of pulp is produced by one of the three methods; it is then divided into four equal portions (samples). Each portion is cooked at a specific level of temperature. The assignment of treatments to plots and subplots is shown in Table 5.29.

The standard ANOVA model for two factors in a split-plot design, in which there are three levels of factor A and four levels of factor B nested within factor A, is described below:

$$y_{ijk} = \mu + \alpha_i + r_k + \alpha(r)_{ik} + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}$$

where y_{ijk} is the observed response at level i ($i = 1, 2, 3$) of factor A and at level j ($j = 1, 2, 3, 4$) of factor B in block k ($k = 1, 3, 3$), μ is the overall mean, α_i is the effect at level i of factor A, r_k is the random effect of blocks assuming $r_k \sim N(0, \sigma_r^2)$, $\alpha(r)_{ik}$ is the random effect of the error of the whole plot assuming $\alpha(r)_{ik} \sim N(0, \sigma_{\alpha(r)}^2)$, β_j is the effect at level j of factor B, $(\alpha\beta)_{ij}$ is the interaction

Table 5.29 Assigning treatments to whole plots and subplots

	Block 1			Block 2			Block 3		
	Preparation method			Preparation method			Preparation method		
Temperature	A_1	A_3	A_2	A_3	A_2	A_1	A_2	A_3	A_1
B_4	AB_{14}	AB_{34}	AB_{24}	AB_{34}	AB_{24}	AB_{14}	AB_{24}	AB_{34}	AB_{14}
B_2	AB_{12}	AB_{32}	AB_{22}	AB_{32}	AB_{22}	AB_{12}	AB_{22}	AB_{32}	AB_{12}
B_1	AB_{11}	AB_{31}	AB_{21}	AB_{31}	AB_{21}	AB_{11}	AB_{21}	AB_{31}	AB_{11}
B_3	AB_{13}	AB_{33}	AB_{23}	AB_{33}	AB_{23}	AB_{13}	AB_{23}	AB_{33}	AB_{13}

Table 5.30 Sources of variation and degrees of freedom for a randomized block design with a split-plot treatment arrangement

Sources of variation	Degrees of freedom
Blocks	$r - 1 = 3 - 1 = 2$
Factor A	$a - 1 = 3 - 1 = 2$
Error _a	$(a - 1)(r - 1) = 4$
Factor B	$b - 1 = 4 - 1 = 3$
A × B	$(a - 1)(b - 1) = 6$
Error	$a(r - 1)(b - 1) = 3 \times 2 \times 3 = 18$
Total	$r \times a \times b - 1 = 3 \times 3 \times 4 - 1 = 35$

fixed effect at level i of factor A and at level j of factor B, and ϵ_{ijk} is the normal random experimental error $\{\epsilon_{ijk} \sim iidN(\sigma^2)\}$. The ANOVA table with sources of variation is shown in Table 5.30 for this experimental design.

Example 5.1 A split-plot design in randomized complete block arrangement with a Poisson response

A split plot is probably the most common design structure in plant and soil research. Such experiments involve two or more treatment factors. Typically, large units called whole plots are grouped into blocks. The levels of the first factor are randomly assigned to whole plots. Each whole plot is divided into smaller units, called subplots (split plots). Next, the levels of the second factor are randomly assigned to units of split plots within each whole plot.

In this example, four blocks were implemented, which were divided into seven parts for the seven levels of the first factor ($A_1, A_2, A_3, A_4, A_5, A_6,$ and A_7), as whole plots. Then, each whole plot was divided into four units for randomly assigning the four levels of factor B, known as subplots ($B_1, B_2, B_3,$ and B_4). Both factors were used to control the growth of a particular weed. Both factors were randomly allocated in each block, as shown below:

Block 1								Block 4						
A_1	A_7	A_3	A_2	A_5	A_4	A_6	...	A_6	A_3	A_7	A_2	A_1	A_5	A_4
B_3	B_3	B_4	B_1	B_2	B_1	B_3		B_3	B_3	B_4	B_1	B_2	B_1	B_3
B_1	B_2	B_3	B_3	B_1	B_2	B_2	...	B_1	B_2	B_3	B_3	B_1	B_2	B_2
B_2	B_4	B_1	B_4	B_3	B_3	B_4		B_2	B_4	B_1	B_4	B_3	B_3	B_4
B_4	B_1	B_2	B_2	B_4	B_4	B_1	...	B_4	B_1	B_2	B_2	B_4	B_4	B_1

Table 5.31 Sources of variation and degrees of freedom for a randomized block design with a split-plot treatment arrangement

Sources of variation	Degrees of freedom
Blocks	$r - 1 = 4 - 1 = 3$
Factor A	$a - 1 = 7 - 1 = 6$
Error _a (A × r)	$(a - 1)(r - 1) = 18$
Factor B	$b - 1 = 4 - 1 = 3$
A × B	$(a - 1)(b - 1) = 18$
Error _b	$a(r - 1)(b - 1) = 7 \times 3 \times 3 = 63$
Total	$r \times a \times b - 1 = 4 \times 7 \times 4 - 1 = 111$

The sources of variation and degrees of freedom for this experiment are shown below in Table 5.31:

In this experiment, the response variable was the number of weeds in each of the plots (Appendix 1: Weed counts). The components that define this GLMM are as shown below:

$$\text{Distribution: } y_{ijk} \mid r_k, \alpha(r)_{ik} \sim \text{Poisson}(\lambda_{ijk})$$

$$r_k \sim N(0, \sigma_r^2), \alpha(r)_{ik} \sim N(0, \sigma_{ar}^2)$$

$$\text{Linear predictor: } \eta_{ijk} = \eta + \alpha_i + r_k + \alpha(r)_{ik} + \beta_j + (\alpha\beta)_{ij}$$

$$\text{Link function: } \log(\lambda_{ijk}) = \eta_{ijk}$$

where η_{ijk} is the linear predictor that relates the effect of factor A with i levels ($i = 1, 2, \dots, 7$) and factor B with j levels ($j = 1, 2, 3, 4$) in block k with ($k = 1, 2, 3, 4$); η is the intercept, α_i is the fixed effect at level i of factor A, β_j is the fixed effect at level j of factor B, $(\alpha\beta)_{ij}$ is the fixed effect of the interaction between level i of factor A and level j of factor B, r_k is the random effect due to block; and $\alpha(r)_{ik}$ is the random error effect of the whole plot, assuming $r_k \sim N(0, \sigma_r^2)$ and $\alpha(r)_{ik} \sim N(0, \sigma_{AR}^2)$, respectively. The model uses the aforementioned linear predictor (η_{ijk}) to estimate the means ($\lambda_{ijk} = \mu_{ijk}$) of the treatments.

The following GLIMMIX program fits a split-plot block design with a Poisson response variable:

```
proc glimmix method=laplace;
class block a b;
model count=a|b / dist=Poisson link=log;
random block block*a;
lsmeans a|b / lines ilink;
run;
```

Part of the output is shown below.

As in the previous examples, the Poisson model was found to be inadequate because the value of Pearson's chi-squared statistic divided by the degrees of

Table 5.32 Results of the analysis of variance

(a) Fit statistics for conditional distribution				
-2 Log L (Conteo r. effects)				1053.96
Pearson's chi-square				504.44
Pearson's chi-square/DF				4.50
(b) Covariance parameter estimates				
Cov Parm	Estimate	Standard error		
Bloque	0.01526	0.03867		
Bloque*A	0.2454	0.07565		
(c) Type III tests of fixed effects				
Effect	Num DF	Den DF	F-value	Pr > F
A	6	18	2.32	0.0775
B	3	63	22.91	<0.0001
A*B	18	63	10.06	<0.0001

freedom is greater than 1 ($\frac{\chi^2}{df} = 4.50$). This indicates that we have probably misspecified either the conditional distribution of $y | \mathbf{b}$ or the linear predictor, but, in this case, there is evidence that we need to look for other distributions for this dataset (part (a), Table 5.32). In addition, in part (b), the values of variance component estimates due to blocks and blocks \times A are tabulated ($\hat{\sigma}_r^2 = 0.01526$; $\hat{\sigma}_{ra}^2 = 0.2454$). On the other hand, the type III tests of fixed effects (part (c)) show a significant effect of factor B and the interaction between both factors.

An alternative to reduce the overdispersion is to keep the same linear predictor, changing the Poisson distribution in the response variable by the negative binomial distribution, that is:

$$\begin{aligned} \text{Distribution: } y_{ijk} | r_k, \alpha(r)_{ik} &\sim \text{Negative binomial } (\lambda_{ijk}, \phi) \\ r_k &\sim iidN(0, \sigma_r^2), \alpha(r)_{ik} \sim iidN(0, \sigma_{AR}^2) \\ \text{Linear predictor: } \eta_{ijk} &= \eta + \alpha_i + r_k + \alpha(r)_{ik} + \beta_j + (\alpha\beta)_{ij} \\ \text{Link function: } \log(\lambda_{ijk}) &= \eta_{ijk} \end{aligned}$$

The following syntax fits a GLMM under a negative binomial distribution.

```
proc glimmix method=Laplace;
class block a b;
model count=a|b / dist=NegBin link=log;
random intercept a /subject=block;
lsmeans a|b/lines ilink;
run;
```

Part of the output is shown below (Table 5.33). According to the results tabulated in (a), they indicate that the overdispersion has been removed from the analysis

Table 5.33 Results of the analysis of variance

(a) Fit statistics for conditional distribution				
–2 log L (Conteo r. effects)				838.51
Pearson's chi-square				79.36
Pearson's chi-square/DF				0.71
(b) Covariance parameter estimates				
Cov Parm	Subject	Estimate	Standard error	
Intercept	Bloque	0.002421	0.02768	
A	Bloque	0.1222	0.07102	
Scale		0.3458	0.06875	
(c) Type III tests of fixed effects				
Effect	Num DF	Den DF	F-value	Pr > F
A	6	18	2.71	0.0473
B	3	63	2.13	0.1054
A*B	18	63	1.18	0.3017

$(\frac{\chi^2}{df} = 0.71)$. The variance components estimates, tabulated in part (b), are $\sigma_r^2 = 0.0024$ and $\sigma_{AR}^2 = 0.1222$ for blocks and $blocks \times A$, respectively. The estimated scale parameter is $\hat{\phi} = 0.3458$. Note that the results under the negative binomial distribution differ from those obtained under the Poisson distribution, which is due, of course, to the fact that the negative binomial distribution better captures overdispersion. The fixed effects F -test for factor A is significant at the 5% significance level (part (c)), whereas factor B and the interaction effect do not significantly influence the response variable.

Example 5.2 A split-split plot in time in a randomized complete block design with a Poisson response.

The propagation of coffee seedlings through grafting in nurseries depends on several factors such as the type of substrate, the rootstock of the plant that will host the graft, type of graft, light intensity, type and size of the container, humidity, temperature, and so forth. The objective of this experiment was to evaluate the effect of shade cloth (light intensity), type of container, and clone on the number of leaves produced by the *Coffea canephora* P. clones grafted with the *Coffea arabica* L. variety Oro azteca.

The factors studied were the color of the shade cloth (black, pearl, and red), container size (tube of 0.5 kg and 1 kg), and five coffee clones of the variety *Coffea canephora* P. plus a franc foot (*Coffea arabica* L. and Var. Oro azteca) over a period of 11 months (Appendix 1: Coffee data). The clones used in the experiment are listed below (Table 5.34). Different physiological parameters were evaluated for 11 months.

This work was implemented in four randomized complete blocks. The following table exemplifies how a block was constructed.

Table 5.34 Clones of *Coffea canephora* P

Graft carrier (<i>Coffea canephora</i> P.)	Grafting (<i>Coffea arabica</i> L.)	Code
Clone 1	Var. Aztec gold	C1
Clone 2	Var. Aztec gold	C2
Clone 3	Var. Aztec gold	C3
Clone 4	Var. Aztec gold	C4
Clone 5	Var. Aztec gold	C5
Franc foot: <i>Coffea arabica</i> L. Var. Aztec gold		Pf

Shade cloth red			Shade cloth Perl			Shade cloth black		
Tray	Container 0.5 kg	Container 1 kg	Tray	Container 0.5 kg	Container 1 kg	Tray	Container 0.5 kg	Container 1 kg
C2	C5	C4	C4	C5	C2	C5	C2	C4
C4	Pf	C3	C3	Pf	C4	Pf	C3	C2
C3	C1	C5	C5	C1	C3	C1	C5	C3
C5	C2	Pf	Pf	C2	C5	C2	Pf	C5
Pf	C4	C1	C1	C4	Pf	C4	C1	Pf
C1	C3	C2	C2	C3	C1	C3	C24	C1

The statistical model describing a split-split plot in time design is described below:

$$\begin{aligned}
 y_{ijklm} = & \mu + \alpha_i + r_m + (ar)_{im} + \beta_j + (\alpha\beta)_{ij} + \gamma_k + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk} + (\alpha\beta\gamma)_{ijk} \\
 & + (raby)_{ijklm} + \tau_l + (\alpha\tau)_{il} + (\beta\tau)_{jl} + (\alpha\beta\tau)_{ijl} + (\gamma\tau)_{kl} + (\alpha\gamma\tau)_{ikl} \\
 & + (\beta\gamma\tau)_{jkl} + (\alpha\beta\gamma\tau)_{ijkl} + \varepsilon_{ijklm} \\
 & i = 1, 2, 3; j = 1, 2, 3, 4, 5; k = 1, 2, 3; l = 1, \dots, 11; m = 1, 2, 3, 4
 \end{aligned}$$

where y_{ijklm} is the response variable in repetition m , shade cloth i , clone j , and tray k in time l ; μ is the overall mean; α_i is the fixed effect due to the type of shade cloth; β_j , γ_k , and τ_l are the fixed effects due to clone type, tray, and sampling time, respectively; $(\alpha\beta)_{ij}$, $(\alpha\gamma)_{ik}$, $(\beta\gamma)_{jk}$, $(\alpha\tau)_{il}$, $(\beta\tau)_{jl}$, and $(\gamma\tau)_{kl}$ are the effects of the double interactions of the factors shade cloth type with clone, tray, and sampling time; $(\alpha\beta\gamma)_{ijk}$, $(\alpha\beta\tau)_{ijl}$, $(\alpha\gamma\tau)_{ikl}$, $(\beta\gamma\tau)_{jkl}$, and $(\alpha\beta\gamma\tau)_{ijkl}$ are the effects of the third and fourth interactions of the factors under study; $(ar)_{im}$ is the random effect of blocks with type of shade cloth with r_m , $(ar)_{im}$, $(raby)_{ijklm}$ are the random effect due to blocks, blocks with type of shade cloth, blocks with type of shade cloth, and time assuming $r_m \sim N(0, \sigma_r^2)$, $(ar)_{im} \sim N(0, \sigma_{ra}^2)$, $(raby)_{ijklm} \sim N(0, \sigma_{\alpha\beta\gamma(\text{rep})}^2)$, and ε_{ijklm} is random error $\{\varepsilon_{ijklm} \sim N(0, \sigma^2)\}$.

The following SAS program fits a GLMM in a split-split plot in time under a randomized complete block design with a Poisson response.

```

proc glimmix data=work.Nhojas_cafe nobound method=laplace;
class shade clone tray rep time;

```

Table 5.35 Fit statistics for choosing the correlation structure

Correlation structure					
Fit statistics	CS	AR(1)	UN	TOEP(1)	ANTE(1)
-2 Log likelihood	29047.38	29043.38	Not converged	29053.85	No converged
AIC (smaller is better)	30236.38	30235.38		30243.85	
AICC (smaller is better)	30338.31	30337.31		30345.44	
BIC (smaller is better)	29871.61	29869.61		29878.70	
CAIC (smaller is better)	30469.61	30465.61		30473.70	
HQIC (smaller is better)	29435.72	29432.72		29442.55	

Table 5.36 Conditional fit statistics and variance component estimates

(a) Fit statistics for conditional distribution			
-2 Log L (y r. effects)			28709.17
Pearson's chi-square			4288.74
Pearson's chi-square/DF			0.56
(b) Covariance parameter estimates			
Cov Parm	Subject	Estimate	Standard error
Variance	Rep	0.008106	0.001392
AR(1)	Rep	-0.3254	0.09437

```

model y = shade | clone | tray | time / dist = poi link = log;
random intercept shade shade * clone * tray / subject = rep type = ar(1) ;
lsmeans shade | clone | tray | time / lines ilink;
run;

```

Some of the results are listed below. To study which correlation structure best fits this experimental design, five types of correlation structures were tested (Table 5.35): compound symmetry (“CS”), autoregression of order 1 (“AR(1)”), unstructured (“UN”), Toeplitz of order 1 (“Toep(1)”), and ante (ANTE(1)). To do this, in the “random” command with the “type” option, the type of correlation to be tested is specified, and it is here where the option of type of variance–covariance structure must be changed. The fit statistics indicate that the variance–covariance structure that best fits the model is the autoregressive structure of order 1 (AR(1)). This can be seen in the following table in which the goodness-of-fit statistics for choosing between all these variance–covariance structures are reported.

Table 5.36 shows the conditional statistics and variance component estimates. The fit statistic Pearson’s chi – square/DF = 0.57 in part (a) indicates that, in a conditional model, there is no evidence of mis-specifying the distribution or linear predictor. In other words, there is no overdispersion in the dataset, and, therefore, it is reasonable that the analysis and inference can be based on the Poisson model.

The analysis of variance for the type III tests of fixed effects (Table 5.37) indicates that there is a highly significant effect of the main effect type of shade cloth ($P = 0.0001$), clone ($P = 0.0001$), and tray ($P = 0.0001$) as well as of most of the interactions, except for the interactions shade_cloth*clone; ($P = 0.3846$),

Table 5.37 Type III fixed effects tests

Type III tests of fixed effects				
Effect	Num DF	Den DF	F-value	Pr > F
Shade	2	6	3.44	0.1011
Clone	5	153	16.38	<0.0001
Shade*clone	10	153	1.08	0.3846
Tray	2	153	56.60	<0.0001
Shade*tray	4	153	8.83	<0.0001
Clone*tray	10	153	2.86	0.0027
Shade*clone*tray	20	153	1.71	0.0363
Time	10	6822	721.20	<0.0001
Shade*time	20	6822	6.91	<0.0001
Clone*time	50	6822	3.17	<0.0001
Shade*clone*time	100	6822	0.80	0.9289
Tray*time	20	6822	9.03	<0.0001
Shade*tray*time	40	6822	2.42	<0.0001
Clone*tray*time	100	6822	0.74	0.9760
Shade*clone*tray*time	200	6822	1.07	0.2484

Table 5.38 Estimated means on the model scale and on the data scale for the shade cloth

(a) Shade cloth least squares means

Shade cloth	Estimate	Standard error	DF	t-value	Pr > t	Mean	Standard error mean
Black	1.6221	0.01542	2	105.17	<0.0001	5.0638	0.07810
Pearl	1.5472	0.01533	2	100.94	<0.0001	4.6981	0.07201
Red	1.7184	0.01301	2	132.09	<0.0001	5.5757	0.07254

(b) T grouping of shade cloth least squares means ($\alpha=0.05$)
 LS means with the same letter are not significantly different

Shade cloth	Estimate ($\hat{\tau}_i$)	
Red	1.7184	A
Black	1.6221	B
Pearl	1.5472	B

shade_cloth*tray*time ($P = 0.9289$), clone*tray*time ($P = 0.9760$), and shade_cloth*clone*tray*time ($P = 0.2484$).

The means and standard errors of each of the main effects, on the data scale, for shade_cloth, tray, and clone are shown in the “Mean” column in part (a) of Table 5.38, whereas in part (b), the mean comparisons for the type of shade cloth are shown.

Table 5.39 presents the estimates of the linear predictor (“Estimates” column) in terms of the model scale and treatment means in terms of the data scale (“Mean” column) for the type of clone (part (a)). In addition, in Table 5.39 (part (b)), the mean comparisons are presented for the type of clone.

Table 5.39 Estimated means on the model scale and on the data scale for the type of clone

(a) Clone least squares means							
Clone	Estimate	Standard error	DF	<i>t</i> -value	Pr > <i>t</i>	Mean	Standard error mean
C1	1.5008	0.04989	153	30.08	<0.0001	4.4854	0.2238
C2	1.4250	0.05080	153	28.05	<0.0001	4.1578	0.2112
C3	1.5064	0.05019	153	30.02	<0.0001	4.5106	0.2264
C4	1.4750	0.05029	153	29.33	<0.0001	4.3709	0.2198
C5	1.5965	0.04970	153	32.12	<0.0001	4.9357	0.2453
Pf	1.6344	0.04943	153	33.07	<0.0001	5.1264	0.2534

(b) T grouping of clone least squares means ($\alpha=0.05$)			
LS means with the same letter are not significantly different			
	Clone	Estimate	
Pf	1.6344		A
C5	1.5965		A
C3	1.5064		B
C1	1.5008		B
C4	1.4750	C	B
C2	1.4250	C	

Table 5.40 Estimated means on the model scale and on the data scale for the tray factor

(a) Tray least squares means							
Tray	Estimate	Standard error	DF	<i>t</i> -value	Pr > <i>t</i>	Mean	Standard error mean
CH1	1.3843	0.04859		28.49	<0.0001	3.9921	0.1940
CH2	1.5665	0.04838		32.38	<0.0001	4.7898	0.2317
CH3	1.6183	0.04819		33.58	<0.0001	5.0443	0.2431

(b) T grouping of tray least squares means ($\alpha=0.05$)			
LS means with the same letter are not significantly different			
	Tray	Estimate	
CH3	1.6183		A
CH2	1.5665		B
CH1	1.3843		C

Table 5.40 presents the estimates for the levels of the tray on both scales (part (a)). Similarly, in this table (part (b)), the treatment mean comparisons are presented for the levels of the tray.

Tables 5.41, 5.42, 5.43, and 5.44 show the means and standard errors on both scales of the two-factor and three-factor interactions.

Interaction type of shade cloth*clone

Interaction type of shade cloth*tray

Interaction clone*tray

Interaction shade*clone*tray

Although it is not the objective of this book, part of the results is discussed below. In Fig. 5.8, it is possible to observe that the red shade cloth significantly stimulates leaf production in coffee grafts, followed by the black and pearl shade cloths. The

Table 5.41 Estimated means on the model scale and on the data scale for the type of shade cloth*clone

Shade cloth*clone least squares means								
Shade cloth	Clone	Estimate	Standard error	DF	t-value	Pr > t	Mean	Standard error mean
Black	C1	1.5109	0.06230	153	24.25	<0.0001	4.5307	0.2823
Black	C2	1.3340	0.06507	153	20.50	<0.0001	3.7961	0.2470
Black	C3	1.4990	0.06354	153	23.59	<0.0001	4.4771	0.2845
Black	C4	1.4485	0.06425	153	22.54	<0.0001	4.2566	0.2735
Black	C5	1.5916	0.06163	153	25.83	<0.0001	4.9118	0.3027
Black	pf	1.6219	0.06114	153	26.53	<0.0001	5.0628	0.3095
Pearl	C1	1.3835	0.07711	153	17.94	<0.0001	3.9889	0.3076
Pearl	C2	1.3926	0.07781	153	17.90	<0.0001	4.0254	0.3132
Pearl	C3	1.4028	0.07589	153	18.49	<0.0001	4.0666	0.3086
Pearl	C4	1.4288	0.07575	153	18.86	<0.0001	4.1736	0.3161
Pearl	C5	1.5216	0.07536	153	20.19	<0.0001	4.5797	0.3451
Pearl	pf	1.5285	0.07458	153	20.50	<0.0001	4.6112	0.3439
Red	C1	1.6081	0.06991	153	23.00	<0.0001	4.9933	0.3491
Red	C2	1.5483	0.07100	153	21.81	<0.0001	4.7036	0.3339
Red	C3	1.6175	0.07055	153	22.93	<0.0001	5.0404	0.3556
Red	C4	1.5477	0.07072	153	21.88	<0.0001	4.7005	0.3324
Red	C5	1.6762	0.06971	153	24.04	<0.0001	5.3451	0.3726
Red	pf	1.7528	0.06923	153	25.32	<0.0001	5.7707	0.3995

Table 5.42 Estimated means on the model scale and on the data scale for the interaction type of shade cloth*tray

Shade*tray least squares means								
Shade cloth	Tray	Estimate	Standard error	DF	t-value	Pr > t	Mean	Standard error mean
Black	CH1	1.4274	0.05961	153	23.94	<0.0001	4.1679	0.2485
Black	CH2	1.5523	0.05846	153	26.55	<0.0001	4.7224	0.2761
Black	CH3	1.5232	0.05824	153	26.15	<0.0001	4.5869	0.2672
Pearl	CH1	1.2070	0.07354	153	16.41	<0.0001	3.3434	0.2459
Pearl	CH2	1.4972	0.07218	153	20.74	<0.0001	4.4691	0.3226
Pearl	CH3	1.6247	0.07145	153	22.74	<0.0001	5.0771	0.3628
Red	CH1	1.5185	0.06733	153	22.55	<0.0001	4.5655	0.3074
Red	CH2	1.6499	0.06714	153	24.57	<0.0001	5.2066	0.3496
Red	CH3	1.7068	0.06732	153	25.35	<0.0001	5.5114	0.3710

production of leaves in coffee grafts shows a bimodal figure that can be due to factors such as humidity and temperature. Extreme conditions of both factors cause stress at the growing points and, therefore, the appearance of leaves.

Regarding the type of clone used as rootstock, the clones showed a better average leaf production in months 5 and 6, whereas the lowest production was observed in

Table 5.43 Estimated means on the model scale and on the data scale for the clone–tray interaction

Clone*tray least squares means								
Clone	Tray	Estimate	Standard error	DF	<i>t</i> -value	Pr > <i>t</i>	Mean	Standard error mean
C1	CH1	1.3916	0.06112	153	22.77	<0.0001	4.0214	0.2458
C1	CH2	1.5502	0.05861	153	26.45	<0.0001	4.7122	0.2762
C1	CH3	1.5607	0.05861	153	26.63	<0.0001	4.7622	0.2791
C2	CH1	1.2242	0.06459	153	18.95	<0.0001	3.4014	0.2197
C2	CH2	1.4780	0.06029	153	24.51	<0.0001	4.3843	0.2644
C2	CH3	1.5727	0.05890	153	26.70	<0.0001	4.8196	0.2839
C3	CH1	1.2924	0.06114	153	21.14	<0.0001	3.6414	0.2226
C3	CH2	1.5433	0.05975	153	25.83	<0.0001	4.6799	0.2796
C3	CH3	1.6836	0.05841	153	28.83	<0.0001	5.3851	0.3145
C4	CH1	1.2982	0.06251	153	20.77	<0.0001	3.6626	0.2289
C4	CH2	1.5829	0.05815	153	27.22	<0.0001	4.8690	0.2831
C4	CH3	1.5439	0.05939	153	26.00	<0.0001	4.6828	0.2781
C5	CH1	1.5311	0.05843	153	26.20	<0.0001	4.6234	0.2702
C5	CH2	1.5981	0.05920	153	26.99	<0.0001	4.9438	0.2927
C5	CH3	1.6602	0.05803	153	28.61	<0.0001	5.2604	0.3053
pf	CH1	1.5684	0.05794	153	27.07	<0.0001	4.7989	0.2781
pf	CH2	1.6464	0.05833	153	28.23	<0.0001	5.1884	0.3026
pf	CH3	1.6884	0.05728	153	29.48	<0.0001	5.4107	0.3099

months 1, 2, 8, and 9. The franc foot showed a higher average of leaves compared to the rest of the clones (Fig. 5.9).

5.3 Exercises

Exercise 5.3.1 A researcher in the area of plant sciences wants to know what is the response of a plant in vitro culture when it is exposed to different concentrations (ppm) of a chemical compound to the number of outbreaks that the explant produces (y_{ij}). The data for this experiment are given below (Table 5.45):

- Write down the analysis of variance table (sources of variation and degrees of freedom).
- Write down the components of the GLMM.
- Analyze the dataset with the model proposed in (b).
- Compare and contrast the results of these analyses. If necessary, reanalyze the dataset using the same model as above, but, now, assume that the data have a negative binomial distribution.
- Summarize the relevant results.

Table 5.44 Estimated means on the model scale and on the data scale for the shade-clone-tray interaction

Shade*clone*tray least squares means									
Shade cloth	Clone	Tray	Estimate	Standard error	DF	t-value	Pr > t	Mean	Standard error mean
Black	C1	CH1	1.2821	0.1528	153	8.39	<0.0001	3.6041	0.5509
Black	C1	CH2	1.4143	0.1521	153	9.30	<0.0001	4.1136	0.6258
Black	C1	CH3	1.2201	0.1538	153	7.93	<0.0001	3.3874	0.5209
Black	C2	CH1	0.8131	0.1615	153	5.04	<0.0001	2.2549	0.3641
Black	C2	CH2	1.1486	0.1543	153	7.45	<0.0001	3.1538	0.4866
Black	C2	CH3	1.3376	0.1533	153	8.72	<0.0001	3.8100	0.5842
Black	C3	CH1	1.1809	0.1548	153	7.63	<0.0001	3.2574	0.5041
Black	C3	CH2	1.1105	0.1550	153	7.17	<0.0001	3.0359	0.4705
Black	C3	CH3	1.3672	0.1528	153	8.95	<0.0001	3.9242	0.5996
Black	C4	CH1	0.7672	0.1608	153	4.77	<0.0001	2.1538	0.3462
Black	C4	CH2	1.4660	0.1517	153	9.66	<0.0001	4.3318	0.6573
Black	C4	CH3	1.3925	0.1523	153	9.14	<0.0001	4.0250	0.6131
Black	C5	CH1	1.2316	0.1538	153	8.01	<0.0001	3.4269	0.5270
Black	C5	CH2	1.6090	0.1507	153	10.67	<0.0001	4.9979	0.7534
Black	C5	CH3	1.4684	0.1515	153	9.70	<0.0001	4.3422	0.6577
Black	Pf	CH1	1.6751	0.1503	153	11.15	<0.0001	5.3393	0.8025
Black	Pf	CH2	1.3126	0.1548	153	8.48	<0.0001	3.7160	0.5753
Black	Pf	CH3	1.5092	0.1511	153	9.99	<0.0001	4.5231	0.6834
Pearl	C1	CH1	0.6441	0.1741	153	3.70	0.0003	1.9043	0.3314
Pearl	C1	CH2	1.3602	0.1639	153	8.30	<0.0001	3.8970	0.6387
Pearl	C1	CH3	1.6030	0.1633	153	9.82	<0.0001	4.9678	0.8111
Pearl	C2	CH1	0.6336	0.1741	153	3.64	0.0004	1.8844	0.3281
Pearl	C2	CH2	1.2050	0.1672	153	7.21	<0.0001	3.3366	0.5579
Pearl	C2	CH3	1.5547	0.1635	153	9.51	<0.0001	4.7335	0.7740
Pearl	C3	CH1	0.8786	0.1684	153	5.22	<0.0001	2.4074	0.4053
Pearl	C3	CH2	1.2777	0.1646	153	7.76	<0.0001	3.5885	0.5905
Pearl	C3	CH3	1.5724	0.1637	153	9.60	<0.0001	4.8184	0.7889
Pearl	C4	CH1	0.9893	0.1680	153	5.89	<0.0001	2.6893	0.4519
Pearl	C4	CH2	1.4198	0.1636	153	8.68	<0.0001	4.1362	0.6769
Pearl	C4	CH3	1.4357	0.1646	153	8.72	<0.0001	4.2026	0.6919
Pearl	C5	CH1	1.4557	0.1631	153	8.93	<0.0001	4.2875	0.6992
Pearl	C5	CH2	1.1672	0.1696	153	6.88	<0.0001	3.2130	0.5449
Pearl	C5	CH3	1.6010	0.1633	153	9.80	<0.0001	4.9582	0.8098
Pearl	Pf	CH1	1.1901	0.1649	153	7.22	<0.0001	3.2875	0.5422
Pearl	Pf	CH2	1.4004	0.1643	153	8.52	<0.0001	4.0570	0.6665
Pearl	Pf	CH3	1.7623	0.1620	153	10.88	<0.0001	5.8260	0.9440
Red	C1	CH1	1.5245	0.1606	153	9.49	<0.0001	4.5930	0.7379
Red	C1	CH2	1.6004	0.1605	153	9.97	<0.0001	4.9548	0.7953
Red	C1	CH3	1.6327	0.1607	153	10.16	<0.0001	5.1178	0.8224
Red	C2	CH1	1.3462	0.1630	153	8.26	<0.0001	3.8430	0.6264

(continued)

Table 5.44 (continued)

Shade*clone*tray least squares means									
Shade cloth	Clone	Tray	Estimate	Standard error	DF	t-value	Pr > t	Mean	Standard error mean
Red	C2	CH2	1.4270	0.1622	153	8.80	<0.0001	4.1663	0.6759
Red	C2	CH3	1.6500	0.1620	153	10.19	<0.0001	5.2071	0.8435
Red	C3	CH1	1.3915	0.1632	153	8.53	<0.0001	4.0207	0.6563
Red	C3	CH2	1.4491	0.1614	153	8.98	<0.0001	4.2592	0.6872
Red	C3	CH3	1.7875	0.1603	153	11.15	<0.0001	5.9747	0.9577
Red	C4	CH1	1.3961	0.1614	153	8.65	<0.0001	4.0394	0.6520
Red	C4	CH2	1.5874	0.1606	153	9.89	<0.0001	4.8910	0.7854
Red	C4	CH3	1.3805	0.1635	153	8.44	<0.0001	3.9768	0.6503
Red	C5	CH1	1.6313	0.1601	153	10.19	<0.0001	5.1103	0.8180
Red	C5	CH2	1.6395	0.1605	153	10.22	<0.0001	5.1527	0.8268
Red	C5	CH3	1.6470	0.1610	153	10.23	<0.0001	5.1912	0.8360
Red	Pf	CH1	1.7075	0.1600	153	10.67	<0.0001	5.5151	0.8825
Red	Pf	CH2	1.7594	0.1594	153	11.04	<0.0001	5.8087	0.9260
Red	Pf	CH3	1.7568	0.1601	153	10.98	<0.0001	5.7938	0.9273

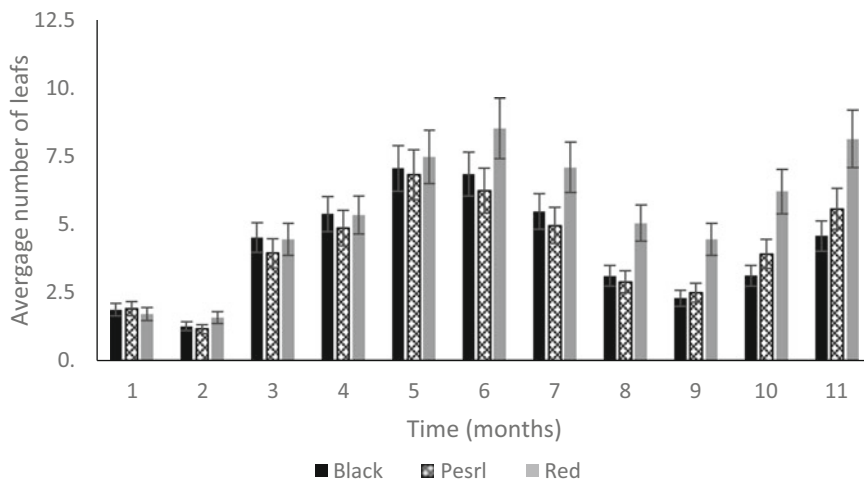


Fig. 5.8 Effect of mesh type on the average number of leaves

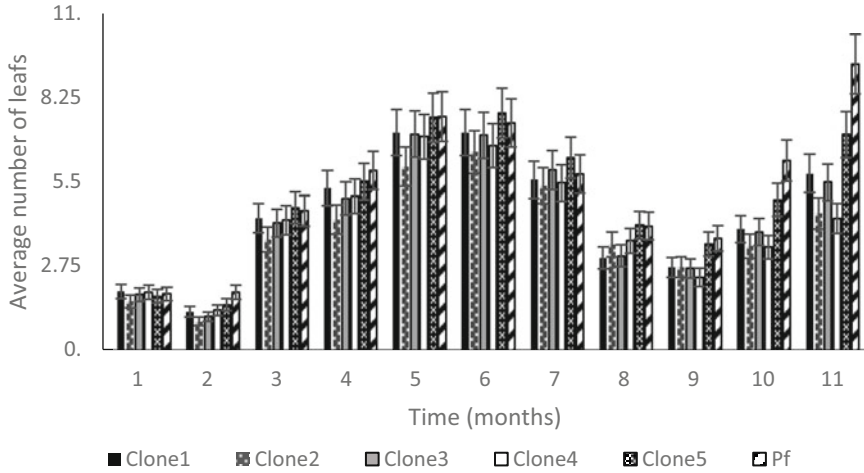


Fig. 5.9 Effect of mesh type on the average number of leaves

Exercise 5.3.2 Earthworms (*Lubricus terrestris* L.) were counted in four replicates of a factorial experiment at the W.K. Kellogg Biological Station in Battle Creek, Michigan, in 1995. A 2^4 factorial experiment was conducted. Factors and treatment levels were plowing (chiseled and unplowed), input level (conventional and low), manure application (yes/no), and crop (corn and soybean). The objective of interest was whether *L. terrestris* density varies according to these management protocols and how various factors act and interact. The data (not pooled) in the table shows the total worm counts (per square foot) in the factorial design 2^4 for the experimental units 64 ($2^4 \times 4$) (juvenile and adult worms). The numbers in each cell of the table correspond to the counts in the replicates (Table 5.46).

- (a) Write down the analysis of variance table (sources of variation and degrees of freedom).
- (b) Write down the components of the GLMM.
- (c) Analyze the dataset with the model proposed in (b).
- (d) Summarize the relevant results.

Exercise 5.3.3 This experiment involves an investigation of genotypic variation within cultivars of pore (*Allium porrum* L.) with respect to adventitious shoot formation in the callus tissue. The data in Table 5.47 refer to 20 genotypes of 1 cultivar. Each genotype is represented by six calluses. These observations are the number of shoots per callus. The data are subject to two sources of variation, i.e., variation between genotypes and variation between the calluses within the genotypes.

Table 5.45 In vitro culture (Conc = concentration in ppm)

Conc	Explant	No. of outbreaks	Conc	Explant	No. of outbreaks
0	1	39	50	13	54
0	2	32	50	15	35
0	3	36	50	16	50
0	4	46	50	17	51
0	5	30	50	18	38
0	6	40	50	19	61
0	7	46	100	1	46
0	8	28	100	2	55
0	9	29	100	3	54
0	10	25	100	4	49
0	11	29	100	5	55
0	12	36	100	6	55
0	13	28	100	7	47
0	14	35	100	8	42
0	15	35	100	9	38
0	16	45	100	10	50
25	1	45	100	11	46
25	2	38	100	12	42
25	3	34	100	13	44
25	4	47	100	14	30
25	5	36	100	15	38
25	6	47	100	16	31
25	7	35	100	17	42
25	8	38	200	1	36
25	9	39	200	2	37
25	10	42	200	3	27
25	11	42	200	4	38
25	12	41	200	5	25
25	13	31	200	6	29
25	14	33	200	7	30
25	15	37	200	8	30
25	16	38	200	9	37
50	1	54	200	10	28
50	2	45	200	11	37
50	3	57	200	12	29
50	4	38	200	13	36
50	5	60	200	14	34
50	6	35	200	15	27
50	7	45	200	16	32
50	8	44	200	17	37
50	9	33	200	18	30
50	10	49	200	19	31
50	11	58	200	20	30
50	12	45			

Table 5.46 Results of the experiment with earthworms

		Tillage			
		Chisel ploughing		No Tillage	
Cultivation	Manure	Entry level		Entry level	
		Low	Conventional	Low	Conventional
Corn	Yes	5, 5, 4, 2	5, 1, 5, 0	8, 4, 6, 4	14, 9, 9, 6
	No	3, 11, 0, 0	2, 0, 6, 1	2, 2, 11, 4	15, 9, 6, 4
Soy	Yes	8, 6, 0, 3	8, 4, 2, 2	2, 2, 13, 7	5, 3, 6, 0
	No	8, 5, 3, 11	2, 6, 9, 4	7, 5, 18, 3	23, 12, 17, 9

Table 5.47 Results of the callus tissue experiment

Callus						
Genotype	1	2	3	4	5	6
1	0	0	0	0	3	0
2	9	0	1	5	2	4
3	2	4	4	0	4	0
4	1	2	5	9	0	4
5	6	3	8	3	5	9
6	6	2	4	4	2	7
7	0	2	0	0	1	0
8	1	1	3	1	0	2
9	3	3	1	0	6	2
10	3	6	4	7	1	8
11	2	6	8	8	7	5
12	0	0	3	2	10	6
13	9	3	5	5	6	4
14	2	3	2	0	3	2
15	0	0	0	0	1	1
16	5	4	4	7	7	1
17	1	0	0	0	0	1
18	0	1	0	0	1	0
19	1	4	6	2	0	7
20	4	3	5	18	4	0

- (a) Write down the analysis of variance table (sources of variation and degrees of freedom).
- (b) Write down the components of the GLMM.
- (c) Analyze the dataset with the model proposed in (b).
- (d) Reanalyze the dataset using the same model as above, but, now, assume that the data have a negative binomial distribution.
- (e) Compare and contrast the results of these analyses.
- (f) Summarize the relevant results.

Exercise 5.3.4 In an experiment at the Research Institute for Animal Production “Schoonoord” in the Netherlands, the effects of active immunization against androstenedione on the fertility of Texel ewes were studied (Engel and te Brake 1993). The number of fetuses per ewe can be considered as the net result of a process that determines the number of ovulations and a probability process for these ovulations to produce fetuses. In this study, the goals are to model and analyze (a) the number of ovulations and the number of fetuses in relation to Fecundin (androstenedione-7 α -carboxyethylthioether) treatment, animal age, mating period and (b) the number of fetuses in relation to treatment, animal age, and number of ovulations observed. A summary of the experiment and a summary of the data are shown below (Table 5.48).

Of the 125 Texel ewes, 63 are treated with Fecundin, whereas the remaining 62 serve as a control group. The ewes are sorted into four age classes (e.g., <0.5, 0.5 – 1.5, 1.5 – 2.5, and > 2.5 years) and two mating periods (starting on October 1 and October 22, 1986, respectively). The interactions with age are interesting and because it is a factor, it is easier to handle than a covariate where age was entered as a factor. The number of animals in the four age classes is 25, 44, 24, and 32, respectively. The age class is evenly distributed in the combinations of mating period and treatment groups. Ewes were slaughtered at 75–80 days after the last mating, and the number of ovulations and number of fetuses were determined. Ovulation numbers ranged from 1 to 5. For six animals, the number of ovulations was not known, so these ewes were excluded from the database.

- (a) Analyze the dataset using a GLMM with the predictor: $\eta_{ijkl} = \eta + \tau_i + \alpha_j + \beta_k + (\tau\alpha)_{ij} + (\tau\beta)_{ik} + (\tau\alpha\beta)_{ijk} + b_l$, where τ , α , and β are the fixed effects of treatment, age, and mating period and b is the random effect due to animal. Assuming that each b has normal distribution with a zero mean and variance σ_b^2 , and under the assumption that the number of ovulations and the number of fetuses have a Poisson distribution.
- (b) From the analyses performed, do you observe the presence of overdispersion in the dataset? If so, propose an alternative distribution for the analysis for this dataset.
- (c) Reanalyze the dataset using the same model as before with the new data distribution.
- (d) Compare and contrast the results of these analyses.
- (e) Summarize the relevant results.

Exercise 5.3.5 The following example deals with one of the most harmful insects in the root system of the main crops, whose common name is “blind hen.” The experiment consisted of six treatments formulated for larval control in a randomized block arrangement (A, B, C, D, E, and F). The count per area shows the number of larvae in two age groups (a and b) (Table 5.49).

Table 5.48 Factors T (Treatment: 1: Fecundin; 2: Control), A (Age: 1: ≤ 0.5 ; 2: $0.5 < - \leq 1.5$; 3: $1.5 < - \leq 2.5$; 4: ≥ 2.5 years); M (Mating period: 1: October 1; 2: October 22); n (number of ovulations), and x (number of fetuses)

T	A	M	N	x	T	A	M	n	x	T	A	M	n	x	T	A	M	n	x
1	1	1	2	1	1	4	1	3	2	2	2	1	3	2	2	3	1	3	2
1	1	1	2	2	1	4	1	3	3	2	2	1	2	1	2	3	1	2	2
1	1	1	2	2	1	4	1	2	2	2	2	1	1	1	2	3	2	2	1
1	1	1	1	1	1	4	1	3	3	2	2	1	2	1	2	3	2	2	2
1	1	1	2	1	1	4	2	2	2	2	2	1	2	2	2	3	2	2	1
1	1	1	2	2	1	4	2	4	4	2	2	1	2	1	2	3	2	2	2
1	1	2	2	1	1	4	2	2	2	2	2	1	2	2	2	3	2	2	2
1	1	2	2	1	1	4	2	4	3	2	2	1	3	2	2	3	2	2	2
1	1	2	1	1	1	4	2	2	2	2	2	1	2	2	2	4	1	2	1
1	1	2	2	1	1	4	2	2	2	2	2	2	2	2	2	4	1	2	2
1	1	2	2	1	1	4	2	5	2	2	2	2	2	2	2	4	1	2	2
1	1	2	2	1	1	4	1	1	1	2	2	1	1	1	2	4	1	2	2
1	1	2	1	1	1	4	1	2	1	2	2	1	2	2	2	4	1	2	2
1	2	1	2	1	1	3	1	1	2	2	2	1	2	2	2	4	1	3	3
1	2	1	2	2	1	3	1	1	1	2	2	1	2	2	2	4	1	2	2
1	2	1	2	1	1	3	1	1	1	2	2	1	1	1	2	4	1	2	2
1	2	1	4	2	1	3	2	1	1	2	2	1	1	1	2	4	2	2	2
1	2	1	3	2	1	3	2	1	1	2	2	1	2	2	2	4	2	2	2
1	2	1	2	2	1	3	1	1	1	2	2	1	1	1	2	4	1	2	2
1	2	1	2	1	1	3	2	1	1	2	2	1	2	1	2	4	2	2	2
1	2	1	2	2	1	3	2	1	1	2	2	1	2	2	2	4	2	2	2
1	2	1	2	1	1	3	2	1	1	2	2	1	1	1	2	4	2	2	1
1	2	1	2	2	1	3	2	1	1	2	2	1	2	1	2	4	2	2	2
1	2	1	2	1	1	3	2	1	1	2	2	1	1	1	2	4	2	2	2
1	2	1	2	2	1	3	2	1	1	2	2	1	2	1	2	4	2	2	2
1	2	1	3	1	1	4	1	3	2	2	2	1	3	3	2	4	2	3	3
1	2	2	2	2	1	4	1	2	2	2	2	1	2	1	2	4	2	3	3
1	2	2	2	1	1	4	1	1	1	2	2	1	1	1	2	4	1	2	2

Table 5.49 Results of the blind hen experiment

Trt	A		B		C		D		E		F	
	A	b	A	B	A	B	A	b	A	b	A	b
1	5	7	5	14	0	3	1	7	1	10	4	13
2	4	2	12	5	2	3	1	6	3	5	4	11
3	4	4	1	14	2	2	1	7	1	8	7	10
4	1	5	5	9	2	7	3	7	0	3	3	12
5	2	2	3	8	0	0	5	4	1	6	1	8

- (a) Write down the analysis of variance table (sources of variation and degrees of freedom).
- (b) Write down the components of the GLMM.
- (c) Analyze the dataset with the model proposed in (b).
- (d) Does the proposed model in (b) adequately describe the variation observed in the dataset? Summarize the relevant results.

Appendix 1

Data: Subcultures

sub1	Rep1	NB	sub1	Rep1	NB	sub1	Rep1	NB	sub1	Rep1	NB
1	1	18	3	2	24	6	1	45	8	9	53
1	2	16	3	3	24	6	2	44	8	10	59
1	3	15	3	4	19	6	3	45	8	11	57
1	4	15	3	5	25	6	4	44	8	12	65
1	5	11	3	6	24	6	5	52	8	13	63
1	6	17	3	7	20	6	6	47	8	14	55
1	7	10	3	8	24	6	7	46	8	15	50
1	8	8	3	9	20	6	8	45	8	16	52
1	9	17	3	10	19	6	9	48	8	17	55
1	10	13	3	11	26	6	10	56	8	18	50
1	11	16	3	12	22	6	11	54	8	19	53
1	12	15	3	13	23	6	12	44	8	20	52
1	13	12	3	14	24	6	13	54	9	1	48
1	14	15	3	15	23	6	14	62	9	2	44
1	15	8	4	1	24	6	15	55	9	3	54
1	16	8	4	2	28	6	16	45	9	4	55
1	17	15	4	3	29	7	1	56	9	5	51
1	18	15	4	4	34	7	2	62	9	6	58
1	19	14	4	5	24	7	3	45	9	7	47
1	20	8	4	6	24	7	4	45	9	8	42
2	1	15	4	7	25	7	5	46	9	9	50
2	2	11	4	8	28	7	6	48	9	10	48
2	3	12	4	9	24	7	7	55	9	11	48
2	4	18	4	10	32	7	8	45	9	12	53

(continued)

sub1	Rep1	NB	sub1	Rep1	NB	sub1	Rep1	NB	sub1	Rep1	NB
2	5	8	4	11	34	7	9	45	9	13	54
2	6	17	4	12	30	7	10	44	9	14	59
2	7	8	4	13	26	7	11	52	9	15	58
2	8	18	4	14	27	7	12	45	10	1	46
2	9	22	4	15	29	7	13	43	10	2	38
2	10	19	5	1	38	7	14	58	10	3	29
2	11	19	5	2	38	7	15	62	10	4	30
2	12	24	5	3	37	7	16	45	10	5	31
2	13	12	5	4	41	7	17	63	10	6	33
2	14	12	5	5	46	7	18	56	10	7	35
2	15	11	5	6	44	7	19	55	10	8	59
2	16	21	5	7	54	7	20	50	10	9	37
2	17	10	5	8	45	8	1	53	10	10	44
2	18	15	5	9	60	8	2	58	10	11	42
2	19	20	5	10	57	8	3	56	10	12	41
2	20	22	5	11	51	8	4	50	10	13	45
2	21	20	5	12	54	8	5	57	10	14	38
2	22	13	5	13	51	8	6	60	10	15	40
2	23	18	5	14	62	8	7	50			
3	1	19	5	15	53	8	8	52			

Data: Beatles

Row	Column	Treatment	Count
1	1	S	3
1	2	U	6
1	3	U	2
1	4	TR	7
1	5	S	1
1	6	TR	5
2	1	TR	5
2	2	S	4
2	3	TR	5
2	4	U	8
2	5	U	6
2	6	S	3
3	1	U	3
3	2	TR	6
3	3	U	4
3	4	S	3
3	5	S	4
3	6	TR	7
4	1	U	3
4	2	TR	4

(continued)

Row	Column	Treatment	Count
4	3	TR	5
4	4	S	2
4	5	U	3
4	6	S	6
5	1	TR	8
5	2	S	5
5	3	S	6
5	4	U	7
5	5	TR	9
5	6	U	4
6	1	S	6
6	2	U	5
6	3	S	6
6	4	TR	9
6	5	TR	9

Data: Weed counts

Block	A	B	Count
1	1	1	14
1	1	2	7
1	1	3	5
1	1	4	7
1	2	1	5
1	2	2	14
1	2	3	5
1	2	4	9
1	3	1	0
1	3	2	0
1	3	3	10
1	3	4	13
1	4	1	20
1	4	2	53
1	4	3	21
1	4	4	7
1	5	1	12
1	5	2	31
1	5	3	32
1	5	4	22
1	6	1	49
1	6	2	16
1	6	3	7
1	6	4	14
1	7	1	20

(continued)

Block	A	B	Count
1	7	2	20
1	7	3	16
1	7	4	6
2	1	1	9
2	1	2	9
2	1	3	9
2	1	4	19
2	2	1	31
2	2	2	11
2	2	3	30
2	2	4	29
2	3	1	25
2	3	2	11
2	3	3	15
2	3	4	23
2	4	1	7
2	4	2	22
2	4	3	20
2	4	4	3
2	5	1	0
2	5	2	28
2	5	3	18
2	5	4	18
2	6	1	55
2	6	2	58
2	6	3	18
2	6	4	19
2	7	1	14
2	7	2	44
2	7	3	19
2	7	4	17
3	1	1	12
3	1	2	8
3	1	3	44
3	1	4	0
3	2	1	29
3	2	2	11
3	2	3	5
3	2	4	49
3	3	1	99
3	3	2	66
3	3	3	11
3	3	4	15

(continued)

Block	A	B	Count
3	4	1	9
3	4	2	8
3	4	3	9
3	4	4	21
3	5	1	49
3	5	2	49
3	5	3	17
3	5	4	22
3	6	1	41
3	6	2	21
3	6	3	48
3	6	4	11
3	7	1	58
3	7	2	34
3	7	3	28
3	7	4	20
4	1	1	6
4	1	2	9
4	1	3	20
4	1	4	0
4	2	1	10
4	2	2	0
4	2	3	7
4	2	4	9
4	3	1	9
4	3	2	29
4	3	3	22
4	3	4	4
4	4	1	22
4	4	2	31
4	4	3	32
4	4	4	41
4	5	1	112
4	5	2	44
4	5	3	24
4	5	4	28
4	6	1	8
4	6	2	8
4	6	3	11
4	6	4	10
4	7	1	117
4	7	2	78
4	7	3	36
4	7	4	38

Coffee data

Shade	Clone	Tray	Rep	y1	y2	y3	y4	y5	y6	y7	y8	y9	y10	y11
Roja	C1	CHI	R1	2	2	4	4	6	4	5	5	5	4	7
Roja	C1	CHI	R1	2	2	4	4	2	2	4	2	4	8	12
Roja	C1	CHI	R1	2	0	4	5	8	7	5	4	4	6	5
Roja	C2	CHI	R1	0	2	5	2	6	3	9	8	8	9	8
Roja	C2	CHI	R1	0	0	3	2	6	6	7	6	7	5	6
Roja	C2	CHI	R1	2	0	2	2	6	6	6	5	5	10	14
Roja	C3	CHI	R1	2	0	4	5	8	3	3	6	2	11	9
Roja	C3	CHI	R1	0	2	6	6	8	10	10	10	9	5	9
Roja	C3	CHI	R1	2	2	4	5	6	7	7	5	5	5	7
Roja	C4	CHI	R1	2	2	6	5	8	9	7	4	3	10	6
Roja	C4	CHI	R1	0	2	4	4	8	8	4	4	4	5	7
Roja	C4	CHI	R1	2	2	4	6	8	6	6	4	4	4	8
Roja	C5	CHI	R1	2	1	5	7	8	6	4	4	2	11	13
Roja	C5	CHI	R1	2	2	4	5	6	4	7	7	8	6	11
Roja	C5	CHI	R1	0	2	6	8	10	10	8	8	7	4	2
Roja	pf	CHI	R1	2	2	6	5	8	7	6	2	2	6	12
Roja	pf	CHI	R1	2	2	4	5	8	8	5	3	2	9	12
Roja	pf	CHI	R1	2	4	6	7	8	10	11	11	10	11	13
Roja	C1	CHI	R2	2	2	5	6	8	9	6	6	6	8	9
Roja	C1	CHI	R2	2	2	4	5	8	10	6	4	2	2	
Roja	C1	CHI	R2	2	2	3	5	6	8	5	4	4	6	7
Roja	C2	CHI	R2	0	2	4	6	8	10	6	7	7	9	8
Roja	C2	CHI	R2	0	2	4	6	8	10	7	4	3	7	7
Roja	C2	CHI	R2	2	2	6	5	2	8	4	5	4	4	4
Roja	C3	CHI	R2	2	0	4	6	8	10	9	8	6	13	13

(continued)

Coffee data (continued)

Shade	Clone	Tray	Rep	y1	y2	y3	y4	y5	y6	y7	y8	y9	y10	y11
Roja	C3	CHI	R2	0	0	4	4	8	10	10	3	3	5	15
Roja	C3	CHI	R2	0	2	5	4	8	10	8	4	2	11	5
Roja	C4	CHI	R2	2	2	4	6	6	10	3	2	2	3	8
Roja	C4	CHI	R2	2	2	6	6	6	6	4	3	7	7	7
Roja	C4	CHI	R2	0	2	5	6	8	10	6	3	3	1	10
Roja	C5	CHI	R2	2	2	6	6	10	8	5	4	3	1	9
Roja	C5	CHI	R2	2	2	4	4	8	10	8	10	12	11	9
Roja	C5	CHI	R2	2	2	4	6	8	10	6	3	3	9	6
Roja	pf	CHI	R2	2	2	4	6	8	8	7	8	12	13	12
Roja	pf	CHI	R2	0	2	4	6	8	10	10	5	5	11	11
Roja	pf	CHI	R2	2	2	4	6	8	8	8	4	4	14	11
Roja	C1	CHI	R3	2	2	6	6	6	8	5	7	1	5	12
Roja	C1	CHI	R3	2	2	4	6	6	8	8	1	7	11	9
Roja	C1	CHI	R3	2	2	6	4	6	8	10	8	7	8	9
Roja	C2	CHI	R3	2	2	4	6	8	10	4	3	6	10	11
Roja	C2	CHI	R3	2	0	2	4	4	4	2	6			
Roja	C2	CHI	R3	2	2	2	2	2	4	2	2	4	6	6
Roja	C3	CHI	R3	2	0	4	4	6	6	3	2	2	8	10
Roja	C3	CHI	R3	2	2	4	6	8	6	2	8	1	4	6
Roja	C3	CHI	R3	0	0	2	4	6	6	4	2	3	8	12
Roja	C4	CHI	R3	2	2	4	4	6	4	2	2	2	7	9
Roja	C4	CHI	R3	2	2	4	6	8	8	7	7	4	1	2
Roja	C4	CHI	R3	2	2	4	5	6	6	5	2	4	8	8
Roja	C5	CHI	R3	2	4	6	5	4	6	2	2	6	12	12
Roja	C5	CHI	R3	2	4	4	6	8	10	9	2	4	8	12

Roja	C5	CHI	R3	2	2	6	6	8	10	12	9	10	11	11
Roja	pf	CHI	R3	2	2	4	5	4	10	10	2	4	11	16
Roja	pf	CHI	R3	2	2	6	6	8	8	6	3	3	12	16
Roja	pf	CHI	R3	2	3	6	6	8	8	6	5	5	18	23
Roja	C1	CHI	R4	2	2	4	5	8	8	2	3	3	2	4
Roja	C1	CHI	R4	2	2	6	6	10	8	10	2	2	11	11
Roja	C1	CHI	R4	0	2	6	6	8	8	12	3	2	2	10
Roja	C2	CHI	R4	2	2	2	2	4	6	6	1			
Roja	C2	CHI	R4	2	0	2	2	4	3	3	2	6	3	3
Roja	C2	CHI	R4	2	2	2	2	2	4	6	6	6	8	9
Roja	C3	CHI	R4	2	0	3	2	4	4	3	1	1	1	2
Roja	C3	CHI	R4	2	0	4	4	8	10	4	8	6	4	6
Roja	C3	CHI	R4	2	0	2	4	6	7	7	6	7	4	5
Roja	C4	CHI	R4	2	0	4	5	4	2	1	1	2	1	1
Roja	C4	CHI	R4	2	0	2	2	2	4	5	2	2	6	7
Roja	C4	CHI	R4	2	2	4	5	6	8	8	6	5	9	9
Roja	C5	CHI	R4	2	4	4	5	10	10	6	5	7	10	5
Roja	C5	CHI	R4	2	0	4	4	4	6	3	3	2	1	1
Roja	C5	CHI	R4	2	0	5	6	8	10	9	4	3	3	2
Roja	pf	CHI	R4	0	4	6	5	4	8	2	3		11	11
Roja	pf	CHI	R4	0	2	4	5	8	8	7	4	4	3	6
Roja	pf	CHI	R4	2	4	6	6	8	10	8	3	3	16	16
Roja	C1	CH2	R1	2	2	3	5	6	8	5	10	4	10	14
Roja	C1	CH2	R1	2	2	4	6	10	10	11	11	9	9	8
Roja	C1	CH2	R1	2	2	4	6	8	10	11	10	11	13	14
Roja	C2	CH2	R1	2	2	4	4	8	8	2	1	1	7	11
Roja	C2	CH2	R1	2	2	4	4	6	6	3	5	5	1	6

(continued)

Coffee data (continued)

Shade	Clone	Tray	Rep	y1	y2	y3	y4	y5	y6	y7	y8	y9	y10	y11
Roja	C2	CH2	R1	2	2	6	6	8	8	9	2	2	2	1
Roja	C3	CH2	R1	2	2	4	6	8	10	7	1	1	1	
Roja	C3	CH2	R1	2	2	6	6	10	10	11	6	5	7	14
Roja	C3	CH2	R1	2	2	4	6	8	10	9	6	6		
Roja	C4	CH2	R1	2	0	4	4	8	10	6	6	5	8	9
Roja	C4	CH2	R1	2	2	4	6	8	10	12	12	14	14	
Roja	C4	CH2	R1	2	2	4	6	8	10	10	10	9	9	11
Roja	C5	CH2	R1	2	2	6	6	10	10	9	5	4	10	9
Roja	C5	CH2	R1	2	2	6	6	10	10	12	9	9	10	15
Roja	C5	CH2	R1	0	2	6	6	8	8	10	8	5	7	16
Roja	pf	CH2	R1	2	2	5	6	8	10	10	12	9	9	10
Roja	pf	CH2	R1	2	2	6	6	8	12	6	5	7	8	7
Roja	pf	CH2	R1	2	4	4	7	10	11	10	9	8		
Roja	C1	CH2	R2	2	0	4	6	8	8	4	7	5	3	4
Roja	C1	CH2	R2	2	0	4	6	8	8	5	6	6	10	11
Roja	C1	CH2	R2	2	2	6	6	8	10	10	3	3	3	5
Roja	C2	CH2	R2	2	2	4	6	6	10	4	2	2	8	9
Roja	C2	CH2	R2	2	2	4	6	8	10	12			1	
Roja	C2	CH2	R2	2	0	2	4	6	8	10	2	2		
Roja	C3	CH2	R2	0	2	6	4	6	10	10				
Roja	C3	CH2	R2	2	0	2	4	4	8	4			5	6
Roja	C3	CH2	R2	2	2	4	6	6	10	10				
Roja	C4	CH2	R2	2	2	4	4	8	10	4	2	2	10	9
Roja	C4	CH2	R2	2	0	4	4	8	8	2	5	9	7	8
Roja	C4	CH2	R2	2	2	4	4	8	8	8	8	3	7	12

Roja	C5	CH2	R2	0	2	4	6	8	8	6	6	5	3	4
Roja	C5	CH2	R2	2	2	6	6	8	10	10	10	8	13	14
Roja	C5	CH2	R2	2	2	6	6	10	10	7	6	5	4	6
Roja	pf	CH2	R2	2	4	6	6	8	12	4	1	1	3	4
Roja	pf	CH2	R2	2	3	6	5	8	10	8	5	6	5	11
Roja	pf	CH2	R2	2	2	4	6	8	10	10	4	4	5	8
Roja	C1	CH2	R3	2	2	6	6	10	10	4	2			
Roja	C1	CH2	R3	2	2	4	6	10	10	2	1	1	1	4
Roja	C1	CH2	R3	2	2	4	5	8	10	6	3	2	1	
Roja	C2	CH2	R3	2	2	6	6	8	10	3	1	1		
Roja	C2	CH2	R3	0	2	4	5	8	12	10				
Roja	C2	CH2	R3	2	0	4	6	8	10	8	9	5	7	8
Roja	C3	CH2	R3	2	2	4	6	8	12	10				
Roja	C3	CH2	R3	2	2	6	6	10	10	10	1			
Roja	C3	CH2	R3	2	2	6	6	10	10	10	2	1	8	11
Roja	C4	CH2	R3	2	2	4	6	8	8	4				
Roja	C4	CH2	R3	2	2	6	5	8	6	6	2	1		
Roja	C4	CH2	R3	2	2	6	6	8	8	4	1	1	1	
Roja	C5	CH2	R3	2	2	6	5	8	10	6	3	3	12	10
Roja	C5	CH2	R3	2	0	4	4	8	10	6	3	2	1	2
Roja	C5	CH2	R3	2	2	6	8	8	10	6	4	3	3	2
Roja	pf	CH2	R3	2	3	5	6	10	12	11	7	7	11	11
Roja	pf	CH2	R3	2	3	5	6	10	12	6	1	1	12	5
Roja	pf	CH2	R3	2	2	6	6	10	10	5	6	7	12	14
Roja	C1	CH2	R4	2	0	4	5	4	8	10	2	2	5	6
Roja	C1	CH2	R4	2	2	4	6	8	10	8	6	5	6	9
Roja	C1	CH2	R4	2	2	4	6	8	10	12	8	12	14	18

(continued)

Coffee data (continued)

Shade	Clone	Tray	Rep	y1	y2	y3	y4	y5	y6	y7	y8	y9	y10	y11
Roja	C2	CH2	R4	2	2	6	6	8	8	8	4	3	7	13
Roja	C2	CH2	R4	2	0	5	6	8	8	5	4	6	10	9
Roja	C2	CH2	R4	0	2	4	4	6	8	8	6	6	4	4
Roja	C3	CH2	R4	2	2	6	6	8	8	8	3	3	10	6
Roja	C3	CH2	R4	2	0	4	6	8	10	12	9	9	7	9
Roja	C3	CH2	R4	2	2	6	6	10	9	11	9	9	8	18
Roja	C4	CH2	R4	2	2	6	6	10	8	9	8	7	5	10
Roja	C4	CH2	R4	2	2	6	6	8	10	10	7	6	7	12
Roja	C4	CH2	R4	2	0	4	6	8	10	9	9	6	4	8
Roja	C5	CH2	R4	0	2	6	6	10	8	8	7	1	7	10
Roja	C5	CH2	R4	2	0	4	4	8	8	6	2	1	3	4
Roja	C5	CH2	R4	2	2	6	8	8	10	10	7	8	7	7
Roja	pf	CH2	R4	2	3	4	5	8	8	8	5	5	3	11
Roja	pf	CH2	R4	2	0	4	6	8	10	9	7	7	6	14
Roja	pf	CH2	R4	2	3	7	7	10	10	9	10	10	12	20
Roja	C1	CH3	R1	2	0	4	6	7	10	9	7	4	4	13
Roja	C1	CH3	R1	2	2	5	6	8	8	10	6	7	7	4
Roja	C1	CH3	R1	2	2	2	4	6	8	7	2	3	5	7
Roja	C2	CH3	R1	2	0	4	6	6	8	8	9	9	9	8
Roja	C2	CH3	R1	2	0	5	5	6	8	10	4	4	2	2
Roja	C2	CH3	R1	2	2	3	5	7	8	8	9	9	9	10
Roja	C3	CH3	R1	2	2	4	6	9	8	10	10	6	5	8
Roja	C3	CH3	R1	2	0	4	4	8	8	5	5	3	3	10
Roja	C3	CH3	R1	2	2	4	6	8	10	9	7	9	11	14
Roja	C4	CH3	R1	2	2	4	4	8	8	9	7	6	7	5

Roja	C4	CH3	R1	2	0	5	4	9	8	7	8		
Roja	C4	CH3	R1	2	2	4	6	8	8	10	7	2	
Roja	C5	CH3	R1	2	2	4	6	8	10	9	6	5	3
Roja	C5	CH3	R1	2	2	4	4	6	8	5	5	7	10
Roja	C5	CH3	R1	0	0	6	6	8	10	5	4		
Roja	pf	CH3	R1	0	2	6	6	8	10	10	6	3	6
Roja	pf	CH3	R1	2	2	4	6	9	10	8	7	8	6
Roja	pf	CH3	R1	2	2	4	6	8	10	6	4	4	1
Roja	C1	CH3	R2	2	0	4	6	8	8	10	5	4	4
Roja	C1	CH3	R2	2	0	4	6	8	8	3	10	9	10
Roja	C1	CH3	R2	2	0	4	6	8	10	12	7	7	14
Roja	C2	CH3	R2	2	2	4	6	8	10	12	11	11	12
Roja	C2	CH3	R2	2	2	2	2	6	8	8	9	9	8
Roja	C2	CH3	R2	2	0	4	4	8	8	6	9		
Roja	C3	CH3	R2	2	2	4	6	10	8	8	1	3	5
Roja	C3	CH3	R2	2	1	6	6	10	10	10	12	11	17
Roja	C3	CH3	R2	2	2	6	8	10	10	9	11	8	12
Roja	C4	CH3	R2	2	2	4	6	8	8	7			
Roja	C4	CH3	R2	2	2	6	6	9	8	2	2		
Roja	C4	CH3	R2	2	2	4	6	8	8	8	6	5	10
Roja	C5	CH3	R2	0	2	6	6	10	10	12	8	8	12
Roja	C5	CH3	R2	2	2	4	6	8	8	8			
Roja	C5	CH3	R2	2	2	6	6	8	10	10	10	8	9
Roja	pf	CH3	R2	2	2	4	6	8	10	10	11	12	15
Roja	pf	CH3	R2	2	0	4	6	8	10	8	9	10	5
Roja	pf	CH3	R2	2	2	4	6	8	10	10	9	10	14
Roja	C1	CH3	R3	2	2	5	6	8	10	10	10	9	12

(continued)

Coffee data (continued)

Shade	Clone	Tray	Rep	y1	y2	y3	y4	y5	y6	y7	y8	y9	y10	y11
Roja	C1	CH3	R3	2	2	4	5	9	10	10	6	5	3	18
Roja	C1	CH3	R3	2	2	5	5	9	12	6		2		8
Roja	C2	CH3	R3	2	2	6	4	9	8	6	1	1		
Roja	C2	CH3	R3	2	2	4	4	8	8	10	12	9	9	15
Roja	C2	CH3	R3	2	2	3	4	6	8	8				
Roja	C3	CH3	R3	2	2	5	6	8	10	11	13	13	11	12
Roja	C3	CH3	R3	2	2	6	6	8	11	4	8	6	3	11
Roja	C3	CH3	R3	2	2	4	6	6	12	7	8	4	1	11
Roja	C4	CH3	R3	2	2	4	5	7	8	4				
Roja	C4	CH3	R3	2	2	4	6	8	10	5	3	3	1	2
Roja	C4	CH3	R3	2	2	4	6	8	6	6	2	2	2	5
Roja	C5	CH3	R3	2	2	4	6	8	10	4	2	2	2	17
Roja	C5	CH3	R3	2	2	4	6	8	10	12	8	7	10	4
Roja	C5	CH3	R3	2	2	5	5	8	10	4	5			
Roja	pf	CH3	R3	2	2	4	6	8	8	3	10	8	1	2
Roja	pf	CH3	R3	2	2	4	6	8	8	5	2	1	10	14
Roja	pf	CH3	R3	2	2	4	6	8	10	6	3	2	6	14
Roja	C1	CH3	R4	2	2	6	6	8	8	8	4	3	3	6
Roja	C1	CH3	R4	2	2	4	5	8	12	6	10	6	7	7
Roja	C1	CH3	R4	2	0	4	6	8	8	10	3	1	6	8
Roja	C2	CH3	R4	2	0	4	6	8	10	12	9	8	16	18
Roja	C2	CH3	R4	2	0	4	6	8	9	11	10	6	10	6
Roja	C2	CH3	R4	2	2	6	6	7	10	6	4	1		
Roja	C3	CH3	R4	2	2	6	6	8	10	11	7	6	13	22
Roja	C3	CH3	R4	2	2	5	5	7	8	5	9	6	16	23

Roja	C3		CH3	R4	2	2	6	6	8	10	10	10	10	11	14	13
Roja	C4		CH3	R4	2	2	4	7	8	12	6	4	4	4	7	11
Roja	C4		CH3	R4	2	2	6	8	10	10	12	5	4	4		
Roja	C4		CH3	R4	2	1	6	8	9	10	8	3			2	2
Roja	C5		CH3	R4	2	2	6	8	10	8	5	9	10	10	14	9
Roja	C5		CH3	R4	2	2	6	8	10	10	7	5	10	10	12	14
Roja	C5		CH3	R4	2	0	6	8	8	10	10	10	9	9	14	16
Roja	pf		CH3	R4	2	2	4	5	8	10	12	10	10	10	21	24
Roja	pf		CH3	R4	2	2	4	6	8	10	12	10	11	11	11	23
Roja	pf		CH3	R4	0	0	5	6	8	6	8	9	7	7	9	13
Perla	C1		CHI	R1	2	0	4	3	2	2	1					
Perla	C1		CHI	R1	2	0	4	5	7	5	1					
Perla	C1		CHI	R1	2	0	2	2	2	4						
Perla	C2		CHI	R1	2	0	2	2	4	5	6	2	2	2	1	1
Perla	C2		CHI	R1	2	2	6	4	4	4						
Perla	C2		CHI	R1	2	0	3	2	4	4						
Perla	C3		CHI	R1	2	2	4	5	8	6	2					
Perla	C3		CHI	R1	2	2	4	2	3	2	1	4				
Perla	C3		CHI	R1	2	0	4	5	7	4	3	2	1	2	2	6
Perla	C4		CHI	R1	2	0	2	2	2	7	4	1				
Perla	C4		CHI	R1	2	3	4	3	6	3	3	3	3	3		
Perla	C4		CHI	R1	2	2	4	5	7	6	3	3				
Perla	C5		CHI	R1	2	2	4	6	5	6	4	2	1	1	4	8
Perla	C5		CHI	R1	2	2	4	4	7	8						
Perla	C5		CHI	R1	2	2	4	5	7	9	7	7	6	7	6	12
Perla	pf		CHI	R1	2	2	4	4	7	9	9	4	4	7	4	14
Perla	pf		CHI	R1	2	0	6	4	6	4		2	2	5	5	14
Perla	pf		CHI	R1	2	2	4	6	8	6	2	2	1	1	4	6

(continued)

Coffee data (continued)

Shade	Clone	Tray	Rep	y1	y2	y3	y4	y5	y6	y7	y8	y9	y10	y11
Perla	pf	CHI	R1	2	2	6	6	7	3	3	3	3	3	2
Perla	C1	CHI	R2	2	2	4	3	2	8	6	1	2	10	10
Perla	C1	CHI	R2	2	2	4	6	8	8	8				
Perla	C1	CHI	R2	2	2	4	3	4	2	2	1	2	5	5
Perla	C2	CHI	R2	0	0	3	3	4	6	8				
Perla	C2	CHI	R2	0	0	2	2	2	8	5				
Perla	C2	CHI	R2	2	0	2	1	2	2	3				
Perla	C3	CHI	R2	2	2	4	4	4	2	3	1		6	9
Perla	C3	CHI	R2	2	2	4	4	8	6	4	2	2	5	6
Perla	C3	CHI	R2	2	0	4	3	4	4	2	2	2	5	5
Perla	C4	CHI	R2	2	2	4	6	6	7	1	11	1	8	11
Perla	C4	CHI	R2	2	0	4	4	5	4	4	1		3	11
Perla	C4	CHI	R2	2	2	4	1	4	6	3	2		2	2
Perla	C5	CHI	R2	2	2	4	4	4	5	4	5	3	3	4
Perla	C5	CHI	R2	2	2	6	6	6	7	5	3	2	3	5
Perla	C5	CHI	R2	2	0	5	3	4	5	4	2	4	2	5
Perla	pf	CHI	R2	2	3	4	4	5	5	2	1			
Perla	pf	CHI	R2	2	2	4	5	5	7	5	2	3	8	16
Perla	pf	CHI	R2	2	2	5	6	8	4	4		2	1	2
Perla	C1	CHI	R3	2	2	6	6	8	7	4	1	1	1	5
Perla	C1	CHI	R3	2	2	2	2	4	5	2	2	2	1	
Perla	C1	CHI	R3	2	0	4	4	4	6	1				7
Perla	C2	CHI	R3	2	2	6	7	4	2	9	7	7	9	12
Perla	C2	CHI	R3	2	2	4	6	8	8	3	4			
Perla	C2	CHI	R3	2	2	4	4	6	4	4		3	6	7

Perla	C3	CHI	R3	2	0	3	3	3	2	2	2	4	4	6
Perla	C3	CHI	R3	2	0	4	6	6	5	5	4	8	9	6
Perla	C3	CHI	R3	2	2	4	5	8	4	5	1	2	2	10
Perla	C4	CHI	R3	2	2	6	5	9	4	3	1	1	5	9
Perla	C4	CHI	R3	2	2	4	7	7	5	5	5	5	5	8
Perla	C4	CHI	R3	2	2	4	8	6	5	2	2	2	3	2
Perla	C5	CHI	R3	2	2	4	6	7	9	5	1	1	5	12
Perla	C5	CHI	R3	2	0	4	6	8	5	4	3	2	6	10
Perla	C5	CHI	R3	2	4	4	5	7	3	5	3	3	6	7
Perla	pf	CHI	R3	2	2	6	7	5	4	3				
Perla	pf	CHI	R3	2	4	3	6	7	7	6	6	6	10	11
Perla	pf	CHI	R3	2	4	6	5	5	4	3	1	1		
Perla	C1	CHI	R4	2	2	4	4	4	4	4				
Perla	C1	CHI	R4	2	0	4	3	5	3					
Perla	C1	CHI	R4	2	0	4	3	6	4	6				
Perla	C2	CHI	R4	2	0	2	2	2	2					
Perla	C2	CHI	R4	2	0	4	4	6	5	5	5	4	4	4
Perla	C2	CHI	R4	2	2	4	4	7	4	3		5	2	2
Perla	C3	CHI	R4	2	0	4	4	7	6					
Perla	C3	CHI	R4	2	0	2	2	4	2	3				
Perla	C3	CHI	R4	2	0	3	3	3	4					
Perla	C4	CHI	R4	2	0	2	2	2	4	4	2	9	12	
Perla	C4	CHI	R4	2	0	3	4	6	3					
Perla	C4	CHI	R4	2	2	4	4	7	3	1		2	4	
Perla	C5	CHI	R4	2	4	10	7	9	8	8	8	7	8	9
Perla	C5	CHI	R4	2	2	4	4	7	8	6	4	3	3	3
Perla	C5	CHI	R4	2	2	4	5	7	7	5	3	3	8	13

(continued)

Coffee data (continued)

Shade	Clone	Tray	Rep	y1	y2	y3	y4	y5	y6	y7	y8	y9	y10	y11
Perla	pf	CH1	R4	2	0	4	5	6	4	3	2	2	1	10
Perla	pf	CH1	R4	2	0	4	4	6	4	2	1	1	1	4
Perla	pf	CH1	R4	2	2	6	6	8	2		4	3	2	4
Perla	C1	CH2	R1	2	2	4	6	8	9	9	2	12	12	12
Perla	C1	CH2	R1	2	2	4	6	8	8	9				
Perla	C1	CH2	R1	2	2	4	4	8	8	5				
Perla	C2	CH2	R1	2	2	6	6	10	10	9	12	12	15	16
Perla	C2	CH2	R1	2	0	4	4	8	8	4	9	11	12	12
Perla	C2	CH2	R1	2	0	4	4	7	6	9	9	10	12	11
Perla	C3	CH2	R1	2	2	4	6	8	6	8	4	4	5	9
Perla	C3	CH2	R1	2	2	4	6	8	8	6				
Perla	C3	CH2	R1	2	2	4	4	8	8					
Perla	C4	CH2	R1	2	0	4	4	8	8	10	10	10	14	15
Perla	C4	CH2	R1	2	2	2	2	4	5	6				
Perla	C4	CH2	R1	2	2	6	6	10	10	12	3			
Perla	C5	CH2	R1	2	0	4	6	8	10	4	1		7	12
Perla	C5	CH2	R1	2	2	4	6	5	7	1				
Perla	C5	CH2	R1	2	0	6	6	10	8	11				
Perla	pf	CH2	R1	2	2	4	4	6	9	2	1	1	5	11
Perla	pf	CH2	R1	2	2	4	6	8	4	6	2	1	7	9
Perla	pf	CH2	R1	2	2	4	6	10	7	11	6	6	9	13
Perla	C1	CH2	R2	2	2	6	6	10	8	10	4	1		
Perla	C1	CH2	R2	2	2	4	6	10	5	3	1	1	1	10
Perla	C1	CH2	R2	2	2	4	6	10	8	5	2	1	1	2
Perla	C2	CH2	R2	2	0	4	6	8	8	8	10			

Perla	C2	CH2	R2	2	0	4	6	7	9	7	3	7	7	6	11
Perla	C2	CH2	R2	2	2	4	6	8	10	8	6	6	6	6	3
Perla	C3	CH2	R2	2	2	4	6	10	8	9	5	4	3	3	10
Perla	C3	CH2	R2	2	2	6	6	10	10	8	6	5	8	8	11
Perla	C3	CH2	R2	2	2	4	6	8	8	9	6	7	9	9	9
Perla	C4	CH2	R2	2	2	6	6	9	9	8	8	8	6	6	9
Perla	C4	CH2	R2	2	0	6	6	10	10	10	6	7	7	7	9
Perla	C4	CH2	R2	2	1	4	7	8	10	7	4	4	5	7	7
Perla	C5	CH2	R2	0	2	4	5	10	8	10	6	7	9	9	12
Perla	C5	CH2	R2	2	2	4	6	10	6	5	5				
Perla	C5	CH2	R2	2	1	4	4	8	8	9	1				
Perla	pf	CH2	R2	2	0	4	6	10	9	6	6	5	4	6	6
Perla	pf	CH2	R2	2	2	4	6	10	4	3	1	1	7	7	7
Perla	pf	CH2	R2	2	5	6	9	11	10	6	3	1	5	6	6
Perla	C1	CH2	R3	2	0	4	6	8	7	5	5	3			
Perla	C1	CH2	R3	2	2	4	6	8	8	5	5	2	8	9	9
Perla	C1	CH2	R3	2	2	6	6	8	8	6	4	2	4	6	6
Perla	C2	CH2	R3	2	0	4	3	7	7						
Perla	C2	CH2	R3	0	0	4	4	8	7						
Perla	C2	CH2	R3	2	2	4	6	8	6	2					
Perla	C3	CH2	R3	2	2	4	4	8	8	5	3	3	3	6	6
Perla	C3	CH2	R3	2	0	4	4	7	5	9	2	2	1	6	6
Perla	C3	CH2	R3	2	0	4	4	8	8	8	4	3	9		
Perla	C4	CH2	R3	2	2	4	8	8	7	9	2	2			
Perla	C4	CH2	R3	2	2	4	8	8	8	1	2	1	5	6	6
Perla	C4	CH2	R3	2	0	4	6	8	4	3	1	1	3	6	6
Perla	C5	CH2	R3	2	0	4	4	8	8	10	5	5	5	11	11

(continued)

Coffee data (continued)

Shade	Clone	Tray	Rep	y1	y2	y3	y4	y5	y6	y7	y8	y9	y10	y11
Perla	C5	CH2	R3	2	0	2	2	6	9	9	5	6	7	7
Perla	C5	CH2	R3	2	0	4	6	8	10	9	2	1	1	2
Perla	pf	CH2	R3	2	2	4	7	12	9	4	4	3	8	12
Perla	pf	CH2	R3	2	2	4	6	8	6	5	1	4	7	10
Perla	pf	CH2	R3	2	2	4	8	8	4	1	5			
Perla	C1	CH2	R4	2	0	4	5	2	7	5	5	4	2	10
Perla	C1	CH2	R4	2	2	4	6	10	9	3	2	2	4	10
Perla	C1	CH2	R4	2	2	4	7	5	3	1	1	1	5	12
Perla	C2	CH2	R4	2	0	4	6	5	6	1	1	1	6	6
Perla	C2	CH2	R4	2	0	4	6	8	6	2				
Perla	C2	CH2	R4	0	0	4	6	7	5					
Perla	C3	CH2	R4	2	0	2	4	4	7	7	1			
Perla	C3	CH2	R4	2	2	2	4	6	8					
Perla	C3	CH2	R4	2	2	4	6	10	7	1		1	4	7
Perla	C4	CH2	R4	2	2	4	4	5	5	7	3	1	5	8
Perla	C4	CH2	R4	2	2	4	6	8	8	8	2	2		
Perla	C4	CH2	R4	2	2	4	6	7	6	4	3	3		
Perla	C5	CH2	R4	2	0	4	6	9	5	5	2		6	7
Perla	C5	CH2	R4	2	2	4	6	2	4		1			
Perla	C5	CH2	R4	2	2	4	5	6	8	8	9	6		
Perla	pf	CH2	R4	2	2	4	6	8	8	4				
Perla	pf	CH2	R4	2	2	4	6	8	5	7				
Perla	pf	CH2	R4	2	4	6	8	10	8	5	5	3		
Perla	C1	CH3	R1	2	0	4	6	8	8	7	9	7	7	9
Perla	C1	CH3	R1	2	2	2	6	8	7	9	9	8	7	13

Perla	C1	CH3	R1	2	0	4	6	7	10	9	12	10	11	16
Perla	C2	CH3	R1	2	2	4	4	8	6	10	9	11	11	13
Perla	C2	CH3	R1	2	2	4	6	8	8	9	9	10	11	11
Perla	C2	CH3	R1	2	2	4	6	8	5	9	8	9	10	11
Perla	C3	CH3	R1	2	2	4	6	8	10	11	3	3	2	4
Perla	C3	CH3	R1	2	0	4	6	8	9	10	10	8	8	10
Perla	C3	CH3	R1	2	0	4	6	8	9	11	7	7	8	12
Perla	C4	CH3	R1	2	0	4	6	8	6	8	5	5	6	8
Perla	C4	CH3	R1	2	2	4	4	8	8	8	6	6	2	8
Perla	C4	CH3	R1	2	2	2	4	6	5	6	4	3	4	12
Perla	C5	CH3	R1	2	0	4	6	10	7	11	12	8	10	16
Perla	C5	CH3	R1	2	2	4	6	10	10	12	11	11	11	11
Perla	C5	CH3	R1	2	2	4	6	10	8	9	5	4	10	11
Perla	pf	CH3	R1	2	2	4	6	8	8	6	10	11	11	17
Perla	pf	CH3	R1	2	2	4	6	8	10	12	7	6	14	20
Perla	pf	CH3	R1	2	2	2	6	8	8	9	3	3	3	5
Perla	C1	CH3	R2	2	2	4	6	8	10	12	12	11	10	14
Perla	C1	CH3	R2	2	2	4	6	9	5	5	8	8	9	14
Perla	C1	CH3	R2	2	0	4	4	8	8	10	10	8	10	16
Perla	C2	CH3	R2	2	4	4	6	10	8	5	9	8	12	16
Perla	C2	CH3	R2	2	0	3	4	7	7	7	7	5	6	9
Perla	C2	CH3	R2	2	0	4	4	8	8	10	9	10	10	15
Perla	C3	CH3	R2	2	2	4	6	8	8	10	10	10	12	14
Perla	C3	CH3	R2	2	0	4	4	8	8	10	6	6	9	10
Perla	C3	CH3	R2	2	0	4	4	8	8	10	9	9	10	16
Perla	C4	CH3	R2	2	2	4	4	8	7	9	8	10	14	9
Perla	C4	CH3	R2	2	0	4	4	8	8	10	10	10	9	6

(continued)

Coffee data (continued)

Shade	Clone	Tray	Rep	y1	y2	y3	y4	y5	y6	y7	y8	y9	y10	y11
Perla	C4	CH3	R2	2	2	4	6	8	10	12	9	9	7	
Perla	C5	CH3	R2	2	0	4	6	8	10	4	7	3	6	11
Perla	C5	CH3	R2	2	0	4	6	8	10	12	5	4	7	16
Perla	C5	CH3	R2	2	0	4	6	8	8	10	5	4	3	8
Perla	pf	CH3	R2	2	0	4	6	8	8	8	6	5	7	21
Perla	pf	CH3	R2	2	2	4	6	8	10	10	12	12	13	22
Perla	pf	CH3	R2	2	0	4	6	8	7	5	12	4	14	18
Perla	C1	CH3	R3	2	2	4	6	8	6	4	5	2	6	12
Perla	C1	CH3	R3	2	0	4	6	8	7	2	2	4	9	14
Perla	C1	CH3	R3	2	2	4	6	8	8	8	6	8	5	4
Perla	C2	CH3	R3	2	0	2	2	2	3	3	3	3	5	5
Perla	C2	CH3	R3	2	0	2	2	4	6	8	4	4	4	6
Perla	C2	CH3	R3	2	0	2	2	4	6	8	7	8	5	7
Perla	C3	CH3	R3	2	2	4	6	10	8	10	8	8	8	9
Perla	C3	CH3	R3	2	2	4	6	8	5	9	3	1		
Perla	C3	CH3	R3	2	0	4	6	10	8	7	2	1	2	6
Perla	C4	CH3	R3	2	2	4	4	8	8	10	6	2	1	4
Perla	C4	CH3	R3	2	0	4	6	8	8	10	6			
Perla	C4	CH3	R3	2	0	4	4	10	7	9	6	5	8	13
Perla	C5	CH3	R3	2	0	2	4	8	6	8	7	9	10	19
Perla	C5	CH3	R3	2	0	4	8	10	10	12	11	11	9	12
Perla	C5	CH3	R3	2	2	4	6	8	8	7	10	4	11	16
Perla	pf	CH3	R3	2	4	6	6	8	8	9	7	6	14	18
Perla	pf	CH3	R3	2	2	4	6	8	10	8	5	5	12	26
Perla	pf	CH3	R3	2	2	4	6	8	9	8	10	10	11	15

Perla	C1	CH3	R4	2	2	4	6	7	5	5	1	1		
Perla	C1	CH3	R4	2	0	4	6	10	7	2	2	1	7	8
Perla	C1	CH3	R4	2	0	6	6	8	5	5	5	5	8	8
Perla	C2	CH3	R4	2	2	3	5	7	8	9	5	5	3	7
Perla	C2	CH3	R4	2	0	5	6	6	8	6	6	5	11	14
Perla	C2	CH3	R4	2	0	4	4	6	2	4	6	5	5	6
Perla	C3	CH3	R4	2	2	4	6	8	8	9	10			
Perla	C3	CH3	R4	2	0	4	6	8	7	5	4	4	4	11
Perla	C3	CH3	R4	2	0	4	6	6	8	10	1	7	5	8
Perla	C4	CH3	R4	2	2	4	4	6	4	2	1	1	3	7
Perla	C4	CH3	R4	2	2	4	6	8	4	4	1	1		
Perla	C4	CH3	R4	2	0	6	6	8	6	5	1			
Perla	C5	CH3	R4	2	2	6	4	9	5	3				
Perla	C5	CH3	R4	2	2	5	5	8	4	1	1	2	5	5
Perla	C5	CH3	R4	2	2	4	4	6	5	5	5	4	7	8
Perla	pf	CH3	R4	2	2	4	6	8	6	4	1	1	16	20
Perla	pf	CH3	R4	2	0	5	5	7	5	4	2	2	4	8
Perla	pf	CH3	R4	2	4	3	5	6	4	5	5	4	10	14
Negra	C1	CH1	R1	2	2	4	6	6	8					
Negra	C1	CH1	R1	2	2	6	6	6	8					
Negra	C1	CH1	R1	2	0	4	6	8	8					
Negra	C2	CH1	R1	2	0	2	4	6	8	4	5	7	6	6
Negra	C2	CH1	R1	0	0	2	2	2	4	4	6	5	4	4
Negra	C2	CH1	R1	0	0	2	2	2	4	9	8	4	7	7
Negra	C3	CH1	R1	2	2	4	4	6	8	7	3	2	6	10
Negra	C3	CH1	R1	2	2	4	5	6	8	5	4	4	3	9
Negra	C3	CH1	R1	2	0	2	3	4	6	3	3	4	9	11

(continued)

Coffee data (continued)

Shade	Clone	Tray	Rep	y1	y2	y3	y4	y5	y6	y7	y8	y9	y10	y11
Negra	C4	CHI	R1	2	0	4	6	6	4	4				
Negra	C4	CHI	R1	2	0	4	4	6	6					
Negra	C4	CHI	R1	2	2	4	6	8	8					
Negra	C5	CHI	R1	2	2	4	6	8	8	4			4	8
Negra	C5	CHI	R1	2	2	4	4	6	8					
Negra	C5	CHI	R1	0	0	2	4	6	8					
Negra	pf	CHI	R1	2	2	6	6	8	8	8	3	3	7	13
Negra	pf	CHI	R1	2	2	4	4	6	6	6	5	6	10	22
Negra	pf	CHI	R1	2	2	6	6	8	8	8	4	4	9	16
Negra	C1	CHI	R2	2	2	6	6	8	8	8	5		7	9
Negra	C1	CHI	R2	2	2	4	6	6	8	3				
Negra	C1	CHI	R2	2	0	4	6	6	6					
Negra	C2	CHI	R2	0	0	4	6	4	4	4				
Negra	C2	CHI	R2	2	2	4	4	6	8	4	3	2		
Negra	C2	CHI	R2	2	2	4	6	8	8					
Negra	C3	CHI	R2	2	2	6	6	4	4	6				
Negra	C3	CHI	R2	2	0	2	2	6	3	3				
Negra	C3	CHI	R2	2	2	6	6	6	8	6			1	2
Negra	C4	CHI	R2	2	2	4	4	6	4	4				
Negra	C4	CHI	R2	2	0	2	2	4	6	6				
Negra	C4	CHI	R2	2	0	4	2	2	2	2				
Negra	C5	CHI	R2	2	0	4	4	6	6	4	3		1	2
Negra	C5	CHI	R2	2	2	6	6	6	8	4				
Negra	C5	CHI	R2	2	2	6	6	8	7	6				
Negra	pf	CHI	R2	2	2	4	6	6	8	2	3	3	11	13

Negra	pf		CHI	R2	2	4	6	6	8	8	6	4	3	4	7	10
Negra	pf		CHI	R2	2	4	6	7	8	8	4	5	2	5	1	
Negra	C1		CHI	R3	2	2	6	6	8	8	10	12	9	9	9	9
Negra	C1		CHI	R3	2	2	6	6	8	8	10	7	5	6	4	15
Negra	C1		CHI	R3	2	0	4	6	8	8	10	6	6	4	7	14
Negra	C2		CHI	R3	2	0	2	6	8	8	4	3	1		4	8
Negra	C2		CHI	R3	2	0	2	2	2	2	2	4			1	
Negra	C2		CHI	R3	2	0	4	4	4	4	8		3			
Negra	C3		CHI	R3	2	2	6	6	8	8	6	3	2			
Negra	C3		CHI	R3	2	2	6	6	8	8	10	6	2		1	4
Negra	C3		CHI	R3	2	2	4	6	8	8	6	2		1	7	10
Negra	C4		CHI	R3	2	2	4	6	8	8	6	5	3			
Negra	C4		CHI	R3	2	2	6	6	8	8	10	5	2			
Negra	C4		CHI	R3	2	2	6	4	6	6	4	4	1			
Negra	C5		CHI	R3	2	2	6	6	8	8	6	6			13	20
Negra	C5		CHI	R3	2	2	6	6	8	8	10	2	4	5	1	2
Negra	C5		CHI	R3	2	2	6	6	8	8	9	8	4	1	7	10
Negra	pf		CHI	R3	2	2	6	6	8	8	10	8	6	6	10	18
Negra	pf		CHI	R3	2	4	6	6	8	8	8	6	6	6	13	17
Negra	pf		CHI	R3	2	4	6	8	8	8	8	4	2	1	18	18
Negra	C1		CHI	R4	2	2	6	6	8	8	6	5	4	3	3	10
Negra	C1		CHI	R4	2	0	4	6	8	8	4	5	3		3	9
Negra	C1		CHI	R4	2	2	6	6	8	8	4	1	4	1	7	12
Negra	C2		CHI	R4	2	2	4	5	8	8	8	6				
Negra	C2		CHI	R4	2	2	6	6	8	8	6	4	2	2	2	2
Negra	C2		CHI	R4	2	2	4	6	8	8	6	4	1			
Negra	C2		CHI	R4	2	2	4	6	8	8	6	4	1			
Negra	C3		CHI	R4	2	2	4	6	8	8	6	3	2	2	5	11

(continued)

Coffee data (continued)

Shade	Clone	Tray	Rep	y1	y2	y3	y4	y5	y6	y7	y8	y9	y10	y11
Negra	C3	CH1	R4	2	0	4	6	6	2	9	5	5	2	2
Negra	C3	CH1	R4	2	0	4	4	8	2	3	3	3	5	6
Negra	C4	CH1	R4	2	2	6	6	8	3	3	2	2	3	4
Negra	C4	CH1	R4	2	0	2	2	2	4	4	6	6	2	5
Negra	C4	CH1	R4	2	2	4	6	8	6	6			9	8
Negra	C5	CH1	R4	2	0	6	4	6	6	8	4	4	8	15
Negra	C5	CH1	R4	2	2	6	5	8	8	3	3	3	5	11
Negra	C5	CH1	R4	2	2	6	6	8	6	5	4	4	7	10
Negra	pf	CH1	R4	2	4	6	6	10	10	6	5	4	11	16
Negra	pf	CH1	R4	2	2	6	5	8	6	6	1	1	8	11
Negra	pf	CH1	R4	2	2	6	4	4	4	4	5	6	10	14
Negra	C1	CH2	R1	2	2	4	6	8	8	10	4	3	7	10
Negra	C1	CH2	R1	2	0	4	6	8	7	9	3	3	4	8
Negra	C1	CH2	R1	2	2	6	6	8	8	9	1	1	1	5
Negra	C2	CH2	R1	0	0	2	4	6	8	8	7	2	2	5
Negra	C2	CH2	R1	2	2	4	4	6	10	6	5	3	9	9
Negra	C2	CH2	R1	0	2	2	2	4	6	8	4	2	1	12
Negra	C3	CH2	R1	2	0	6	6	8	10	5				
Negra	C3	CH2	R1	2	0	4	6	8	8	6				
Negra	C3	CH2	R1	2	2	4	6	8	10	7	7	3	2	3
Negra	C4	CH2	R1	2	2	6	6	8	10	5	8	7	9	8
Negra	C4	CH2	R1	2	0	4	10	8	7	10		1	4	5
Negra	C5	CH2	R1	2	2	6	6	8	10	10	5	5	1	7
Negra	C5	CH2	R1	2	2	4	7	10	10	9				

Negra	C5	CH2	R1	2	2	4	8	10	10	7	6	6	2
Negra	pf	CH2	R1	2	0	4	6	8	6	5	4	4	10
Negra	pf	CH2	R1	2	3	6	6	8	6	4	3	1	11
Negra	pf	CH2	R1	2	2	6	6	8	6	6	3	2	12
Negra	C1	CH2	R2	2	0	4	6	8	7	8	8	3	8
Negra	C1	CH2	R2	2	2	6	6	8	9	8	2	2	2
Negra	C1	CH2	R2	2	2	4	6	6	8	6	2	7	14
Negra	C2	CH2	R2	2	2	6	6	8	5	8	8	6	
Negra	C2	CH2	R2	2	2	4	4	8	7	9	9	7	10
Negra	C2	CH2	R2	0	2	4	4	6	8	5			
Negra	C3	CH2	R2	2	2	4	6	8	6	3			
Negra	C3	CH2	R2	2	0	4	4	6	8	10	2	4	
Negra	C3	CH2	R2	2	2	4	6	8	8	10			
Negra	C4	CH2	R2	2	0	4	4	8	9	10	7	6	2
Negra	C4	CH2	R2	2	0	6	6	8	7	9	10	7	4
Negra	C4	CH2	R2	2	2	6	6	10	10	7	10	10	6
Negra	C5	CH2	R2	2	2	6	6	8	10	9	7	6	7
Negra	C5	CH2	R2	2	2	6	6	8	10	12	10	8	11
Negra	C5	CH2	R2	2	2	4	8	8	10				
Negra	pf	CH2	R2	2	2	4	6	8	10	11	8	9	20
Negra	pf	CH2	R2	2	0	4	6	8	10	3			
Negra	pf	CH2	R2	2	0	4	6	8	7	6			
Negra	C1	CH2	R3	2	0	6	6	8	8	7	7	3	12
Negra	C1	CH2	R3	2	2	6	6	10	4	10	7	5	14
Negra	C1	CH2	R3	2	2	4	6	8	8	10	5	2	12
Negra	C2	CH2	R3	2	0	4	6	8	8	2	5	2	
Negra	C2	CH2	R3	2	0	4	6	8	6	6	2		

(continued)

Coffee data (continued)

Shade	Clone	Tray	Rep	y1	y2	y3	y4	y5	y6	y7	y8	y9	y10	y11
Negra	C2	CH2	R3	2	2	4	6	6	5	7	4	2		
Negra	C3	CH2	R3	2	2	6	6	8	10	9	4	3	12	14
Negra	C3	CH2	R3	2	2	4	4	6	4	5	5	2	8	12
Negra	C3	CH2	R3	2	2	4	6	8	6	8	4	4	4	6
Negra	C4	CH2	R3	2	2	6	6	8	4	7	6	4	5	5
Negra	C4	CH2	R3	2	2	6	6	8	10	7				
Negra	C4	CH2	R3	2	2	4	6	6	6	7	2	1	2	7
Negra	C5	CH2	R3	2	2	4	6	8	10	11	3	5	9	11
Negra	C5	CH2	R3	2	2	4	4	8	8	8	7			
Negra	C5	CH2	R3	2	2	6	8	8	12	8	8	6	9	13
Negra	pf	CH2	R3	2	0	4	6	8	10	12	11	7	10	14
Negra	pf	CH2	R3	2	0	4	6	8	10	8	6	5	13	18
Negra	pf	CH2	R3	2	0	6	8	8	8	9	9	4	10	12
Negra	C1	CH2	R4	2	2	6	6	8	8					
Negra	C1	CH2	R4	2	2	6	6	8	8	7	2	2		
Negra	C1	CH2	R4	2	2	6	6	8	3					
Negra	C2	CH2	R4	2	0	2	4	6	5	5	4	4	4	7
Negra	C2	CH2	R4	2	0	4	4	4	2	1				
Negra	C2	CH2	R4	2	2	6	6	8	2	2				
Negra	C3	CH2	R4	2	2	6	6	8	8	2				
Negra	C3	CH2	R4	2	2	6	6	8	6	5				
Negra	C3	CH2	R4	2	0	4	4	6	2	8		5	3	4
Negra	C4	CH2	R4	2	2	6	6	6	8	9	3	1		
Negra	C4	CH2	R4	2	0	6	6	8	10	7	4	3	2	4
Negra	C4	CH2	R4	2	2	6	4	8	4	8	6	4	2	5

Negra	C5	CH2	R4	2	2	6	6	10	10	8	4	4	6	8
Negra	C5	CH2	R4	2	0	4	8	10	6	6	3	4	2	8
Negra	C5	CH2	R4	2	2	6	6	8	8	5	10	5	5	8
Negra	pf	CH2	R4	2	0	4	6	8	8	3	1			
Negra	pf	CH2	R4	2	2	4	6	8	8					
Negra	pf	CH2	R4	2	2	4	6	8	10	5	2			
Negra	C1	CH3	R1	2	0	6	6	8	9	7	2	1		
Negra	C1	CH3	R1	2	2	6	6	8	6	5	5	5	6	6
Negra	C1	CH3	R1	2	2	4	4	4	4	6	6	6	4	4
Negra	C2	CH3	R1	2	2	6	4	6	6	10	10	9	10	10
Negra	C2	CH3	R1	2	2	6	6	8	8	12	12	9	12	15
Negra	C2	CH3	R1	0	0	2	4	2	6	6	8	6	1	5
Negra	C3	CH3	R1	2	2	6	6	8	10	4	3	3	2	7
Negra	C3	CH3	R1	2	2	4	6	6	8	8	7	9	7	16
Negra	C3	CH3	R1	2	0	4	6	8	10	8	10	10	11	6
Negra	C4	CH3	R1	2	4	4	6	8	10	9	8	6	7	9
Negra	C4	CH3	R1	2	2	4	4	6	8	6	6			
Negra	C4	CH3	R1	2	2	4	6	7	6	9	8	9	11	11
Negra	C5	CH3	R1	2	2	6	6	8	10	9	8	7	4	3
Negra	C5	CH3	R1	2	2	6	6	8	9	9	8	7	9	9
Negra	C5	CH3	R1	2	2	6	8	8	5	6	8	7	10	12
Negra	pf	CH3	R1	2	2	6	8	6	8	9	7	9	3	9
Negra	pf	CH3	R1	2	2	6	6	8	10	7	7	10	10	10
Negra	pf	CH3	R1	2	0	4	6	8	10	8	5	3	1	7
Negra	C1	CH3	R2	2	0	4	6	8	10	8	2			
Negra	C1	CH3	R2	2	0	4	6	8	10	8	2			
Negra	C1	CH3	R2	2	0	4	6	8	6	6	4	4	1	2
Negra	C1	CH3	R2	2	2	6	6	10	6	7	5	3	2	6

(continued)

Coffee data (continued)

Shade	Clone	Tray	Rep	y1	y2	y3	y4	y5	y6	y7	y8	y9	y10	y11
Negra	C2	CH3	R2	2	0	2	2	4	6	6	2	2		
Negra	C2	CH3	R2	2	2	6	6	6	6	7	4	4	2	2
Negra	C2	CH3	R2	2	0	4	4	6	7	6	5	3	1	2
Negra	C3	CH3	R2	2	2	6	6	8	10	10	12	10	9	11
Negra	C3	CH3	R2	2	2	6	6	8	6	7	6	5	10	8
Negra	C3	CH3	R2	2	2	6	6	8	7	6	7	5	4	5
Negra	C4	CH3	R2	2	0	6	6	8	7	7	3	2	1	2
Negra	C4	CH3	R2	2	2	4	6	6	4	4	6	3		
Negra	C4	CH3	R2	2	2	6	6	6	5	7	5	6	4	2
Negra	C5	CH3	R2	2	2	6	6	10	7	6	3	6	4	6
Negra	C5	CH3	R2	2	2	6	6	8	7	8	5	4	2	4
Negra	C5	CH3	R2	2	2	6	6	8	8	9	4	6	7	7
Negra	pf	CH3	R2	2	2	4	6	8	5	6	4	5	4	7
Negra	pf	CH3	R2	2	0	4	6	8	2	4	6	6	5	6
Negra	pf	CH3	R2	2	0	4	6	6	7	6	4	4	3	7
Negra	C1	CH3	R3	2	2	6	6	8	6	2				
Negra	C1	CH3	R3	2	0	4	6	8	8	4				
Negra	C1	CH3	R3	2	2	4	6	8	4	5	4	4	5	4
Negra	C2	CH3	R3	2	0	4	6	8	9	4	11	10	6	8
Negra	C2	CH3	R3	2	0	4	6	8	8	5	2	2	3	14
Negra	C2	CH3	R3	2	2	6	6	10	8	9	6	7	6	2
Negra	C3	CH3	R3	2	0	4	6	7	8	4				
Negra	C3	CH3	R3	2	0	4	6	7	8	8				
Negra	C3	CH3	R3	2	0	4	6	8	5	3				
Negra	C4	CH3	R3	2	0	4	4	8	8	9	7	6	4	7

Negra	C4	CH3	R3	2	2	6	6	8	7	8									
Negra	C4	CH3	R3	2	0	2	6	8	8	8	8	10	7						9
Negra	C5	CH3	R3	2	2	4	4	8	3	2									
Negra	C5	CH3	R3	2	2	4	6	8	6	5									
Negra	C5	CH3	R3	2	2	6	6	6	3	4	1								
Negra	pf	CH3	R3	2	2	4	6	8	5	5	2	2	6						10
Negra	pf	CH3	R3	2	2	4	8	8	8	4	5	7	7						7
Negra	pf	CH3	R3	2	2	4	8	8	5	3	4	2	6						4
Negra	C1	CH3	R4	2	2	4	4	8	8	9	8	8	8						9
Negra	C1	CH3	R4	2	0	4	6	8	7	9	1		2						2
Negra	C1	CH3	R4	2	0	4	6	8	8	1									
Negra	C2	CH3	R4	2	0	4	4	6	5	6	6								
Negra	C2	CH3	R4	2	2	2	6	6	7	7									
Negra	C2	CH3	R4	2	0	2	2	6	5	7		2	1						1
Negra	C3	CH3	R4	2	0	4	4	6	6	6	10								
Negra	C3	CH3	R4	2	2	6	6	10	10	10									
Negra	C3	CH3	R4	2	0	4	4	6	7	7									
Negra	C4	CH3	R4	2	0	6	6	8	6	8	4	2	1						2
Negra	C4	CH3	R4	2	0	4	4	6	8	10	7								
Negra	C4	CH3	R4	2	0	4	4	8	7	6	2	2	3						2
Negra	C5	CH3	R4	2	2	6	8	8	6	4	3	2	2						2
Negra	C5	CH3	R4	2	0	4	6	8	8	8	6	6	3						5
Negra	C5	CH3	R4	2	2	4	6	8	8	8	7	8	5						8
Negra	pf	CH3	R4	2	2	4	6	8	10	11	4	4	4						6
Negra	pf	CH3	R4	2	4	4	8	8	8	3	1								
Negra	pf	CH3	R4	2	4	6	8	10	10	5	2	2	4						6

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



Chapter 6

Generalized Linear Mixed Models for Proportions and Percentages



6.1 Response Variables as Ratios and Percentages

In this chapter, we will review generalized linear mixed models (GLMMs) whose response can be either a proportion or a percentage. For proportion and percentage data, we refer to data whose expected value is between 0 and 1 or between 0 and 100. For the remainder of this book, we will refer to this type of data only in terms of proportion, knowing that it is possible to change it to a percentage scale only when multiplying it by 100. Proportions can be classified into two types: discrete and continuous. Discrete proportions arise when the unit of observation consists of N distinct entities, of which individuals have the attribute of interest “ y ”. N must be a nonnegative integer and “ y ” must be a positive integer; here, $y \leq N$. Therefore, the observed proportion must be a discrete fraction, which can take values $\frac{0}{N}, \frac{1}{N}, \dots, \frac{N}{N}$. A binomial distribution is the sum of a series of m independent binary trials (i.e., trials with only two possible outcomes: success or failure), where all trials have the same probability of success. For binary and binomial distributions, the target of inference is the value of the parameter such that $0 \leq E\left(\frac{y}{N}\right) = \pi \leq 1$. Continuous proportions (ratios) arise when the researcher measures responses such as the fraction of the area of a leaf infested with a fungus, the proportion of damaged cloth in a square meter, the fraction of a contaminated area, and so on. As with the binomial parameter π , the continuous rates (fractions) take values between 0 and 1, but, unlike the binomial, the continuous proportions do not result from a set of Bernoulli tests. Instead, the beta distribution is most often used when the response variable is in continuous proportions. In the following sections, we will first address issues in modeling when we have binary and binomial data. When the response variable is binomial, we have the option of using a linearization method (pseudo-likelihood (PL)) or the Laplace or quadrature integral approximation (Stroup 2012).

6.2 Analysis of Discrete Proportions: Binary and Binomial Responses

A binomial distribution is the number of successes from a series of N independent binary trials – Bernoulli trials (i.e., trials with two possible outcomes: success or failure), where all trials have the same probability of success. In the context of a GLMM, there are N binomial responses, each of which is the result of binary trials. The i th response consists of two pieces of information: the number of trials n_i and the number of successes y_i , as shown in the following example.

6.2.1 Completely Randomized Design (CRD): Methylation Experiment

An agent to induce demethylation is applied to plants; this agent converts methylated nucleotides to their unmethylated forms, thus causing epigenetic changes that produce or induce abnormal phenotypes such as deformation or stunting (Amoah et al. 2008). A pilot study was implemented to investigate the relationship between the dose of the demethylating agent and the observed proportion of plants with a normal phenotype. Seeds were treated with the demethylating agent at six different doses, including the control. Plants were sown in trays, with each tray containing seeds previously treated with the same dose of the demethylating agent. Each dose was replicated 4 times: 2 with 60 plants and 2 with 100 plants. The trays were allocated following a completely randomized design (CRD). The plants with a normal phenotype in each tray are shown (in Table 6.1) with the number of plants per tray (N). The notation 59(60) indicates that 59 normal plants were found out of 60 plants under study. In the same way, the notation 14(100) indicates that 14 normal plants were found out of 100 plants under study.

The sources of variation and degrees of freedom (DFs) for this experiment are shown in Table 6.2.

Table 6.1 Number of normal plants out of a total of N plants per tray and dose of the demethylating agent

Dose					
0	0.01	0.1	0.5	1.0	1.5
59(60)	58(60)	54(60)	4(60)	3(60)	3(60)
58(60)	59(60)	53(60)	11(60)	2(60)	3(60)
99(100)	98(100)	88(100)	14(100)	2(100)	1(100)
98(100)	99(100)	87(100)	15(100)	1(100)	3(100)

Table 6.2 Sources of variation and degrees of freedom

Sources of variation	Degrees of freedom
Dose	$t - 1 = 6 - 1 = 5$
Error	$t(r - 1) = 6 \times (4 - 1) = 18$
Total	$t \times r - 1 = 6 \times 4 - 1 = 23$

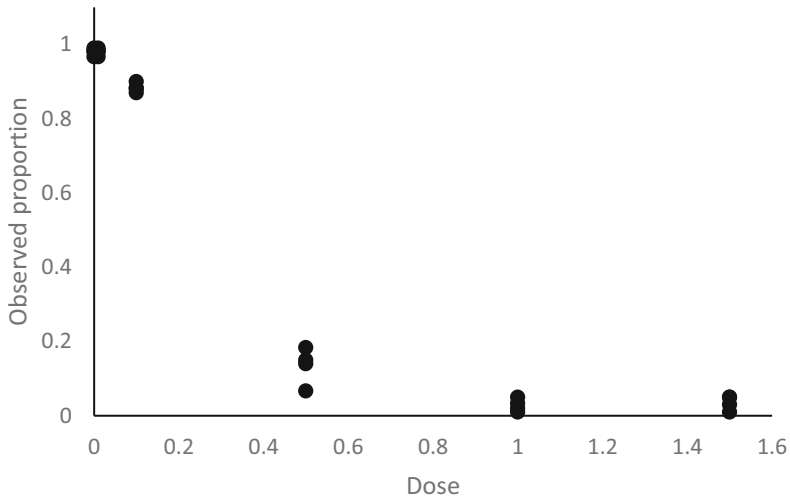


Fig. 6.1 Effect of the demethylating agent on the proportion of normal plants

The statistical model of a completely randomized design (CRD) is

$$y_{ij} = \mu + \tau_i + \varepsilon_{ij}$$

where y_{ij} is the number of observed normal plants in the tray j ($j = 1, 2, 3, 4$) at the dose i ($i = 1, 2, \dots, 6$), μ is the overall mean, τ_i is the effect of dose i of the demethylating agent, and ε_{ij} are non-normal errors.

The expected value (normal plants) of a set of tests n_i follows a binomial distribution $y_i \sim \text{Binomial}(n_i, \pi_i)$, where π_i is the probability of success in each trial, with $0 \leq \pi_i \leq 1$, where $\pi_i = y_i/n_i$. Thus, the probability of observing an outcome y_i can be written as

$$P(Y_i = y_i | n_i, \pi_i) = \binom{n_i}{y_i} \pi_i^{y_i} (1 - \pi_i)^{n_i - y_i}, y_i = 0, 1, \dots, n_i.$$

This probability depends on the number of known tests n_i , whereas the probability of success (π_i) is an unknown parameter. In Fig. 6.1, we observe that the probability of obtaining a normal plant depends on the applied dose of the demethylating agent. Given that y_i has a binomial distribution, the expected value (the mean) is the product of the number of trials and the probability of success in each trial, that is, $E(Y_i) = n_i \pi_i$. Since the number of trials is fixed (once the data have been obtained), modeling the probability of success is equivalent to modeling the expected value as well as the variance since it is also a function of the number of trials and the probability of success. So, the expected value and variance of y_i are

$$E(y_i) = \mu_i = n_i\pi_i; \text{Var}(y_i) = n_i\pi_i(1 - \pi_i).$$

This variance is small if the value π_i is close to 0 or 1, and this increases to its maximum when $\pi_i = 0.5$. This can be seen in Fig. 6.1, where proportions close to 0 or 1 show less variance than do proportions between 0.1 and 0.2 for a demethylating agent dose of 0.5. This variance can also be written in terms of the expected value as:

$$\text{Var}(y_i) = \frac{\mu_i}{n_i} (n_i - \mu_i).$$

In this CRD, the fixed number of treatments t (doses) were randomly assigned to r experimental units (trays). The linear predictor describing the structure of the mean of this GLMM is

$$\eta_i = \eta + \tau_i$$

where η_i denotes the i th linear predictor, η is the intercept, and τ_i is the fixed effect due to treatments i ($i = 1, 2, \dots, t$) with t treatments and r_i replicates in each treatment.

The components that define this GLMM are shown below:

Distribution: $y_i \sim \text{Binomial}(N_{ij}, \pi_i)$

Linear predictor: $\eta_i = \eta + \tau_i$

Link function: $\text{logit}(\pi_i) = \text{logit}\left(\frac{\pi_i}{1 - \pi_i}\right) = \eta_i$

where η_i is the linear predictor that relates the effect of dose i ($i = 1, 2, \dots, 6$) to probability π_i . The model uses the linear predictor (η_i) to estimate the means ($\pi_i = \mu_i$) of the observations for each treatment.

The following GLIMMIX program fits a CRD with a binomial response:

```
proc glimmix nobound method=Laplace;
class Dose Rep;
model y/N= dose/link=logit;
lsmeans dose/lines ilink;
run;
```

In this example, the distribution of the dataset was not specified to GLIMMIX in the model specification because by using the expression "Y/N," proc GLIMMIX automatically infers that this dataset has a binomial distribution. It is also important to note that variable dose and repetition were declared as class variables in the "class" command, which Statistical Analysis Software (SAS) interprets as explanatory variables that are nonnumerical factors. However, the variable declared "Rep" is not used in the model specification.

Table 6.3 Results of the analysis of variance

(a) Fit statistics for conditional distribution							
-2 Log L (y r. effects)							83.46
Pearson's chi-square							11.95
Pearson's chi-square/DF							0.50
(b) Type III tests of fixed effects							
Effect	Num DF	Den DF	F-value	Pr > F			
Dose	5	15	132.53	<0.0001			
(c) Dose least squares (LS) means							
Dose	Estimate	Standard error	DF	t-value	Pr > t	Mean	Standard error mean
0	3.9580	0.4122	15	9.60	<0.0001	0.9813	0.007581
0.01	3.9580	0.4122	15	9.60	<0.0001	0.9813	0.007581
0.1	2.0049	0.1728	15	11.60	<0.0001	0.8813	0.01808
0.5	-1.8360	0.1623	15	-11.31	<0.0001	0.1375	0.01925
1	-3.6633	0.3580	15	-10.23	<0.0001	0.02501	0.008729
1.5	-3.4337	0.3212	15	-10.69	<0.0001	0.03126	0.009728

Part of the results is shown in Table 6.3. Pearson's chi-squared statistic value divided by the degrees of freedom in part (a) (Pearson's chi - square/DF = 0.5) indicates that there is no evidence of extra-dispersion in the dataset. The analysis of variance (ANOVA) tabulated in part (b) in Table 6.3, with the type III tests of fixed effects, indicates that there is a highly significant difference ($P = 0.0001$) in the average proportion of normal plants with respect to the dose applied to the seeds.

The output when using the "lsmeans" command in conjunction with the "link" option is in the "Mean" column (part (c) in Table 6.3). These values are the values of π_i 's, i.e., the estimated probabilities $\hat{\pi}_0 = 0.9813$ and $\hat{\pi}_{0.01} = 0.9813$ of normal plants for the treatments whose doses are 0 and 0.01, respectively. For treatments with doses of 0.1 and 0.5, the observed probabilities of normal plants are $\hat{\pi}_{0.1} = 0.8813$ and $\hat{\pi}_{0.5} = 0.1375$, respectively, whereas for the 1 and 1.5 doses, the observed probabilities of normal plants decrease dramatically with $\hat{\pi}_1 = 0.02501$ and $\hat{\pi}_{1.5} = 0.03126$, respectively.

Figure 6.2 shows the mean comparisons (least significance difference (LSD)) of the estimated probabilities according to the dose applied to the seeds in trays. In this figure, we can observe that in the treatments with dose = 0 (control) and dose = 0.01, the observed proportions of normal plants are not statistically different from each other, but they do differ with the other applied doses. At a dose of 0.1, the observed proportion of normal plants was 88.13%, and this was statistically different from all the doses used. Finally, doses at 0.5, 1, and 1.5 of the demethylating agent in the observed proportion of normal plants decreased drastically to 13.75%, 2.501%, and 3.12%, respectively. The doses of 1 and 1.5 produced statistically equal proportions of normal plants.

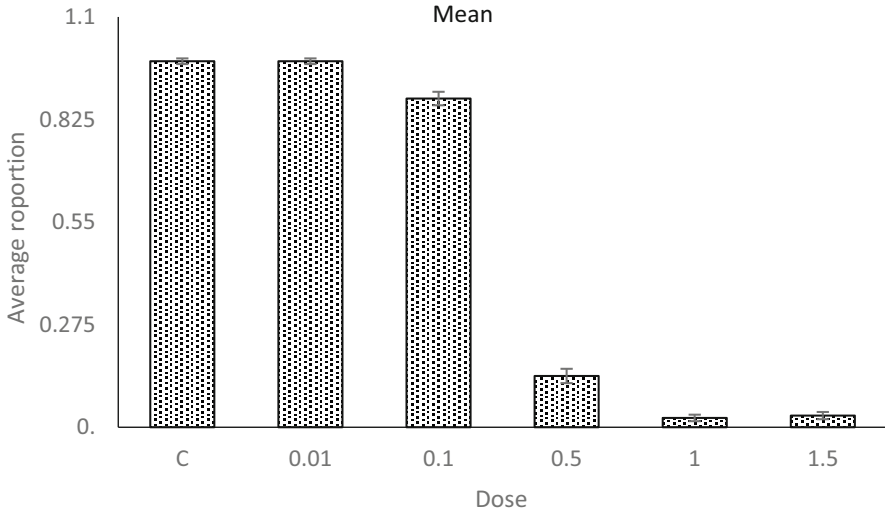


Fig. 6.2 Comparison of the estimated probabilities per dose of the demethylating agent

If the researcher wishes to model how dose levels of the demethylating agent affect normal plant proportions, then the dose must be declared as a continuous variable. The following SAS syntax with `proc GLIMMIX` runs a binomial regression:

```
proc glimmix data=crd_bin method=Laplace plots=all;
class rep;
model y/N= dose/solution;
random rep;
run;quit.
```

Most of the commands and options have already been discussed throughout this book; the “`model y/N`” command indicates that the response variable is in a ratio. Therefore, this dataset is modeled with a binomial distribution, which is affected by the different number of individuals in each repetition. `proc GLIMMIX` interprets the distribution of the data as binomial, whereas the “`solution`” option requests the parameter estimates of the model (intercept and slope).

The components that define this GLMM are shown below:

Distribution: $y_i \sim \text{Binomial}(N_{ij}, \pi_i)$

Linear predictor: $\eta_i = \eta + \beta * \text{dose}_i$

Link function: $\text{logit}(\pi_i) = \text{logit}\left(\frac{\pi_i}{1-\pi_i}\right) = \eta_i$

Thus, the model can be written as

Table 6.4 Regression analysis results

(a) Fit statistics					
-2 Log likelihood					231.58
Akaike information criterion (AIC) (smaller is better)					235.58
AICC (smaller is better)					236.15
Bayesian information criterion (BIC) (smaller is better)					237.93
CAIC (smaller is better)					239.93
HQIC (smaller is better)					236.20
Pearson's chi-square					2317.12
Pearson's chi-square/DF					96.55
(b) Type III tests of fixed effects					
Effect	Num DF	Den DF	F-value	Pr > F	
Dose	1	19	475.97	<0.0001	
(c) Solutions for fixed effects					
Effect	Estimate	Standard error	DF	t-value	Pr > t
Intercept	2.7927	0.1302	3	21.46	0.0002
Dose	-7.6232	0.3494	19	-21.82	<0.0001

$$\eta_i = \log \left(\frac{\mu_i}{n_i - \mu_i} \right) = \log \left(\frac{n_i \pi_i}{n_i - n_i \pi_i} \right) = \log \left(\frac{\pi_i}{1 - \pi_i} \right) = \text{logit}(\pi_i) = \eta + \beta \text{dose}_i$$

and the logit function can be written in terms of the probability of success, π_i , as

$$\pi_i = \frac{1}{1 + \exp(-\eta_i)}$$

Part of the SAS output of the GLIMMIX syntax is shown below. The goodness-of-fit statistics, type III tests of fixed effects, and parameter estimates are shown in Table 6.4. The analysis of variance indicates that the demethylating agent has a highly significant effect on the observed proportion of normal plants ($P < 0.0001$) (part (b)). The maximum likelihood estimates for the intercept and slope are $\eta = 2.7927$ and $\beta = -7.6232$, respectively.

Figure 6.3 shows that as the value of the linear predictor increases (η_i), the value of the residuals rapidly decreases. We can also see that the residuals plotted against the quantiles clearly do not follow a normal distribution because this model is not a linear function of the explanatory variable “dose.”

Figure 6.4 shows that the proportions studied and fitted are not so far apart, and, as such, the binomial model is suitable for this dataset. The estimated linear predictor of this model is as follows:

$$\hat{\eta}_i = \hat{\eta} + \hat{\beta} \times \text{dose}_i = 2.7927 - 7.6232 \times \text{dose}_i.$$

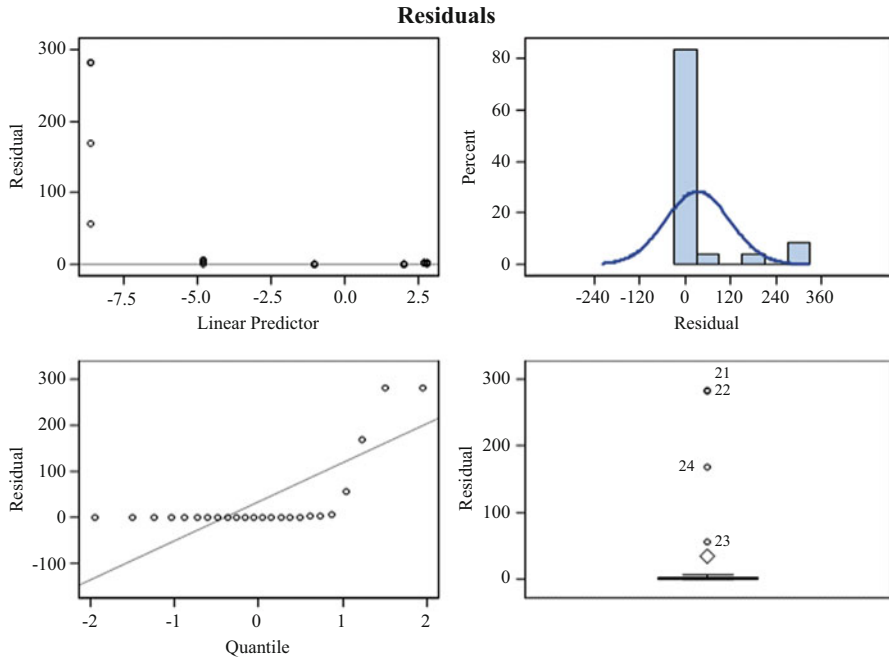


Fig. 6.3 A graph of residuals versus the linear predictor, quantiles

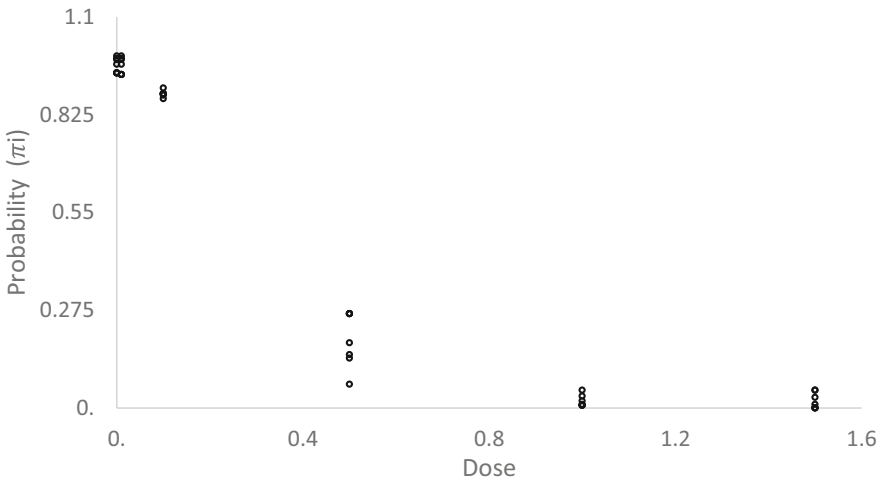


Fig. 6.4 Observed and estimated proportion

The logit of the probability of success is a linear function of the explanatory variables, so the model can be written in terms of the probability of success (observing normal plants) as

$$\pi_i = \frac{1}{1 + \exp(-\eta_i)}$$

Given the parameter estimates, we can predict the success probability of observing a normal plant, and given a certain concentration of the demethylating agent, this estimated probability (using the estimated linear predictor) can be seen plotted in Fig. 6.4.

$$\hat{\pi}_i = \frac{1}{1 + \exp(\hat{\eta}_i)} = \frac{1}{1 + \exp(-2.7927 + 7.6232 \times \text{dose}_i)}$$

6.3 Factorial Design in a Randomized Complete Block Design (RCBD) with Binomial Data: Toxic Effect of Different Treatments on Two Species of Fleas

A group of researchers wishes to study the toxic effect of certain treatments (Trts) on two flea species (SP) (*Daphnia magna* and *Ceriodaphnia dubia*). To compare the toxicity effect of treatments on both flea species, a randomized complete block design (RCBD bioassay) was implemented with three replicates per treatment, with each replicate consisting of 10 fleas (Appendix: Fleas). The linear predictor describing this experiment is described below:

$$\eta_{ijkl} = \eta + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \text{bioassay}_k + \text{rep}(\text{bioassay})_{l(k)}$$

where η is the intercept, α_i is the fixed effect due to species i , β_j is the fixed effect of treatment j , $(\alpha\beta)_{ij}$ is the fixed effects interaction between the flea species and treatment, bioassay_k is the random effect due to bioassay k assuming $\text{bioassay}_k \sim N(0, \sigma_{\text{bioassay}}^2)$, and $\text{rep}(\text{bioassay})_{l(k)}$ is the random effect due to repetition bioassay assuming $\text{rep}(\text{bioassay})_{l(k)} \sim N(0, \sigma_{\text{rep}(\text{bioassay})}^2)$.

The remaining components of this GLMM with a binomial response (N_{ijk}, π_{ijk}) are described below:

Distribution: $y_{ijkl} \mid \text{bioassay}_k, \text{rep}(\text{bioassay})_{l(k)} \sim \text{Binomial}(N_{ijk}, \pi_{ijk})$

$\text{bioassay}_k \sim N(0, \sigma_{\text{bioassay}}^2)$, $\text{rep}(\text{bioassay})_{l(k)} \sim N(0, \sigma_{\text{rep}(\text{bioassay})}^2)$, where N_{ijk} is the number of dead fleas, observed in species i in replicate l in bioassay k under treatment j ,

Link function: $\text{logit}(\pi_{ijk}) = \log\left[\frac{\pi_{ijk}}{1 - \pi_{ijk}}\right] = \eta_{ijk}$.

The following SAS syntax allows us to fit the GLMM with a binomial response.

Table 6.5 Results of the analysis of variance

(a) Fit statistics				
–2 Log likelihood				145.33
AIC (smaller is better)				173.33
AICC (smaller is better)				177.85
BIC (smaller is better)				160.71
CAIC (smaller is better)				174.71
HQIC (smaller is better)				147.97
(b) Fit statistics for conditional distribution				
–2 Log L (Sobrevi r. effects)				145.33
Pearson's chi-square				10.72
Pearson's chi-square/DF				0.10
(c) Covariance parameter estimates				
Cov Parm	Estimate	Standard error		
Bioen	–0.1051	.		
Bioen*SP (Rep)	–0.1192	.		
(d) Type III tests of fixed effects				
Effect	Num DF	Den DF	F-value	Pr > F
SP	1	14	0.02	0.8829
Trt	5	80	15.08	<0.0001
SP*trt	5	80	4.66	0.0009

```
proc glimmix data=pulgass nobound method=laplace;
class Bioen SP Trt Rep ;
Model Sobrevi/n = SP|Trat/dist=binomial;
random Bioen sp*bioen(rep) ;
lsmeans SP|Trt/lines ilink;
run;
```

Part of the results is listed in Table 6.5. The fit statistics in part (a) and the conditional statistics in part (b) are useful for model comparison, whereas the variance component estimates are shown in part (c). The value of the statistic Pearson's chi – square/DF = 0.10 indicates that the binomial model gives a good fit to the dataset. The variance component estimates for bioassays and replication nested in bioassays are $\hat{\sigma}_{\text{bioassay}}^2 = -0.1051$ and $\hat{\sigma}_{\text{rep}(\text{bioassay})}^2 = -0.1192$, respectively. The type III tests of fixed effects (part (d)) show the significance tests of the fixed effects in the model. The treatment effect and the interaction between the flea species (SP) and treatment are clearly significant with $P < 0.0001$ and $P = 0.0009$, respectively.

Since survival was statistically similar in both flea species, we will focus on the factors that were significant. Part (a) in Table 6.6 shows the means and standard errors of treatments on the model scale (“Estimate” column) and on the data scale (“Mean” column), obtained with “lsmeans” and the “ilink” option as well as the mean comparisons, which are on the model scale (part (b)).

Table 6.6 Means and standard errors on the model scale and on the data scale

(a) Trt least squares means							
Trt	Estimate	Standard error	DF	t-value	Pr > t	Mean	Standard error mean
T1	8.1179	4.3180	80	1.88	0.0637	0.9997	0.001287
T2	4.3564	3.0554	80	1.43	0.1578	0.9873	0.03820
T3	1.0081	0.1924	80	5.24	<0.0001	0.7326	0.03768
T4	-1.0509	0.1712	80	-6.14	<0.0001	0.2591	0.03286
T5	-4.7187	3.0570	80	-1.54	0.1266	0.008848	0.02681
T6	-8.1182	4.3184	80	-1.88	0.0638	0.000298	0.001286

(b) Conservative T grouping of Trt least squares means ($\alpha=0.05$)

LS means with the same letter are not significantly different

Trt	Estimate			
T1	8.1179		A	
T2	4.3564	B	A	
T3	1.0081	B	A	C
T4	-1.0509	B	D	C
T5	-4.7187		D	C
T6	-8.1182		D	

The LINES display does not reflect all significant comparisons. The following additional pairs are significantly different: (T3,T4)

Based on the fixed effects tests, the flea species \times treatment interaction is significant. The means on the model scale are listed under the “Estimate” column, followed by their standard errors, “Standard error” (Table 6.7). The output of the “ilink” option in “lsmeans” applies the inverse function of the link function to the estimates on the model scale to obtain the estimates on the data scale. The probabilities, on the data scale, are given under the “Mean” column with their respective standard errors and correspond to the probability of insect (flea) survival.

Figure 6.5 shows that the survival of both species is different in treatments 2–5; the *Daphnia* species showed more resistance in treatments 2 and 3, whereas the *Ceriodaphnia* species showed greater resistance in treatments 4 and 5. On the other hand, in treatments 1 and 6, survival was similar in both species.

6.4 A Split-Plot Design in an RCBD with a Normal Response

A split plot is the most common treatment structure design in agricultural and agro-industrial research areas. These experiments generally involve two or more factors under study. Typically, large or primary experimental units, commonly known as the whole plot, are grouped into blocks. The levels of the first factor are randomly assigned to the whole plots. Then, each whole plot is divided into smaller units, known as split or secondary plots. The levels of the second factor are randomly assigned to the subplots within each whole plot.

Table 6.7 Means and standard errors on the model scale and on the data scale of the interaction between both factors

SP*treatment least squares means		Estimate	Standard error	DF	t-value	Pr > t	Mean	Standard error mean
SP	Treatment							
	T1	8.1179	6.1065	80	1.33	0.1875	0.9997	0.001820
<i>Daphnia</i>	T2	8.1180	6.1068	80	1.33	0.1875	0.9997	0.001820
<i>Daphnia</i>	T3	1.9717	0.3218	80	6.13	<0.0001	0.8778	0.03452
<i>Daphnia</i>	T4	-1.2537	0.2536	80	-4.94	<0.0001	0.2221	0.04381
<i>Daphnia</i>	T5	-8.1186	6.1085	80	-1.33	0.1876	0.0002	0.001819
<i>Daphnia</i>	T6	-8.1182	6.1073	80	-1.33	0.1875	0.0002	0.001819
<i>Ceriodaphnia</i>	T1	8.1178	6.1064	80	1.33	0.1875	0.9997	0.001820
<i>Ceriodaphnia</i>	T2	0.5947	0.2202	80	2.70	0.0084	0.6444	0.05046
<i>Ceriodaphnia</i>	T3	0.04446	0.2109	80	0.21	0.8336	0.5111	0.05269
<i>Ceriodaphnia</i>	T4	-0.8480	0.2301	80	-3.69	0.0004	0.2999	0.04830
<i>Ceriodaphnia</i>	T5	-1.3188	0.2583	80	-5.11	<0.0001	0.2110	0.04301
<i>Ceriodaphnia</i>	T6	-8.1182	6.1071	80	-1.33	0.1875	0.0002	0.001819

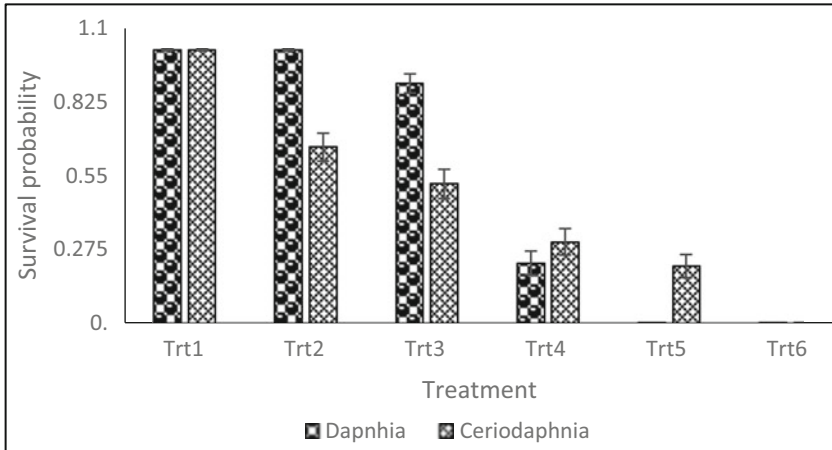


Fig. 6.5 The average survival rate of both species

The model equation for the analysis of variance assuming normality in the response is

$$y_{ijk} = \eta + \alpha_i + r_k + (ra)_{ik} + \beta_j + (\alpha\beta)_{ij} + e_{ijk}$$

$$i = 1, 2, \dots, a; j = 1, 2, \dots, b; k = 1, 2, \dots, r$$

where y_{ijk} is the observed response variable in the k th block at the i th level of factor A and at the j th level of factor B, α and β refer to the fixed treatment effects due to factors A and B, respectively, r is the random effect due to the blocks, $(ra)_{ik}$ is the random error term due to the whole plot that is an interaction between the blocks and factor A, and e_{ijk} is the random residual effect. Normally, the errors and other random terms are also assumed to be normal; however, when the response variable is not normally distributed, this way of specifying the model is not the most appropriate. Thus, under the assumption that the response variable is normal, this way of specifying the model is valid.

6.4.1 An RCBD Split Plot with Binomial Data: Carrot Fly Larval Infestation of Carrots

Data were obtained from an experiment that was designed to compare a number of carrot genotypes with respect to their resistance to infestation by carrot fly larvae. The data involved 16 genotypes that were compared at 2 pest levels to be controlled. The experiment was conducted in three randomized blocks. Each block consisted of

Table 6.8 The notation 44/53 denotes that 44 carrots were infected (*y*) out of a sample size of 53 studied (*N*)

	Treatment (level of infestation)					
	1			2		
Genotype	Block1	Block2	Block3	Block1	Block2	Block3
G1	44/53	42/48	27/51	16/60	9/52	26/54
G2	24/48	35/42	45/52	13/44	20/48	16/53
G3	8/49	16/49	16/50	4/52	6/51	12/43
G4	4/51	5/42	12/46	15/52	10/56	6/48
G5	11/52	13/51	15/44	4/51	6/43	9/46
G6	15/50	5/49	7/50	1/51	8/49	3/54
G7	18/52	13/47	7/47	2/52	4/52	6/52
G8	5/47	15/49	8/50	6/56	4/50	6/42
G9	11/52	6/45	5/51	3/54	8/51	3/53
G10	0/51	10/39	14/48	3/50	0/50	10/51
G11	6/52	4/46	10/37	1/52	7/38	4/48
G12	0/52	4/55	1/40	1/50	3/50	1/45
G13	14/45	18/43	4/40	4/51	7/46	7/45
G14	3/52	12/53	4/55	3/52	7/48	12/49
G15	11/52	6/54	5/49	2/50	4/46	14/53
G16	4/53	1/40	4/52	4/56	1/44	3/42

Table 6.9 Sources of variation and degrees of freedom

Sources of variation	Degrees of freedom
Blocks	$r - 1 = 3 - 1 = 2$
Factor A (infestation)	$a - 1 = 2 - 1 = 1$
Error _a ($A \times \text{blocks}$)	$(r - 1)(a - 1) = 2$
Factor B (genotypes)	$b - 1 = 16 - 1 = 15$
Infestation*genotype ($A \times B$)	$(a - 1)(b - 1) = 15$
Error _b	$a(r - 1)(b - 1) = 2 \times 2 \times 15 = 60$
Total	$r \times a \times b - 1 = 3 \times 2 \times 16 - 1 = 95$

32 plots, 1 for each combination of genotype and pest infestation level. At the end of the experiment, about 50 carrots were taken from each plot and assessed for infestation by carrot fly larvae. The data obtained are shown in Table 6.8.

Table 6.9 shows the analysis of variance summarizing the sources of variation and degrees of freedom.

Rewriting in terms of the linear predictor

$$\eta_{ijk} = \eta + \alpha_i + r_k + (ra)_{ik} + \beta_j + (\alpha\beta)_{ij}$$

Since the observations were taken at the subplot level, conditioned on the structural effects of the design, these observations have a variance associated with the subplot. Therefore, α and β refer to the treatment fixed effects due to factors A

Table 6.10 Results of the analysis of variance

(a) Fit statistics for conditional distribution				
-2 Log L (y r. effects)				527.82
Pearson's chi-square				189.09
Pearson's chi-square/DF				1.97
(b) Covariance parameter estimates				
Cov Parm	Subject	Estimate	Standard error	
Intercept	Bloque	0.004272	0.02741	
Trt	Bloque	0.03344	0.03545	
(c) Type III tests of fixed effects				
Effect	Num DF	Den DF	F-value	Pr > F
Genotype	15	60	28.28	<0.0001
Trt	1	2	16.24	0.0564
Genotype*Trt	15	60	4.45	<0.0001

and B, respectively; $(\alpha\beta)_{ij}$ refers to the interaction of the above factors; r_k is the random effect due to blocks; and blocks \times whole plot $(ra)_{ik}$ is assumed to contribute to the variation such that $r_k \sim N(0, \sigma_r^2)$ and $(ra)_{ik} \sim N(0, \sigma_{\text{block} \times A}^2)$. This model uses the linear predictor η_{ijk} to estimate the mean of the observations μ_{ijk} .

The specification of the this GLMM is as follows:

Distribution: $y_{ijk} \mid r_k, (ra)_{rk} \sim \text{Binomial}(N_{ijk}, \pi_{ijk})$
 $r_k \sim N(0, \sigma_r^2)$,
 $(ra)_{rk} \sim N(0, \sigma_{\text{block} \times A}^2)$
 Link function: $\text{logit}(\pi_{ijk}) = \eta_{ijk}$.

The following SAS GLIMMIX program allows the fitting of a GLMM with a split-plot structure in a randomized complete block design with a binomial response.

```
proc glimmix data=spd_pp nobound method=quadrature;
class Genotype Trt Block ;
model y/N = Genotype | Trt;
random intercept trt /subject=block;
lsmeans Genotype | Trt /lines ilink;
run;
```

The program uses the quadrature estimation method (**method=quadrature**). This estimation method produces similar results as the Laplace method. Part of the results is provided in Table 6.10. Pearson's chi-squared/DF value in part (a) gives an idea of whether there is overdispersion or extra-variation in the dataset. In this case, Pearson's chi - square/DF = 1.97 indicates that there is overdispersion in the dataset, so it is feasible to use either the pseudo-likelihood (PL) estimation method or a different distribution. In addition to these results, the variance component estimated due to blocks and blocks \times genotype (the whole plot) in part (b) are $\sigma_{\text{block}}^2 = 0.004272$ and $\sigma_{(\text{block} \times A)}^2 = 0.03344$, respectively. The results of the fixed

effects tests (part (c)) indicate that the effect of genotype and the interaction between genotype and treatment are significant.

The appropriate method for model evaluation depends on whether or not there is evidence of overdispersion, so we consider this issue below. The residual variance incorporates systematic discrepancies between the model and the observed responses, variation between replicates (observations in independent experimental units with the same values of the explanatory variables) and sampling variation arising from the distribution of the data; in this case, it is the binomial distribution. If there are no duplicate observations and the fitted model provides an adequate description of the systematic trend, then only sampling variation contributes to the residual variance. If this is true, then the residual deviation has an approximate chi-squared distribution with degrees of freedom similar to the mean squared error (MSE) (the residual).

Since there is overdispersion in the data using the binomial distribution, there are three alternatives we can explore: (1) review the linear predictor, which involves carefully revising the analysis of variance table; (2) add a scale parameter; or (3) use another distribution for the dataset. Each of these three possible alternatives is discussed below, in this order.

6.4.1.1 Linear Predictor Review (η_{ijk})

If the proportion of normal plants (π_{ijk}) is being affected by the genotype within each infestation level ($\text{trt} = \alpha_i$) from plot to plot within each of the blocks, then a nested factorial effect of genotype within infestation levels (trt) could be included in the analysis of variance. Thus, the linear predictor would be defined as

$$\eta_{ijk} = \eta + \alpha_i + r_k + (ra)_{ik} + \beta(\alpha)_{j(i)}$$

where α_i , $\beta(\tau)_{j(i)}$, r_k , and $(ra)_{ik}$ are the fixed effects due to treatments, the effect of genotypes nested within a treatment, random effects due to blocks ($r_k \sim N(0, \sigma_r^2)$), and the interaction between blocks and treatment ($(ra)_{ik} \sim N(0, \sigma_{RA}^2)$), respectively.

The following GLIMMIX syntax estimates the above linear predictor:

```
proc glimmix data=spd_pp method=laplace;
class Genotype Trt Block ;
model y/N = Trt genotype (trt) ;
random trt/subject=block;
lsmeans genotype (trt)/lines ilink slice=trt slicediff=trt;
run;
```

The only difference between this proc GLIMMIX and the previous one is that in this program, we have included the nested effect of genotypes within treatment, genotype (trt), and removed only the fixed effects of genotypes. Part of the results is shown in Table 6.11. The value of Pearson's chi-squared/DF statistic (part (a)) as

Table 6.11 Results of the analysis of variance, under a new linear predictor

(a) Fit statistics for conditional distribution				
-2 Log L (y r. effects)				527.82
Pearson's chi-square				189.07
Pearson's chi-square/DF				1.97
(b) Covariance parameter estimates				
Cov Parm	Subject	Estimate	Standard error	
Intercept	Bloque	0.004265	0.02740	
Trt	Bloque	0.03343	0.03544	
(c) Type III tests of fixed effects				
Effect	Num DF	Den DF	F-value	Pr > F
Trt	1	2	16.20	0.0565
Genotype (Trt)	30	60	15.83	<0.0001

well as the fit statistics did not decrease when modifying the linear predictor. However, the *F*-values calculated for treatments and genotypes within treatments (part (c)) are smaller than those obtained in the split-plot design.

Since the overdispersion is still present (Pearson's chi - square/DF = 1.97), another alternative is to add a scaling parameter to the model. This alternative is presented below.

6.4.1.2 Scale Parameter

If the residual deviation is larger than expected when compared to critical values of the appropriate chi-squared distribution, and if this cannot be corrected by redefining the linear predictor of the model, then there is more variation present than can be accounted for by the distributional likelihood assumption. In this case, we say that the data show overdispersion. The simplest way to deal with overdispersion is to extend the model for scaling the variance function. Adding the scale parameter replaces $\text{Var}(y_{ij}) = \pi_{ij}(1 - \pi_{ij})$ with $\text{Var}(y_{ij}) = \phi\pi_{ij}(1 - \pi_{ij})$. The rationale for this approach is discussed by Collett (2002). The parameter ϕ is a scale factor, called the dispersion parameter, which is used to summarize the degree of overdispersion present in the observations. Clearly, $\phi = 1$ corresponds to the original distribution model. This parameter can be estimated in several different ways. The logarithm of the likelihood of the binomial distribution is given by

$$\log \binom{N}{y_{ij}} + y_{ij} \log \left(\frac{\pi_{ij}}{1 - \pi_{ij}} \right) + N \log(1 - \pi_{ij})$$

In the logarithm of the likelihood, the term “ $y_{ij} \log \left(\frac{\pi_{ij}}{1 - \pi_{ij}} \right)$ ” is very important; any quantity that multiplies y_{ij} is known as the natural or canonical parameter, and this parameter is always a function of the mean. For the binomial distribution, the mean

$N_{ij}\pi_{ij}$ and the natural parameter is $\log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right)$, and, in categorical data, it is known as “log odds.” The generalized estimating equation (GEE) method provides a valid analysis for marginal means, since under a binomial distribution, in the quasi-likelihood, the variance of the distribution is given by $\phi\pi_{ij}(1-\pi_{ij})$. This is achieved by adding the “random _residual_” command in the following SAS syntax.

The following GLIMMIX commands are used to invoke the scale parameter but using the first predictor proposed for these data.

```
proc glimmix data=spd_pp nobound;
class GenotypeTrtBlock ;
model y/N = Trt | genotype;
random intercept trt/subject=block;
random _residual_;
lsmeans Trt | genotype/lines ilink ;
run;
```

In this syntax, we still keep the binomial distribution (y/N is equivalent to telling GLIMMIX in SAS that it is a binomial response) but will add the “random _residual_” command. In this case, we cannot obtain the maximum likelihood estimators because we cannot implement the Laplace method (“method = laplace”) or adaptive quadrature (“method = quad”) approximation method, so the estimation is performed through the pseudo-likelihood (PL) method. This causes the scale parameter to be estimated, and, consequently, it is used in the adjustment of all standard errors and statistical tests. Proc GLIMMIX uses the generalized statistics of McCullagh and Nelder (1989), i.e., χ^2/df as the estimator of the scale parameter ($\hat{\phi}$). All standard errors from the analysis under a binomial distribution are multiplied by $\sqrt{\hat{\phi}}$, and all F -tests are divided by $\hat{\phi}$ to account for overdispersion. Part of the output is shown below.

The value of Pearson’s statistic in part (a) indicates that overdispersion has not been eliminated. Chi – square/DF = 3.13, on the contrary, indicates that this value has increased. This result indicates that adding a scale parameter to the model does not decrease the extra-variation present in the dataset, since the binomial assumption forces a relationship between the mean and variance of the data that might not contain the data being analyzed. On the other hand, the estimated scale parameter is $\hat{\phi} = 3.1263$ (part (b)). Pearson’s residual analysis showed that its variance is 3.6257, which is considerably larger than 1, implying a large overdispersion. In addition, the results of the fixed effects tests (part (c)) vary from those above (Table 6.12).

Therefore, the third option based on assuming an alternative distribution (beta distribution) on the response variable is discussed below.

Table 6.12 Results of the analysis of variance, adding a scale parameter to the model

(a) Fit statistics				
-2 Res log pseudo-likelihood				182.52
Generalized chi-square				200.09
Gener. chi-square/DF				3.13
(b) Covariance parameter estimates				
Cov Parm	Subject	Estimate	Standard error	
Intercept	Bloque	0.005416	0.04750	
Trt	Bloque	0.03202	0.06338	
Residual variance component (VC)		3.1263	0.5719	
(c) Type III tests of fixed effects				
Effect	Num DF	Den DF	F-value	Pr > F
Trt	1	2	10.20	0.0856
Genotype	15	60	9.04	<0.0001
Genotype*Trt	15	60	1.42	0.1674

6.4.1.3 Alternative Distribution

Another approach to control the overdispersion would be to use a different distribution in the interval [0, 1], such as the beta distribution, to model the data. Generally, this distribution yields good results when all experiments have the same number of observations (successes and failures), i.e., when $N_{ijk} = N$. When N_{ijk} varies a little, even in many cases, the beta distribution yields acceptable results. It is important to mention that the proportions come from binomial counts, and, therefore, we now define the response variable as $p_{ijk} = \frac{y_{ijk}}{N_{ijk}}$ so that it can be modeled as the beta distribution. The components of the beta response model are listed below:

Distribution: $p_{ijk} \mid r_k, (ra)_{rk} \sim \text{Beta}(\pi_{ijk}, \phi)$ with ϕ as the scale parameter

$r_k \sim N(0, \sigma_r^2), (ra)_{rk} \sim N(0, \sigma_{RA}^2)$

Linear predictor: $\eta_{ijk} = \eta + \alpha_i + r_k + (\alpha r)_{ik} + \beta_j + (\alpha\beta)_{ij}$

Link function: $\text{logit}(\pi_{ijk}) = \text{logit}\left(\frac{\pi_{ijk}}{1 - \pi_{ijk}}\right) = \eta_{ijk}$

As mentioned before, we now use the response variable $p_{ijk} = \frac{y_{ijk}}{N_{ijk}}$. This new response variable p_{ijk} is not the same as the one used in the binomial distribution. The following SAS commands fit a GLMM in a split-plot randomized complete block design with a beta response. It is important to mention that before implementing this model in SAS GLIMMIX, the variable $p = p_{ijk} = \frac{y_{ijk}}{N_{ijk}}$ was defined.

```
proc glimmix data=spd_pp nobound method=laplace;
class GenotypeTrtBlock ;
model p = Genotype | Trt / dist=beta;
random intercept trt / subject=block;
lsmeans Genotype | Trt / lines ilink;
run;
```

Table 6.13 Fit statistics assuming binomial and beta distributions

(a) Fit statistics		
Distribution	Binomial	Beta
-2 Log likelihood	541.85	-246.49
AIC (smaller is better)	609.85	-176.49
AICC (smaller is better)	648.87	-132.28
BIC (smaller is better)	579.20	-208.04
CAIC (smaller is better)	613.20	-173.04
HQIC (smaller is better)	548.24	-239.91
(b) Fit statistics for conditional distribution		
Distribution	Binomial	Beta
-2 Log L (y r. effects)	527.82	-254.68
Pearson's chi-square	189.09	93.95
Pearson's chi-square/DF	1.97	1.01

Table 6.14 Results of the analysis of variance, assuming binomial and beta distributions

(a) Covariance parameter estimates						
Cov Parm	Subject	Binomial		Beta		
		Estimate	Standard error	Estimate	Standard error	
Intercept	Bloque	0.004272	0.02741	-0.00524	.	
Trt	Bloque	0.03344	0.03545	0.02175	0.1475	
Scale ($\hat{\phi}$)			.	25.7070		
(b) Type III tests of fixed effects						
Effect	Num DF	Den DF	Binomial		Beta	
			F-value	Pr > F	F-value	Pr > F
Trt	1	4	16.24	0.0564	9.98	0.0342
Genotype	15	60	28.28	<0.0001	13.25	<0.0001
Genotype*Trt	15	60	4.45	<0.0001	2.23	0.0146

Some of the SAS GLIMMIX output is listed below. Based on the fit statistics under the binomial (first alternative) and beta distributions (Table 6.13), clearly the values of the statistics related to the degree of overdispersion are lower in the beta distribution than in the binomial distribution, indicating that the beta distribution provides a better fit (part (a)). Looking at the fit statistics for the conditional model in part (b), the values of the three fit statistics in the binomial model are higher than the values in the beta model. The value of Pearson's chi - square/DF under the beta distribution is 1.01. This value indicates that the overdispersion has been virtually eliminated from the data and that therefore the beta distribution is a better candidate model for this dataset.

Adding the scale parameter (ϕ) to the model, the variance components and standard errors in Table 6.14 cause (part (a)) variation for each of the results and, therefore, the F- and t-tests are affected (part (b)). The estimated value of the scale

Table 6.15 Estimated means and standard errors on the model scale and the data scale

(a) Trt least squares means							
Trt	Estimate	Standard error	DF	<i>t</i> -value	Pr > <i>t</i>	Mean	Standard error mean
Trt1	-1.2362	0.01768	2	-69.94	0.0002	0.2251	0.003083
Trt2	-1.9327	0.01768	2	-109.34	<0.0001	0.1264	0.001952

(b) Genotype least squares means							
Genotype	Estimate	Standard error	DF	<i>t</i> -value	Pr > <i>t</i>	Mean	Standard error mean
G1	0.1524	0	57	Infty	<0.0001	0.5380	0
G10	-1.4143	0	57	-Infty	<0.0001	0.1956	0
G11	-1.8698	0	57	-Infty	<0.0001	0.1336	0
G12	-2.8971	0.03885	57	-74.58	<0.0001	0.05230	0.001925
G13	-1.4336	0	57	-Infty	<0.0001	0.1925	0
G14	-1.8761	0.1304	57	-14.39	<0.0001	0.1328	0.01502
G15	-1.8618	0	57	-Infty	<0.0001	0.1345	0
G16	-2.6686	0	57	-Infty	<0.0001	0.06485	0
G2	0.2225	0	57	Infty	<0.0001	0.5554	0
G3	-1.3329	0	57	-Infty	<0.0001	0.2087	0
G4	-1.5897	0	57	-Infty	<0.0001	0.1694	0
G5	-1.3696	0	57	-Infty	<0.0001	0.2027	0
G6	-2.0173	0	57	-Infty	<0.0001	0.1174	0
G7	-1.7001	0.1356	57	-12.53	<0.0001	0.1545	0.01771
G8	-1.7161	0	57	-Infty	<0.0001	0.1524	0
G9	-1.9796	0	57	-Infty	<0.0001	0.1214	0

parameter is $\hat{\phi} = 25.7018$. The variance components based on the binomial model and beta are listed below.

The treatment means (part (a)) and genotypes (part (b)) are presented in Table 6.15. The estimates on the model scale are listed under the column “Estimate” with their respective standard errors “Standard error,” and the values on the data scale are listed under the column “MEAN” with their respective standard errors “Standard error mean.” In the table of least squares means for the effect of genotypes, inconsistencies are observed in the values of *t* and in the standard error values of the means, so other estimation alternatives should be sought.

In large samples, both binomial and normal distributions are quite similar. Logically, the latter two analyses, binomial and beta, are attractive because of their consistency with the nature of the data. Because of the inconsistencies in the estimates of the mean for genotypes (*t*value = Infty and standard error of the mean), a robust method of estimation could be used; in this case, this is the normal distribution.

Assuming that p_{ijk} has a normal distribution with a mean μ_{ijk} and constant variance σ^2 , the components of this model are as follows:

Table 6.16 Results of the analysis of variance, assuming a normal distribution

(a) Fit statistics				
-2 Res log likelihood		-79.38		
AIC (smaller is better)		-73.38		
AICC (smaller is better)		-72.98		
BIC (smaller is better)		-76.08		
CAIC (smaller is better)		-73.08		
HQIC (smaller is better)		-78.81		
Generalized chi-square		0.60		
Gener. chi-square/DF		0.01		
(b) Covariance parameter estimates				
Cov Parm	Estimate	Standard error		
Bloque	0.000123	0.000742		
Trt*bloque	0.000329	0.000925		
Residual	0.009442	0.001724		
(c) Type III tests of fixed effects				
Effect	Num DF	Den DF	F-value	Pr > F
Genotype	15	60	12.59	<0.0001
Trt	1	2	20.46	0.0456
Genotype*Trt	15	60	2.93	0.0016

Distribution: $\text{pct}_{ijk} | r_k, (ra)_{ik} \sim \text{Normal}(\mu_{ijk}, \sigma^2)$

$r_k \sim N(0, \sigma_r^2), (ra)_{ik} \sim N(0, \sigma_{RA}^2)$

Linear predictor: $\eta_{ijk} = \eta + \alpha_i + r_k + (ar)_{ik} + \beta_j + (\alpha\beta)_{ij}$

Link function: $\eta_{ijk} = \mu_{ijk}$; identity

Similarly, in this example, the response variable used was $\text{pct}_{ijk} = \frac{y_{ijk}}{N_{ijk}}$. This new response variable pct_{ijk} is not the same as the response variable used in the binomial distribution. The following SAS GLIMMIX commands adjust a linear mixed model (LMM) under a split plot in a randomized complete block design with a normal response.

```
proc glimmix data=spd_pct nobound;
class Genotype Trt Block ;
model pct = Genotype|Trt;
random block block*trt;
lsmeans Genotype|Trt/lines;
run;
```

Part of the results is shown below. The values of fit statistics in part (a) of Table 6.16 for the model are clearly lower than those estimated in the previous options. This indicates that the normal distribution is reasonable, even though the response is a proportion. The estimated variance components, tabulated in part (b) due to blocks, blocks x treatment, and the mean squared error (MSE) (Residual = Gener. chi-square/DF) are $\hat{\sigma}_{\text{block}}^2 = 0.000123$, $\hat{\sigma}_{\text{block} \times \text{trt}}^2 = 0.00039$, and $\hat{\sigma}^2 = \text{MSE} = 0.009442 \cong 0.01$, respectively.

Table 6.17 Means and standard errors for genotypes and treatments

(a) Genotype least squares means							
Genotype	Estimate	Standard error	DF	<i>t</i> -value	Pr > <i>t</i>	Mean	Standard error mean
G1	0.5260	0.04086	60	12.87	<0.0001	0.5260	0.04086
G10	0.1340	0.04086	60	3.28	0.0017	0.1340	0.04086
G11	0.1522	0.04086	60	3.73	0.0004	0.1522	0.04086
G12	0.03332	0.04086	60	0.82	0.4179	0.0333	0.04086
G13	0.2026	0.04086	60	4.96	<0.0001	0.2026	0.04086
G14	0.1342	0.04086	60	3.28	0.0017	0.1342	0.04086
G15	0.1360	0.04086	60	3.33	0.0015	0.1360	0.04086
G16	0.05625	0.04086	60	1.38	0.1737	0.0562	0.04086
G2	0.5355	0.04086	60	13.11	<0.0001	0.5355	0.04086
G3	0.2139	0.04086	60	5.24	<0.0001	0.2139	0.04086
G4	0.1751	0.04086	60	4.28	<0.0001	0.1751	0.04086
G5	0.2035	0.04086	60	4.98	<0.0001	0.2035	0.04086
G6	0.1301	0.04086	60	3.18	0.0023	0.1301	0.04086
G7	0.1671	0.04086	60	4.09	0.0001	0.1671	0.04086
G8	0.1504	0.04086	60	3.68	0.0005	0.1504	0.04086
G9	0.1187	0.04086	60	2.90	0.0051	0.1187	0.04086
(b) Trt least squares means							
Trt	Estimate	Standard error	DF	<i>t</i> -value	Pr > <i>t</i>	Mean	Standard error mean
Trt1	0.2478	0.01863	2	13.30	0.0056	0.2478	0.01863
Trt2	0.1358	0.01863	2	7.29	0.0183	0.1358	0.01863

The *F*-statistics for the fixed effects of genotype, treatments, and the interaction between both factors provide significant statistical evidence on the proportion of infested carrots in each of the genotypes (part (c)). Overall, the least squares means for genotypes and treatments are reported in Table 6.17 in parts (a) and (b). The genotypes showing the highest fraction of infested carrots were 1, 2, 3, 5, and 13, whereas genotypes 12 and 16 showed the lowest percentage of infested carrots. Now, for treatments, the highest proportion of infested carrots was observed in treatment 1 with 24.78%, whereas in treatment 2, it was 13.58%.

Based on the fixed effects tests, the interaction effect of genotype x treatment on the proportion of infested carrots was statistically different. Genotypes 9 and 16 showed higher susceptibility in treatment 1 followed by treatment 2, whereas genotypes 5, 11, 13, and 15 showed the same proportions of infested carrots in both treatments (Fig. 6.6). On the other hand, genotypes that showed higher resistance to infestation levels were genotypes 1, 2, and 6 followed by genotypes 3, 4, 7, 8, 10, and 12.

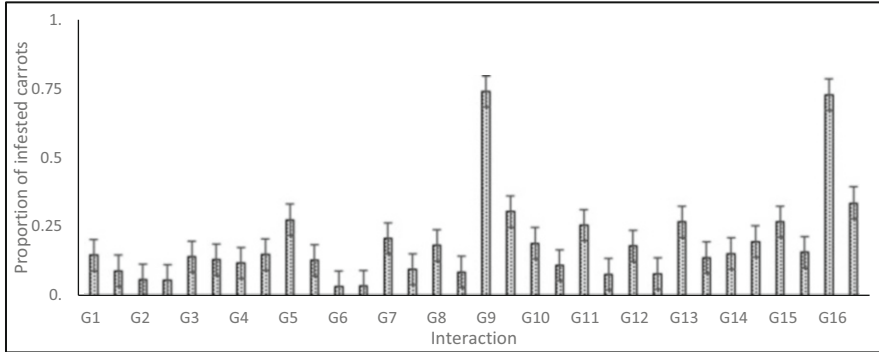


Fig. 6.6 The average proportion of infested carrots in genotypes as a function of treatment

6.5 A Split-Split Plot in an RCBD:- In Vitro Germination of Seeds

The growth of a plant in a tissue culture can be explained by various combined effects of A, B, and C factors. For this, the availability and efficient use of chemical resources (factors) is of great relevance when availability is scarce or too expensive. In light of this, the combination of three reagents (A, B, and C), reagent A at three levels and reagents B and C at two levels, were tested on the in vitro germination of orchid seeds. The combination of the levels of each of the factors is schematized below.

Block 1											
A ₃				A ₁				A ₂			
B ₁		B ₂		B ₁		B ₂		B ₂		B ₁	
C ₂	C ₂	C ₂	C ₂	C ₂	C ₂	C ₂	C ₂	C ₂	C ₂	C ₂	C ₂
C ₁	C ₁	C ₁	C ₁	C ₁	C ₁	C ₁	C ₁	C ₁	C ₁	C ₁	C ₁

Block 2											
A ₂				A ₁				A ₃			
B ₁		B ₂		B ₁		B ₂		B ₂		B ₁	
C ₁	C ₁	C ₁	C ₁	C ₁	C ₁	C ₁	C ₁	C ₁	C ₁	C ₁	C ₁
C ₂	C ₂	C ₂	C ₂	C ₂	C ₂	C ₂	C ₂	C ₂	C ₂	C ₂	C ₂

In each of the factor combinations, N orchid seeds were placed to germinate for a period of time. Let y_{ijk} be the number of seeds germinated at the i th level of factor A, at the j th level of factor B, and at the k th level of factor C. Since the observations are made at the sub-subplot level, conditional on the structural effects of the design, these observations have a variance associated with the subplot. Therefore, the statistical model for this experiment is given below:

Table 6.18 Number of seeds that germinated (y_{ijkl}) in each of the factor combinations

Block	A	B	C	Y	N
1	1	1	1	15	73
2	1	1	1	10	86
1	1	1	2	17	69
2	1	1	2	19	32
1	1	2	1	26	125
2	1	2	1	21	62
1	1	2	2	14	81
2	1	2	2	12	21
1	2	1	1	10	92
2	2	1	1	12	108
1	2	1	2	30	44
2	2	1	2	32	33
1	2	2	1	37	91
2	2	2	1	30	42
1	2	2	2	32	98
2	2	2	2	37	44
1	3	1	1	18	52
2	3	1	1	18	73
1	3	1	2	23	108
2	3	1	2	21	55
1	3	2	1	24	106
2	3	2	1	27	92
1	3	2	2	37	64
2	3	2	2	37	97

Distribution: $y_{ijkl} \mid r_l, (r\alpha)_{il}, (r\alpha\beta)_{ijl} \sim \text{Binomial}(N_{ijk}, \pi_{ijk})$
 $r_l \sim N(0, \sigma_r^2), (r\alpha)_{rk} \sim N(0, \sigma_{RA}^2), (r\alpha\beta)_{ijl} \sim N(0, \sigma_{rab}^2)$

Linear predictor:

$\eta_{ijk} = \eta + \alpha_i + r_l + (r\alpha)_{il} + \beta_j + (\alpha\beta)_{ij} + (r\alpha\beta)_{ijl} + \gamma_k + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk} + (\alpha\beta\gamma)_{ijk}$,
 where blocks (r_l), blocks \times A ($(r\alpha)_{il}$), and blocks \times A \times B ($(r\alpha\beta)_{ijl}$) are assumed to contribute to the variation such that $r_l \sim N(0, \sigma_r^2), (r\alpha)_{il} \sim N(0, \sigma_{r \times A}^2), (r\alpha\beta)_{ijl} \sim N(0, \sigma_{rab}^2)$, respectively, and ε_{ijkl} experimental errors are distributed as $N(0, \sigma^2)$. This model uses the linear predictor η_{ijk} to estimate the mean of the observations μ_{ijk} .

Link function: $\text{logit}(\pi_{ijkl}) = \eta_{ijkl}$

Table 6.18 below shows the data obtained from this experiment.

Table 6.19 presents the analysis of variance and shows the sources of variation and degrees of freedom for this experimental design.

The following SAS GLIMMIX program allows a GLMM with a split-split plot structure to be fitted in an RCBD with a binomial response.

Table 6.19 Sources of variation and degrees of freedom for the randomized block design with an arrangement of treatments under the split-split-plot structure

Sources of variation	Degrees of freedom
Blocks	$r - 1 = 2 - 1 = 1$
Factor A	$a - 1 = 3 - 1 = 2$
Error _a (Bloque*A)	$(r - 1)(a - 1) = 2$
Factor B	$b - 1 = 2 - 1 = 1$
A*B	$(a - 1)(b - 1) = 2$
Error _b (A*B(Bloque))	$a(b - 1)(r - 1) = 3 \times 1 \times 1 = 3$
Factor C	$(c - 1) = 2 - 1 = 1$
A*C	$(3 - 1)(2 - 1) = 2$
B*C	$(b - 1)(c - 1) = 1$
A*B*C	$(a - 1)(b - 1)(c - 1) = 2$
Error	$ab(c - 1)(r - 1) = 3 \times 2 \times 1 \times 1 = 6$
Total	$r \times a \times b \times c - 1 = 2 \times 3 \times 2 \times 2 - 1 = 23$

```
proc GLIMMIX data=germ nobound method=laplace;
class Block A B C;
model Y/N = A | B | C / dist=binomial link=logit;
random block block*A block*A block*A*B;
lsmeans A | B | C / lines ilink;
run;
```

Part of the output is shown in Table 6.20. The value of the conditional statistic Pearson's chi-square/DF = 1.81 (part (a)) indicates that there is an overdispersion in the dataset since these values are greater than 1. The estimated variance components tabulated in part (b) correspond to blocks, blocks × factor A, and blocks × factor A × factor B, which are $\sigma_r^2 = 0.0752$, $\sigma_{rA}^2 = 0.088$, and $\sigma_{rab}^2 = 0.0425$, respectively. The type III tests of fixed effects are shown in part (c). Here, we see that the test of equality of treatments is not significant for factors A and B and the interaction AB ($A, P = 0.1917, B, P = 0.0897; AB, P = 0.6262$), whereas for factor C and the interactions AC, BC, and ABC, it is significant at a level of 5%.

Since there is overdispersion in the dataset, the binomial distribution does not provide a good fit for the dataset (Pearson's chi-square/DF = 1.81). An alternative to model this dataset could be the beta distribution. Under this assumption, let the response variable be $p_{ijk} = \frac{y_{ijk}}{N_{ijk}}$, the proportion of seeds that germinated, then p_{ijk} is assumed to have a beta distribution rather than a binomial distribution for the success count y_{ijk} out of a total of N_{ijk} Bernoulli trials.

The components of the model are listed below:

Distribution: $p_{ijk} | r_i, (ra)_{il}, (ra\beta)_{ijl} \sim \text{Beta}(\pi_{ijk}, \phi)$, with ϕ as the scale parameter.

$$r_i \sim N(0, \sigma_r^2), (ra)_{rk} \sim N(0, \sigma_{rA}^2), (ra\beta)_{ijl} \sim N(0, \sigma_{rab}^2)$$

Linear predictor:

$$\eta_{ijk} = \eta + \alpha_i + r_i + (r\alpha)_{il} + \beta_j + (\alpha\beta)_{ij} + (ra\beta)_{ijl} + \gamma_k + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk} + (\alpha\beta\gamma)_{ijk}$$

Link function: $\text{logit}(\pi_{ijk}) = \text{logit}\left(\frac{\pi_{ijk}}{1 - \pi_{ijk}}\right) = \eta_{ijk}$

Table 6.20 Results of the analysis of variance of the RCBD in the split-split plot under the binomial distribution

(a) Fit statistics for conditional distribution				
-2 Log L (y r. effects)				146.19
Pearson's chi-square				43.49
Pearson's chi-square/DF				1.81
(b) Covariance parameter estimates				
Cov Parm	Estimate		Standard error	
Bloque	0.07521		0.1180	
Bloque*A	0.08847		0.09319	
Bloque*A*B	0.02205		0.04258	
(c) Type III tests of fixed effects				
Effect	Num DF	Den DF	F-value	Pr > F
A	2	2	4.22	0.1917
B	1	3	6.12	0.0897
A*B	2	3	0.55	0.6262
C	1	6	65.73	0.0002
A*C	2	6	11.68	0.0085
B*C	1	6	29.38	0.0016
A*B*C	2	6	31.69	0.0006

The following SAS commands fit a GLMM on a split-split plot in a randomized complete block design assuming a beta distribution for the response variable.

```
proc glimmix data=germ nobound method=laplace;
class BlockABC ;
model p = A|B|C/dist=beta ;
random block block*A block*A*B; /*intercept A /subject=block*/;
lsmeans A|B|C/lines ilink;
run;
```

Part of the results is listed in Table 6.21 under a beta distribution. The value of the fit statistic for the conditional model tabulated in (a) (Pearson's chi - square/DF = 1.01) indicates that overdispersion has been removed and that the beta distribution is a good model to fit the dataset. Part (b) shows the variance component estimates for blocks, $blockxA$, and $blockxAxB$ ($\hat{\sigma}_r^2 = -0.157$, $\hat{\sigma}_{rA}^2 = -0.05558$, and $\hat{\sigma}_{rab}^2 = -0.227$, respectively) and the value of the estimated scale parameter ($\hat{\phi} = 19.2789$). According to the type III tests of fixed effects in part (c), the main effect of factor C ($P = 0.0128$) and interaction $A \times B \times C$ ($P = 0.0424$) are statistically significant at a level of 5%.

The estimates of the interactions are shown in Table 6.22 on the model scale under the "Estimate" column and as probabilities on the data scale under the "Mean" column with its corresponding standard errors under the "Standard error mean" column.

Table 6.21 Results of the analysis of variance of the RCBD in the split-split plot structure under the beta distribution

(a) Fit statistics for conditional distribution				
-2 Log L (p r. effects)				-37.51
Pearson's chi-square				21.31
Pearson's chi-square/DF				1.01
(b) Covariance parameter estimates				
Cov Parm	Estimate	Standard error		
Bloque	-0.1570	.		
Bloque*A	-0.05558	.		
Bloque*A*B	-0.2270	.		
Scale	19.2789	5.8703		
(c) Type III tests of fixed effects				
Effect	Num DF	Den DF	F-value	Pr > F
A	2	2	1.21	0.4521
B	1	2	0.00	0.9687
A*B	2	2	1.08	0.4799
C	1	4	18.34	0.0128
A*C	2	4	1.50	0.3257
B*C	1	4	6.56	0.0626
A*B*C	2	4	7.72	0.0424

Table 6.22 Estimated least mean squares on the model scale ("Estimate" column) and the data scale ("Mean" column)

A*B*C least squares means									
A	B	C	Estimate	Standard error	DF	t-value	Pr > t	Mean	Standard error mean
1	1	1	-0.3769	0.3194	4	-1.18	0.3034	0.4069	0.07709
1	1	2	0.9506	0.3445	4	2.76	0.0509	0.7212	0.06927
1	2	1	0.1721	0.3147	4	0.55	0.6135	0.5429	0.07810
1	2	2	0.7010	0.3308	4	2.12	0.1014	0.6684	0.07331
2	1	1	-0.6521	0.3296	4	-1.98	0.1190	0.3425	0.07422
2	1	2	2.9148	0.8071	4	3.61	0.0225	0.9486	0.03937
2	2	1	0.7430	0.4699	4	1.58	0.1890	0.6776	0.1026
2	2	2	0.4056	0.4515	4	0.90	0.4198	0.6000	0.1084
3	1	1	0.2695	0.3161	4	0.85	0.4419	0.5670	0.07761
3	1	2	0.2752	0.3163	4	0.87	0.4334	0.5684	0.07759
3	2	1	0.1236	0.3143	4	0.39	0.7143	0.5309	0.07827
3	2	2	1.1726	0.3614	4	3.24	0.0315	0.7636	0.06523

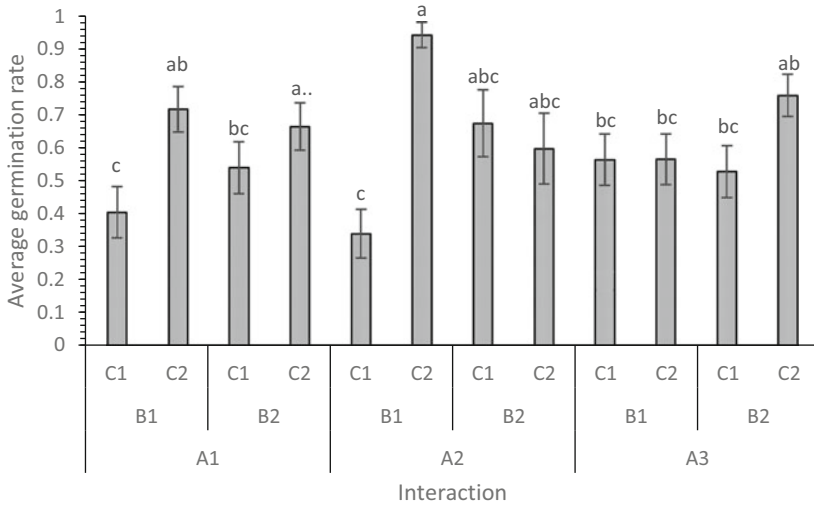


Fig. 6.7 The average seed germination rate

The simple effects of factors show that the best combination of factor levels was $A2*B1*C2$, showing the highest seed germination proportion followed by the combination of factors $A1*B1*C2$, $A3*B2*C2$, and lower proportion, which were observed in the combination of factors $A1*B2*C2$, $A2*B2*C1$ and $A2*B2*C2$ (Fig. 6.7). Finally, the combination of the factor levels $A2 \times B1 \times C1$ showed the lowest proportion of seed germination.

6.6 Alternative Link Functions for Binomial Data

In previous chapters, we used proc GLIMMIX with binomial data and, by default, it works with the link function “logit.” However, in certain applications with binomial data, other link functions are acceptable, either because they make it easier to interpret or because for certain binomial datasets, the link function “logit” cannot accurately model the data and, as a result, produce biased (misleading) results. In this section, we consider two alternative link functions to the logit for binomial data: the link “probit” and the complementary log-log link.

The probit model is also used to model dichotomous (Bernoulli) or binomial (sum of Bernoulli trials) responses. For this model, the link function, called the probit link, uses the inverse of the cumulative distribution function of a standard normal distribution to transform probabilities to the standard normal variable. That is, $\Phi^{-1}(\pi_i) = \eta_i$, which implies that $\pi_i = \Phi(\eta_i)$, where $\Phi(Z) = \int_{-\infty}^z \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}t^2} dt$.

The use of the probit regression model dates back to Bliss (1934). Bliss was interested in finding an effective pesticide to control insects that fed on grape leaves.

He discovered that the relationship between the response and a dose of pesticide was sigmoid, and he applied the probit link function to transform the dose–response curve from a sigmoid to a linear relationship.

The complementary function $\log - \log$ defined as $\eta_i = \log(-\log(1 - \pi_i))$, whose inverse is $\pi_i = 1 - e^{-e^{\eta_i}}$, is useful for data in which most of the probabilities are near zero or near one. For small values of π_i , the log-log transformation produces results highly similar to those produced when using a logit link. As the probability increases, the transformation approaches infinity more slowly than the probit or logit model.

6.6.1 Probit Link: A Split-Split Plot in an RCBD with a Binomial Response

This example takes the dataset of the split-split plot in an RCBD (Exercise 6.8.5). In this example, the data were modeled using the function “logit.” In this exercise, we will fit the dataset using the link function “probit,” and we will compare and contrast the results using a logit link. The components of the GLMM are identical to those in Example 6.5, except for the link function. That is, we replace:

Link function: $\text{logit}(\pi_{ijk}) = \text{logit}\left(\frac{\pi_{ijk}}{1 - \pi_{ijk}}\right) = \eta_{ijk}$ by $\Phi^{-1}(\pi_{ijk}) = \eta_{ijk}$.

The following GLIMMIX syntax implements the fitting of the binomial data using the link function “probit.”

```
proc glimmix data=germ nobound method=laplace;
class Block A B C;
model Y/N = A | B | C / link=probit;
random block block*A block*A*B;
lsmeans A | B | C / lines ilink;
run;
```

Table 6.23 shows part of the results under the binomial distribution with the “probit” link function. In parts (a) and (b), we see the mean squared error and variance component estimates for blocks, whole plot, subplot, and sub-subplot, where it can be observed that these values are positive and not negative, as the ones obtained with the link function “logit.” Since the variance components are positive, this analysis makes more sense than the one based on the logit link.

The type III tests of fixed effects are tabulated in part (c) of Table 6.23; the main effects of factors A and B and the interactions $A*B$, $A*C$, and $B*C$ are not significant in both link functions, whereas the main effect of factor C and the interaction $A*B*C$ are statistically significant under the “probit” link.

The estimated probabilities ($\hat{\pi}_{ijk}$) and their respective standard errors are presented in Table 6.24 for each of the combinations of the three factors, which

Table 6.23 Results of the analysis of variance of the RCBD in the split-split plot structure under the binomial distribution using the “probit” link

(a) Fit statistics for conditional distribution					
-2 Log L (y r. effects)					146.43
Pearson’s chi-square					43.01
Pearson’s chi-square/DF (CME = $\hat{\sigma}^2$)					1.09
(b) Covariance parameter estimates					
Cov Parm	Estimate			Standard error	
Block ($\hat{\sigma}_{\text{block}}^2$)	0.02411			0.03707	
Block*A ($\hat{\sigma}_{\text{block} \times A}^2$)	0.02128			0.02830	
Block*A*B ($\hat{\sigma}_{\text{block} \times A \times B}^2$)	0.01617			0.01896	
(c) Type III tests of fixed effects					
Effect	Num DF	Den DF	F-value	Probit Pr > F	Logit Pr > F
A	2	2	5.49	0.1541	0.4521
B	1	3	4.17	0.1339	0.9687
A*B	2	3	0.36	0.7226	0.4799
C	1	6	67.13	0.0002	0.0128
A*C	2	6	12.34	0.0075	0.3257
B*C	1	6	29.16	0.0017	0.0626
A*B*C	2	6	33.93	0.0005	0.0424

are very similar in both link functions. However, the average standard error is slightly higher with the “logit” link function ($\text{standar.error.mean}_{\text{logit}} = 0.0711$) compared to the “probit” link ($\text{standar.errormean}_{\text{probit}} = 0.0693$).

6.6.2 Complementary Log-Log Link Function: A Split Plot in an RCBD with a Binomial Response

Researchers studied three different micro-minerals (A, B, and C) on the attachment of explants of a commercial culture. In this vein, micro-mineral A was tested at three levels ($i = 1, 2, \text{ and } 3$), and micro-minerals B and C at two levels ($j, k = 1, 2 \text{ and}$). The combination of the different levels yielded a total of 12 combinations. Since the researchers wanted to study factor C with greater precision, a split-plot treatment structure was designed in which micro-minerals A and B were placed in the whole plot (a large plot) and micro-mineral C in the subplot (a small plot). Treatment factor combinations were placed in an RCBD manner ($r = 1, 2$). The outcome of interest was the number of live plants ($y_{ijk r}$) out of the total number of plants growing in the

Table 6.24 Means and standard errors using the probit and logit link functions

A*B*C least squares means						
A	B	C	Probit		Logit	
			Mean	Standard error mean	Mean	Standard error mean
1	1	1	0.1543	0.05050	0.1494	0.04796
1	1	2	0.3723	0.08296	0.3780	0.08767
1	2	1	0.2724	0.06746	0.2694	0.06896
1	2	2	0.2954	0.07798	0.2953	0.08053
2	1	1	0.1023	0.03805	0.09593	0.03409
2	1	2	0.8255	0.06338	0.8292	0.06135
2	2	1	0.5684	0.08306	0.5703	0.08845
2	2	2	0.5529	0.08327	0.5530	0.08847
3	1	1	0.2844	0.07196	0.2844	0.07418
3	1	2	0.2751	0.06868	0.2733	0.07041
3	2	1	0.2568	0.06452	0.2563	0.06589
3	2	2	0.4612	0.08017	0.4608	0.08553

unit (n_{ijkl}). The data can be referred to in the Appendix (Data: Commercial crop explant attachment).

The GLMM for this experiment is described below (log-log data):

Distribution: $y_{ijkl} | r_l, r(a\beta)_{ijl} \sim \text{Binomial}(N_{ijk}, \pi_{ijk})$

$r_l \sim N(0, \sigma_r^2), r(a\beta)_{ijl} \sim N(0, \sigma_{rab}^2),$

Linear predictor: $\eta_{ijkl} = \eta + r_l + \alpha_i + \beta_j + (\alpha\beta)_{ijl} + r(a\beta)_{il} + \gamma_k + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk} + (\alpha\beta\gamma)_{ijk},$
 $i + \beta_j + (\alpha\beta)_{ijl} + r(a\beta)_{il} + \gamma_k + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk} + (\alpha\beta\gamma)_{ijk},$ where blocks (r_l) and blocks x ($A \times B$) ($(r(a\beta))_{ijl}$) are assumed to contribute to the variation such that $r_l \sim N(0, \sigma_r^2)$ and $r(a\beta)_{ijl} \sim N(0, \sigma_{rab}^2),$ respectively.

Link function: $\log - \log(\pi_{ijkl}) = \eta_{ijkl}$

The following GLIMMIX code adjusts the binomial proportions with a complementary link function $\log - \log$ in an RCBD manner.

```
proc glimmix data=spp nobound method=laplace;
class block A B C;
model y/n = A | B | C / link=ccll;
random block block (A*B);
lsmeans A | B | C / lines ilink;
run;
```

The “link = ccll” option specifies that “proc GLIMMIX” will fit the model using the complementary ($\log - \log$) link function. The “lsmeans A|B|C / lines ilink” command calls for estimation of the linear predictors η_{ijk} , whereas the “lines” and “ilink” options provide the comparison between the linear predictors and their inverse. Part of the output is shown below. Table 6.25 shows the variance component estimates of blocks and blocks ($A \times B$) using alternative link functions. Under

Table 6.25 Variance component estimates using the same distribution but a different link function

Covariance parameter estimates						
	Log – log		Logit		Probit	
Cov Parm	Estimate	Standard error	Estimate	Standard error	Estimate	Standard error
Block	0.05808	0.07112	0.08144	0.1042	0.02676	0.03494
Block (A*B)	0.05065	0.03121	0.09203	0.05754	0.03374	0.02111

Table 6.26 Type III tests of fixed effects using the same distribution but with a different link function

Type III tests of fixed effects								
Effect	Num DF	Den DF	Log – log		Logit		Probit	
			F-value	Pr > F	F-value	Pr > F	F-value	Pr > F
A	2	5	6.27	0.0434	7.44	0.0318	8.17	0.0266
B	1	5	4.85	0.0789	3.13	0.1370	2.81	0.1543
A*B	2	5	0.65	0.5613	0.28	0.7693	0.24	0.7971
C	1	6	68.84	0.0002	65.29	0.0002	66.70	0.0002
A*C	2	6	11.94	0.0081	11.53	0.0088	12.12	0.0078
B*C	1	6	27.51	0.0019	28.88	0.0017	28.77	0.0017
A*B*C	2	6	32.44	0.0006	32.36	0.0006	33.93	0.0005

Table 6.27 Fit statistics using the same distribution but a different link function

Covariance parameter estimates			
	Log – log	Logit	Probit
–2 Log likelihood	164.85	172.57	170.88
AIC (smaller is better)	192.85	200.57	198.88
AICC (smaller is better)	239.51	247.24	245.55
BIC (smaller is better)	174.55	182.27	180.59
CAIC (smaller is better)	188.55	196.27	194.59
HQIC (smaller is better)	154.58	162.31	160.62

the link “probit,” the variance components are smaller compared to those obtained with the link functions “log – log” and “logit.”

The values of the hypothesis tests for the fixed effects, both main effects and interactions, are shown in Table 6.26. The three link functions behave similarly.

One tool that might be useful in choosing which link function provides a better fit, or which best describes the variability of a dataset, is the model fit statistics. The fit statistics indicate that the model with the complementary “log – log” link function provides the best fit (Table 6.27).

Table 6.28 shows the maximum likelihood estimators ($\hat{\pi}_{ijk}$) for each of the link functions and the combination of factor levels, and it can be verified that they provide very similar estimates. It is important to mention that the correct

Table 6.28 Means and standard errors using the same distribution but with a different link function

A*B*C least squares means								
A	B	C	Log – log		Logit		Probit	
			Mean	Standard error mean	Mean	Standard error mean	Mean	Standard error mean
1	1	1	0.1494	0.04259	0.1513	0.04732	0.1547	0.05030
1	1	2	0.3776	0.08554	0.3727	0.08510	0.3696	0.08223
1	2	1	0.2661	0.06257	0.2706	0.06744	0.2737	0.06718
1	2	2	0.3001	0.07718	0.2993	0.07951	0.2980	0.07789
2	1	1	0.1020	0.03079	0.1023	0.03451	0.1047	0.03829
2	1	2	0.8389	0.08212	0.8188	0.06189	0.8196	0.06375
2	2	1	0.5558	0.09578	0.5733	0.08633	0.5700	0.08251
2	2	2	0.5578	0.09596	0.5560	0.08635	0.5546	0.08273
3	1	1	0.2770	0.06780	0.2805	0.07192	0.2827	0.07131
3	1	2	0.2782	0.06574	0.2779	0.06929	0.2778	0.06855
3	2	1	0.2555	0.05987	0.2561	0.06416	0.2569	0.06410
3	2	2	0.4599	0.08735	0.4610	0.08331	0.4609	0.07965

specification of the linear predictor as well as the distribution of the response variable are the most important elements for obtaining a good fit.

6.7 Percentages

In this section, we consider proportions that have been calculated from discrete counts, for example, the number of infected plants in treatment i of total N_i plants that are likely to have a binomial distribution. This class of models allows the response to arise from different distributions and probabilities.

6.7.1 RCBD: Dead Aphid Rate

An experiment was designed to study the effect of conidial density on the transmission of a fungus that attacks aphids. Aphid carcasses killed by the fungus, and from which the fungus released spores, were placed on bean plants at three densities ($A = 1$, $B = 5$, or $C = 10$ carcasses per plant) to provide different doses of fungal conidia. Densities were assigned to individual bean plants in a completely randomized design with six replicates. A total of 20 live uninfected (N) aphids were placed on each plant with a ladybug that was allowed to forage (feed on the bean plants) to facilitate the transfer of conidia between the carcasses and the live aphids. For each plant, the number of aphids infected with the fungus was counted (n_{ij}) and the proportion of aphids infected with the fungus was calculated 7 days after the

Table 6.29 Proportion of infested aphids

Plant	Density	p_{ij}
1	C	0.34299
2	A	0.16659
3	B	0.47004
4	C	0.62481
5	B	0.21926
6	B	0.16659
7	C	0.47502
8	C	0.52747
9	A	0.41581
10	B	0.42556
11	A	0.19466
12	A	0.34299
13	C	0.677
14	C	0.76674
15	A	0.13124
16	B	0.58419
17	B	0.38225
18	A	0.28905

Table 6.30 Sources of variation and degrees of freedom

Sources of variation	Degrees of freedom
Trt	$t - 1 = 2$
Error	$t(r - 1) = 15$
Total	$t \times r - 1 = 17$

inoculum was placed. The results shown below correspond to the proportion of infected aphids calculated at each of the inoculum concentrations ($p_{ij} = n_{ij}/N$; $N = 20$) to each of the conidial concentrations (density) tested (Table 6.29).

The sources of variation and degrees of freedom for this experiment are shown in Table 6.30.

The components of the GLMM having a beta response are listed below:

Distributions: $p_{ij} \mid \text{density(plant)}_{i(j)} \sim \text{Beta}(\pi_{ij}, \phi)$

$\text{density(plant)}_{i(j)} \sim N\left(0, \sigma_{\text{density(plant)}}^2\right)$

Linear predictor: $\eta_{ij} = \mu + \text{density}_i + \text{density(plant)}_{i(j)}$; $i = 1, 2, 3$; $j = 1, \dots, 6$

Link function: $\log\left(\frac{\pi_{ij}}{1 - \pi_{ij}}\right) = \text{logit}(\pi_{ij}) = \eta_{ij}$

The following GLIMMIX program fits a GLMM in a completely randomized design with a beta distribution. Here, density is conc_ino.

```
proc glimmix data=thumbs nobound method=laplace;
class plant conc_ino;
model p = conc_ino /dist=beta link=logit;
random conc_ino(plant);
```

Table 6.31 Results of the analysis of variance

(a) Fit statistics for conditional distribution				
-2 Log L (P r. effects)			-24.13	
Pearson's chi-square			18.45	
Pearson's chi-square/DF			1.02	
(b) Covariance parameter estimates				
Cov Parm	Estimate	Standard error		
Conc_Ino (Planta)	-0.1833	.		
Scale	12.9999	4.1954		
(c) Type III tests of fixed effects				
Effect	Num DF	Den DF	F-value	Pr > F
Conc_Ino	2	15	8.25	0.0038

Table 6.32 Means and standard errors on the model scale and the data scale

Conc_Ino least squares means							
Conc_Ino	Estimate	Standard error	DF	t-value	Pr > t	Mean	Standard error mean
A	-1.0340	0.2438	15	-4.24	0.0007	0.2623	0.04717
B	-0.5282	0.2246	15	-2.35	0.0328	0.3709	0.05241
C	0.2775	0.2197	15	1.26	0.2259	0.5689	0.05388

```
lsmeans conc_ino/lines ilink;
run;
```

Part of the results is shown in Table 6.31. The value of the conditional fit statistic in part (a), Pearson's chi - square/DF = 1.02, indicates that there is no overdispersion in the data and that the beta distribution is a good model for this dataset. The estimated variance of the plants' nested inoculum density is $\hat{\sigma}_{\text{density(plant)}}^2 = -0.1833$ and the estimated scale parameter is $\hat{\phi} = 12.999$; both are tabulated in part (b). In part (c) of the same table, the type III tests of fixed effects are shown, indicating that the density (concentration) of the inoculum has a significant effect ($P = 0.0038$) on the proportion of infested aphids with the fungus.

The values under the column "Estimates" are estimated mean proportions on the model scale, whereas the column "Mean" shows the estimated mean proportions on the data scale with their respective standard errors (Table 6.32). These estimates were obtained with the "lsmeans" and "ilink" option.

Figure 6.8 shows a linear trend in the proportion of aphids infested as conidial density increases. Conidia densities A and B showed statistically equal proportions of infested aphids compared to density C. Finally, the highest proportion of infested aphids was observed at density C.

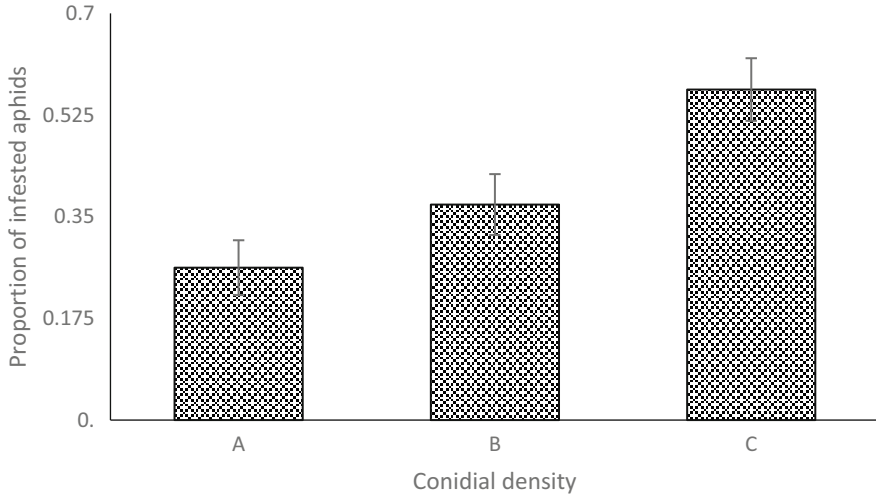


Fig. 6.8 Proportion of aphids infected at different conidia concentration densities

6.7.2 RCBD: Percentage of Quality Malt

An agro-industrial engineer is interested in studying the effect of germination time in minutes (48, 96, and 144) on the percentage of quality malt obtained from six sorghum varieties (sorghum bicolor): Gambella 1107, Macia, Meko, Red Swazi, Teshale, and 76T1#23 (Bekele et al. 2012). The percentage of quality malt (y) as a function of both factors is shown in Table 6.33.

For this purpose, an RCBD was implemented with a treatment factorial structure (variety \times germination time). The statistical model to analyze the dataset is the following:

Distributions: $y_{ijk} | r_k \sim \text{Beta}(\pi_{ijk}, \phi)$; $i = 1, \dots, 6$; $j, k = 1, 2, 3$

$r_k \sim N(0, \sigma_{\text{block}}^2)$, where y_{ijk} is the k th percentage of malt quality observed at the i th variety with the j th fermentation time.

Linear predictor: $\eta_{ijk} = \mu + r_k + \alpha_i + \beta_j + (\alpha\beta)_{ij}$, where μ is the overall mean, α_i is the fixed effect due to variety i , β_j is the fixed effect due to germination time j , and $(\alpha\beta)_{ij}$ is the interaction effect between variety and germination time.

Link function: $\text{logit}(\pi_{ijk}) = \eta_{ijk}$

Table 6.34 shows the sources of variation and degrees of freedom for this experiment.

The following GLIMMIX commands adjust a GLMM with a beta response.

```
proc glimmix data=malting nobound method=laplace;
class var_sorghum ger_time block;
model p = var_sorghum|ger_time/dist=beta link=logit;
random block;
lsmeans var_sorghum|ger_time/lines ilink;
run;
```

Table 6.33 Percentage of quality malt as a function of both factors (variety and germination time)

Variety	Time	Block	y	Variety	Time	Block	y
Gambella	T1	1	7.25	Red Swazi	T2	1	21
Gambella	T1	2	11.16	Red Swazi	T2	2	15.09
Gambella	T1	3	15.9	Red Swazi	T2	3	24.84
Macia	T1	1	10.91	Teshale	T2	1	25.42
Macia	T1	2	8.75	Teshale	T2	2	26.86
Macia	T1	3	10.87	Teshale	T2	3	26.64
Meko	T1	1	24.65	76 T1#23	T2	1	23.69
Meko	T1	2	23.63	76 T1#23	T2	2	20.71
Meko	T1	3	28.75	76 T1#23	T2	3	26.14
Red Swazi	T1	1	20.95	Gambella	T3	1	12.45
Red Swazi	T1	2	15.82	Gambella	T3	2	15.34
Red Swazi	T1	3	25.24	Gambella	T3	3	17.32
Teshale	T1	1	25.92	Macia	T3	1	8.51
Teshale	T1	2	27.64	Macia	T3	2	8.15
Teshale	T1	3	28.03	Macia	T3	3	13.07
76T1#23	T1	1	23.39	Meko	T3	1	22.09
76T1#23	T1	2	19.43	Meko	T3	2	24.11
76T1#23	T1	3	25.55	Meko	T3	3	24.47
Gambella	T2	1	10.03	Red Swazi	T3	1	20.81
Gambella	T2	2	12.9	Red Swazi	T3	2	16.05
Gambella	T2	3	17.84	Red Swazi	T3	3	23.7
Macia	T2	1	7.88	Teshale	T3	1	26.42
Macia	T2	2	9.14	Teshale	T3	2	27.07
Macia	T2	3	11.99	Teshale	T3	3	28.01
Meko	T2	1	22.97	76 T1#23	T3	1	24.18
Meko	T2	2	25.37	76 T1#23	T3	2	19.58
Meko	T2	3	25.71	76 T1#23	T3	3	25.74

Table 6.34 Sources of variation and degrees of freedom

Sources of variation	Degrees of freedom
Blocks	$r - 1 = 3 - 1 = 2$
Variety	$a - 1 = 6 - 1 = 5$
Time_Germination	$b - 1 = 3 - 1 = 2$
Variety*germ_time	$(a - 1)(b - 1) = 10$
Error	$(ab - 1)(r - 1) = 17 \times 2 = 34$
Total	$r \times a \times b - 1 = 54 - 1 = 53$

Part of the results of the above program is shown in Table 6.35. In part (a), the value of Pearson’s chi-square/DF is tabulated ($\frac{\chi^2}{df} = 0.92$), which indicates that the beta distribution is a good distribution for modeling malt percentage since the t -value of Pearson’s chi-square/DF is close to 1. The estimated variance due to blocks is

Table 6.35 Results of the analysis of variance of the RCBD with a beta distribution

(a) Fit statistics for conditional distribution				
-2 Log L (p r. effects)			-280.89	
Pearson's chi-square			49.66	
Pearson's chi-square/DF			0.92	
(b) Covariance parameter estimates				
Cov Parm	Estimate	Standard error		
Block	0.01210	0.01055		
Scale	431.54	85.4922		
(c) Type III tests of fixed effects				
Effect	Num DF	Den DF	F-value	Pr > F
Var_sorghum	5	34	106.51	<0.0001
Ger_time	2	34	0.26	0.7722
Var_sorghum*ger_time	10	34	1.08	0.4041

Table 6.36 Means and standard errors on the model scale and the data scale for sorghum varieties

Var_sorghum least squares means							
Var_sorghum	Estimate	Standard error	DF	t-value	Pr > t	Mean	Standard error mean
76 T1#23	-1.2011	0.07401	34	-16.23	<0.0001	0.2313	0.01316
Gambella	-1.8898	0.07929	34	-23.83	<0.0001	0.1313	0.009042
Macia	-2.2067	0.08295	34	-26.60	<0.0001	0.09915	0.007409
Meko	-1.1201	0.07364	34	-15.21	<0.0001	0.2460	0.01366
Red Swazi	-1.3685	0.07493	34	-18.26	<0.0001	0.2029	0.01212
Teshale	-1.0025	0.07314	34	-13.71	<0.0001	0.2685	0.01436

$\hat{\sigma}_{\text{block}}^2 = 0.012$ and the estimated scale parameter is $\hat{\phi} = 431$ (part (b)), whereas the type III fixed effects hypothesis tests in part (c) show that sorghum variety has a significant effect on malt quality percentage ($P = 0.0001$).

The least squares means on the model scale and the data scale for the factor variety are listed under the columns “Estimate” and “Mean” with their respective standard errors “Standard error” in Table 6.36.

Figure 6.9 shows that Teshale produced the highest average malt percentage (0.2685 ± 0.01436), followed by the varieties 76 T1#23 and Meko ($0.2313 \pm 0.01316, 0.246 \pm 0.01366$), whereas the variety Macia produced the lowest malt percentage (0.09915 ± 0.0074).

6.7.3 A Split Plot in an RCBD: Cockroach Mortality (Blattella germanica)

An entomologist is interested in testing six isolates of insect pathogenic fungi: five obtained from different hosts and one already known isolate (Control) of a fungus

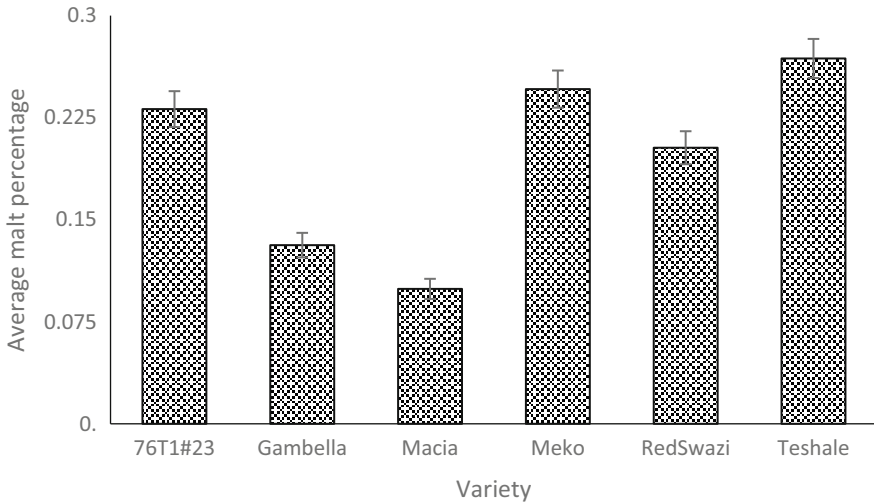


Fig. 6.9 Percentage of quality malt of bicolor sorghum varieties

Table 6.37 Analysis of variance with sources of variation and degrees of freedom for this experiment

Sources of variation	Degrees of freedom
Blocks	$r - s1 = 2 - 1 = 1$
Isolation	$a - 1 = 6 - 1 = 5$
Block (insulation)	$a(r - 1) = 6$
Age	$b - 1 = 3 - 1 = 2$
Isolation*age	$(a - 1)(b - 1) = 5 \times 2 = 10$
Error	$(a - 1)(b - 1)(r - 1) = 2 \times 5 \times 1 = 10$
Total	$r \times a \times b - 1 = 2 \times 6 \times 3 - 1 = 35$

with potential for biological control of a particular species of cockroaches. To do so, the entomologist decides to test these fungal isolates on three different insect ages (age1 = E1, age2 = E2, and age3 = E3). Each of the isolates was placed in a Petri dish with 10 insects of a specific age. Each set (isolate–age) was randomly assigned to two blocks (Appendix: Data: Cockroaches).

The analysis of variance table (Table 6.37) with the sources of variation and degrees of freedom for this experiment is presented below. The response variable (percentage mortality) for this experiment is assumed to have a beta distribution.

The components that describe the model of this experiment are listed below:

Distributions: $y_{ijk} \mid r_k, r(\alpha)_{k(i)} \sim \text{Beta}(\pi_{ijk}, \phi)$; $i = 1, \dots, 6$; $j = 1, 2, 3$; $k = 1, 2$.

$$r_k \sim N(0, \sigma_r^2), r(\alpha)_{k(i)} \sim N(0, \sigma_{r(\alpha)}^2)$$

Linear predictor: $\eta_{ijk} = \mu + r_k + \alpha_i + r(\alpha)_{k(i)} + \beta_j + (\alpha\beta)_{ij}$

Link function: $\text{logit}(\pi_{ijk}) = \eta_{ijk}$

Table 6.38 Results of the analysis of variance of the RCBD with a factorial structure in treatments

(a) Fit statistics for conditional distribution				
-2 Log L (y r. effects)			-74.53	
Pearson's chi-square			34.02	
Pearson's chi-square/DF			1.00	
(b) Covariance parameter estimates				
Cov Parm	Subject	Estimate	Standard error	
Aislamiento	Block	-0.03125	.	
Scale		24.1882	5.7925	
(c) Type III tests of fixed effects				
Effect	Num DF	Den DF	F-value	Pr > F
Isolation	5	6	16.48	0.0019
Age	2	10	30.01	<0.0001
Isolation*age	10	10	4.83	0.0102

The following GLIMMIX commands adjust a GLMM with a beta response.

```
proc glimmix nobound method=laplace;
class block Isolation Age;
model y = Isolation|Age/dist=beta link=logit;
random Isolation/subject=block;
lsmeans Insulation|Age/slice=Insulation lines ilink;
run;
```

Some of the outputs are listed below (Table 6.38). The conditional statistic Pearson's chi - square/DF = 1 indicates that the distribution used is appropriate for these datasets (part (a)). The variance component estimates are tabulated in part (b), and, for blocks, the estimate is $\hat{\sigma}_r^2 = -0.03125$ and the estimated scale parameter is $\hat{\phi} = 24.1882$. The hypothesis test is in part (c) with type III fixed effects of equality of means for type of isolation, age of the insect, and the interaction between both factors. These outputs indicate that they have a significant effect on insect mortality.

We see the expected proportions with their respective standard errors of both factors on the data scale under the "Mean" column (Tables 6.39 and 6.40). These values arise by applying the inverse link to estimates under "Estimate" on the model scale. Table 6.39 shows the estimated average mortality probabilities for the isolates; for example, for isolate A1, applying the inverse link to the linear predictor estimate $\hat{\eta}_{1.} = 0.1722$ we get $\hat{\pi}_{1.} = 1/1 + e^{-0.1722} = 0.5429$. In this manner, we see that the expected proportions for isolates 2 and 4 are $\hat{\pi}_{2.} = 0.6555$ and $\hat{\pi}_{4.} = 0.5762$, respectively, whereas for the control $\hat{\pi}_{\text{control}} = 0.1157$.

Regarding the age of the insect (Table 6.40), the expected average probability of mortality was higher at age three (adults) with a higher mortality rate $\hat{\pi}_{.3} = 0.6435$, whereas insects at age two (E2) had a higher resistance to the isolations, showing a mortality of $\hat{\pi}_{.2} = 0.2598$.

In general, fungal isolates A1, A2, A3, and A4 showed an average mortality of more than 75% for adult insects (E3), whereas isolates A1, A2, and A5 showed a

Table 6.39 Means and standard errors on the model scale and the data scale for isolation

Isolate least squares means							
Isolate	Estimate	Standard error	DF	<i>t</i> -value	Pr > <i>t</i>	Mean	Standard error mean
A1	0.1722	0.1859	6	0.93	0.3900	0.5429	0.04614
A2	0.6442	0.2100	6	3.07	0.0220	0.6557	0.04740
A3	-0.1489	0.1952	6	-0.76	0.4746	0.4629	0.04853
A4	0.3073	0.2088	6	1.47	0.1915	0.5762	0.05098
A5	-0.2023	0.1806	6	-1.12	0.3053	0.4496	0.04468
Control	-2.0339	0.2418	6	-8.41	0.0002	0.1157	0.02473

Table 6.40 Means and standard errors on the model scale and the data scale for insect age

Age least squares means							
Age	Estimate	Standard error	DF	<i>t</i> -value	Pr > <i>t</i>	Mean	Standard error mean
E1	-0.1747	0.1310		-1.33	0.2120	0.4564	0.03251
E2	-1.0468	0.1374		-7.62	<0.0001	0.2598	0.02643
E3	0.5908	0.1634		3.61	0.0047	0.6435	0.03749

mortality rate of around 65% for cockroaches of age E1 (juvenile insects). On the other hand, all isolates showed lower lethal effectiveness on insects of age E2 (Fig. 6.10).

6.7.4 A Split-Plot Design in an RCBD: Percentage Disease Inhibition

A plant pathologist wishes to compare the response of two plant varieties to different doses/amounts of a pesticide formulated to protect plants against a disease. Five racks (blocks) were chosen to account for local variation within the greenhouse. Each rack was divided into four sections or rooms and were randomly assigned one of four pesticide levels to each rack. The four pesticide levels were 1, 2, 4, and 8 mg/L. One plant of each variety was placed in each section of the rack. Of the two plant varieties, one variety was susceptible, labeled S, and the other variety was resistant, labeled R (Table 6.41). The response variable (*y*) is the percentage of disease inhibition in the plant.

The sources of variation and degrees of freedom for this experiment are shown in Table 6.42.

Following the same reasoning used in the examples above, the components of the GLMM with a beta response that models the observed disease inhibition proportion (p_{ijk}) under dose *i* with variety *j* in block *k* are listed as follows:.

Distributions: $y_{ijk} | r_k, (\alpha)_{ik} \sim \text{Beta}(\pi_{ijk}, \phi)$; $i = 1, \dots, 4$; $j = 1, 2$; $k = 1, \dots, 5$
 $r_k \sim N(0, \sigma_r^2)$, $(\alpha)_{ik} \sim N(0, \sigma_{rA}^2)$

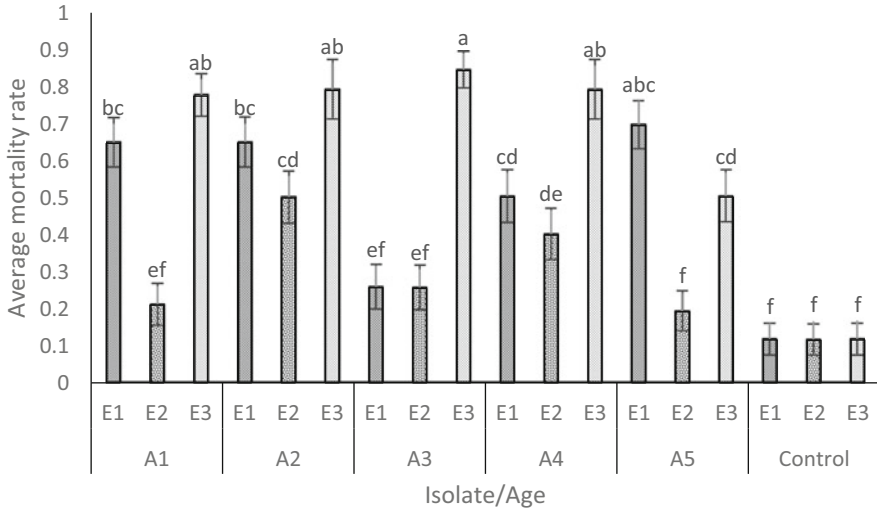


Fig. 6.10 Cockroach mortality percentage

Table 6.41 Percentage of inhibition

Block	Variety	Dose	y	Block	Variety	Dose	y
1	R	1	15.7	1	S	1	19.8
2	R	1	23.1	2	S	1	17.8
3	R	1	15.9	3	S	1	13.2
4	R	1	20.8	4	S	1	14.8
5	R	1	24.5	5	S	1	19.7
1	R	2	25.1	1	S	2	21.2
2	R	2	29.2	2	S	2	29.3
3	R	2	29.7	3	S	2	26
4	R	2	28.6	4	S	2	27.5
5	R	2	26.6	5	S	2	22
1	R	4	27.9	1	S	4	29.3
2	R	4	29.7	2	S	4	27.2
3	R	4	24	3	S	4	26
4	R	4	29.7	4	S	4	31.5
5	R	4	29.6	5	S	4	27.9
1	R	8	23.8	1	S	8	22.8
2	R	8	31.2	2	S	8	33
3	R	8	21.8	3	S	8	25.2
4	R	8	23.3	4	S	8	27.2
5	R	8	23.9	5	S	8	20.8

Table 6.42 Sources of variation and degrees of freedom

Sources of variation	Degrees of freedom
Blocks	$r - 1 = 5 - 1 = 4$
Dose	$a - 1 = 4 - 1 = 3$
Error _a (Bloque*Dose)	$(r - 1)(a - 1) = 12$
Variety	$b - 1 = 2 - 1 = 1$
Dose*variety	$(a - 1)(b - 1) = 3$
Error _b	$a(b - 1)(r - 1) = 4 \times 1 \times 4 = 16$
Total	$r \times a \times b - 1 = 5 \times 4 \times 2 - 1 = 39$

Linear predictor: $\eta_{ijk} = \mu + r_k + \alpha_i + (r\alpha)_{ik} + \beta_j + (\alpha\beta)_{ij}$, where r_k is the random block effect, α_i is the fixed dose effect, β_j is the fixed variety effect, $(r\alpha)_{ik}$ is the random effect due to block by dose interaction, and $(\alpha\beta)_{ij}$ is the interaction of fixed effects due to dose variety.

Link function: $\text{logit}(\pi_{ijk}) = \eta_{ijk}$

The following GLIMMIX commands adjust a GLMM.

```
proc glimmix nobound method=laplace;
class Variety dose block;
model y = dose variety dose*variety /dist=beta link=logit;
random Block Block*dose;
contrast 'Linear dose' dose -3 -1 1 3;
contrast 'Quadratic dose' dose 1 -1 -1 -1 1;
contrast 'dose Cubic' dose -1 3 -3 1;
lsmeans variety|dose / slice=(variety dose) lines ilink;
ods output lsmeans=dose_means;
run;
```

The “contrast” command in the program can perform a hypothesis testing to see what trend (linear, quadratic, or cubic) the “dose” factor has on the percentage of disease inhibition. Part of the output is shown in Table 6.43. The value of the conditional goodness-of-fit statistic Pearson’s chi – square/DF = 0.59 indicates that we have no evidence of overdispersion, and, therefore, the beta distribution is adequate to model this dataset (part (a)). The variance component estimates in part (b) for block and block \times dose are $\hat{\sigma}_r^2 = 0.004898$ and $\hat{\sigma}_{r \cdot \text{dose}}^2 = 0.002372$, respectively. Finally, the F -value provides sufficient statistical evidence of the effect of dose on disease decline in plants ($P = 0.0001$), whereas the effect of variety and dose \times variety do not provide sufficient evidence.

Table 6.44 shows the polynomial contrasts for the effect of “dose,” which indicate that there is a significant quadratic effect on the percentage of disease inhibition.

The inhibition percentage has almost a linear trend as the dose increases from 1 to 4 ml/L in both varieties, but when the dose is higher than 4 ml/L, the inhibition of the disease decreases in both varieties (Fig. 6.11).

Table 6.43 Results of the analysis of variance

(a) Fit statistics for conditional distribution				
-2 Log L (y r. effects)			-184.32	
Pearson's chi-square			23.63	
Pearson's chi-square/DF			0.59	
(b) Covariance parameter estimates				
Cov Parm	Estimate	Standard error		
Block	0.004898	.		
Block*dose	0.002372	0.007513		
Scale	205.52	67.7447		
(c) Type III tests of fixed effects				
Effect	Num DF	Den DF	F-value	Pr > F
Dose	3	12	17.67	0.0001
Variety	1	16	1.74	0.2057
Dose*variety	3	16	1.22	0.3337

Table 6.44 Polynomial contrasts

Contrasts				
Label	Num DF	Den DF	F-value	Pr > F
Linear dose	1	12	25.48	0.0003
Quadratic dose	1	12	30.93	0.0001
Cubic dose	1	12	0.30	0.5948

6.7.5 Randomized Complete Block Design with a Binomial Response with Multiple Variance Components

The dataset corresponds to an experiment implemented by Madden and Hughes (1995) on the incidence of the disease caused by the fungus *Plasmopara viticola* on grape plants (*Vitis labrusca*). Six different treatments in a randomized block design ($b = 3$) were tested, where treatment 1 was the control, to study the disease with three grape plants ($v = 3$). On a single date in autumn, five sprouts were ($r = 5$) randomly selected from each of the three grape plants and the number of leaves with at least one mildew lesion was counted (m) out of a total n leaves. The number of leaves per shoot ranged from 7 to 21. The data for this experiment can be found in the Appendix (Data: Disease incidence on grape plants).

The statistical model that could describe the incidence of disease in this experiment, if the response variable p_{ijkl} were treated as a normal variable, would be as described below:

$$p_{ijkl} = \eta + \tau_i + b_j + (bv)_{jk} + (bvr)_{jkl} + \varepsilon_{ijkl}$$

$$i = 1, 2, \dots, 6; j = 1, 2, 3; k = 1, 2, \dots, 3; l = 1, 2, \dots, 5$$

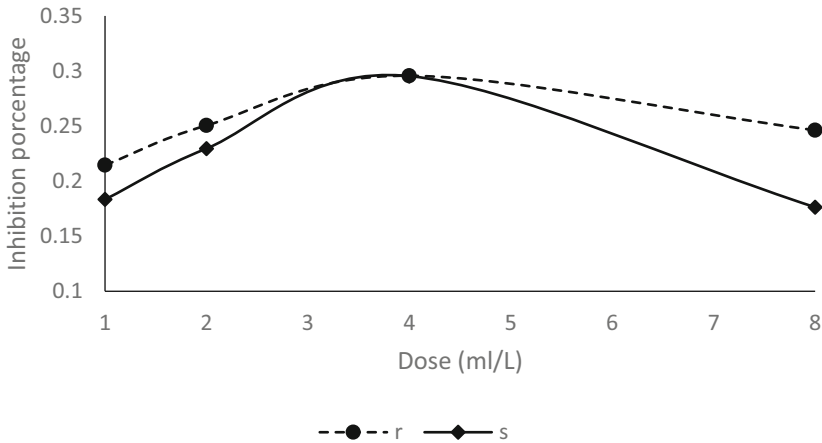


Fig. 6.11 Percentage of disease inhibition in both varieties

where p_{ijkl} is the $ijkl$ proportion of diseased leaves, η is the intercept, τ_i is the fixed treatment effect i , b_j is the random effect of blocks assuming $b_j \sim N(0, \sigma_{\text{block}}^2)$, $(\text{bv})_{jk}$ is the block–plant random effect assuming $(\text{bv})_{jk} \sim N(0, \sigma_{\text{block} \times \text{plant}}^2)$, $(\text{bvr})_{jkl}$ is the random effect due to block–plant–sprouts assuming $(\text{bvr})_{jkl} \sim N(0, \sigma_{\text{block} \times \text{plant} \times \text{sprout}}^2)$, and ε_{ijkl} is the experimental error assuming $\varepsilon_{ijkl} \sim N(0, \sigma^2)$.

For the disease incidence data, the assumption of a normal distribution for p_{ijkl} is not recommended. A good starting point for the analysis is to assume that the observed number of diseased leaves in the sprouts (y_{ijkl}) follows a binomial distribution with parameter π_{ijkl} and n_{ijkl} , the total number of leaves on the sprout.

Therefore, the components of the GLMM with a binomial distribution in the response variable are as follows:

Distribution: $p_{ijkl} \mid b_j, (\text{bv})_{jk}, (\text{bvr})_{jkl} \sim \text{binomial}(\pi_{ijkl}, n_{ijkl})$

$b_j \sim N(0, \sigma_{\text{block}}^2), (\text{bv})_{jk} \sim N(0, \sigma_{\text{block} \times \text{plant}}^2), (\text{bvr})_{jkl} \sim N(0, \sigma_{\text{block} \times \text{plant} \times \text{sprout}}^2)$

Linear predictor: $\eta_{ijkl} = \eta + \tau_i + b_j + (\text{bv})_{jk} + (\text{bvr})_{jkl}$.

Link function: $\text{logit}(\pi_{ijkl}) = \eta_{ijkl}$

The following GLIMMIX syntax fits a GLMM with a binomial response.

```
proc glimmix method=laplace nobound;
class v r b t;
model m/n = t /dist=bin;
random intercept v*v*r/subject=b;
lsmeans t/lines ilink;
run;
```

Table 6.45 Results of the analysis of variance under the binomial distribution

(a) Fit statistics				
-2 Log likelihood				723.17
AIC (smaller is better)				741.17
AICC (smaller is better)				741.87
BIC (smaller is better)				733.06
CAIC (smaller is better)				742.06
HQIC (smaller is better)				724.87
(b) Fit statistics for conditional distribution				
-2 Log L (m r. effects)				665.02
Pearson's chi-square				398.21
Pearson's chi-square/DF				1.47
(c) Covariance parameter estimates				
Cov Parm	Subject	Estimate	Standard error	
Intercept	b	-0.00408	.	
V	b	0.01917	.	
v*r	b	0.1960	.	
(d) Type III tests of fixed effects				
Effect	Num DF	Den DF	F-value	Pr > F
t	2	220	1837.99	<0.0001

Part of the results based on the aforementioned model is shown in Table 6.45. By default, proc GLIMMIX provides the fit statistics useful for selecting the best model from a group of models (part (a)).

In addition to accuracy considerations, the Laplace (or quadrature) analysis allows us to obtain the “conditional distribution fit statistics,” specifically Pearson’s χ^2/df . Recall that this statistic helps assess the goodness of fit of the model. If the value of $\chi^2/df \gg 1$ is an indicator that there is overdispersion in the dataset, then this may be because the linear predictor is incomplete or the assumed distribution is not suitable (mis-specified) for this dataset. In part (b), we can see that the value of the conditional distribution statistic of Pearson’s $\chi^2/df = 1.47$. This value indicates that we have evidence of overdispersion. The variance component estimates due to block, block \times plant, and block \times plant \times sprout are tabulated in part (c), whereas the type III tests of fixed effects (part (d)) indicate that there is a significant difference ($P < 0.0001$) between treatments.

Since there is overdispersion in the data in the binomial model, an alternative distribution is the beta distribution. The components of the GLMM are as follows:

Distribution: $p_{ijkl} \mid b_j, (bv)_{jk}, (bvr)_{jkl} \sim \text{beta}(\pi_{ijkl}, \phi)$;
 $b_j \sim N(0, \sigma_{\text{block}}^2), (bv)_{jk} \sim N(0, \sigma_{\text{block} \times \text{plant}}^2), (bvr)_{jkl} \sim N(0, \sigma_{\text{block} \times \text{plant} \times \text{sprout}}^2)$
 Linear predictor: $\eta_{ijkl} = \eta + \tau_i + b_j + (bv)_{jk} + (bvr)_{jkl}$
 Link function: $\text{logit}(\pi_{ijkl}) = \eta_{ijkl}$.

Table 6.46 Results of the analysis of variance under the beta distribution

(a) Fit statistics				
-2 Log likelihood				-231.10
AIC (smaller is better)				-211.10
AICC (smaller is better)				-209.30
BIC (smaller is better)				-220.11
CAIC (smaller is better)				-210.11
HQIC (smaller is better)				-229.22
(b) Fit statistics for conditional distribution				
-2 Log L (m r. effects)				-231.10
Pearson's chi-square				136.55
Pearson's chi-square/DF				1.07
(c) Covariance parameter estimates				
Cov Parm	Subject	Estimate	Standard error	
Intercept	b	-0.6308	.	
V	b	-0.2215	.	
v*r	b	-0.1843	.	
Scale ($\hat{\phi}$)		9.8397	1.1926	
(d) Type III tests of fixed effects				
Effect	Num DF	Den DF	F-value	Pr > F
t	2	220	1837.99	<0.0001

The following SAS commands adjust an GLMM under a beta distribution.

```
proc GLIMMIX method=laplace nobound;
class v r b t;
model pct = t /dist=beta link=logit;
random intercept v v*r/subject=b;
lsmeans t/lines ilink;
run;
```

Some of the outputs are shown below. Table 6.46 shows that the values of the fit statistics, as well as the conditional distribution statistics (parts (a) and (b)), are much smaller than when the binomial distribution was used.

This indicates that the beta distribution is more appropriate for the dataset, as the value of Pearson's statistic is $\chi^2/df = 1.03$, indicating that the problem of overdispersion was almost totally controlled. The variance component estimates as well as the estimated scale parameter ($\hat{\phi}$) are tabulated in part (c). Similar to the previous analysis, the type III tests of fixed effects indicate that there is a highly significant difference (part (d)) in treatments on the average proportion of leaves with fungal disease.

The least mean squares (means) on the model scale (column "Estimate") and on the data scale (column "Mean") are tabulated in Table 6.47. The results indicate that

Table 6.47 Estimated means (least squares means) on the model scale and on the data scale

Least squares means							
t	Estimate	Standard error	DF	t -value	Pr > t	Mean	Standard error mean
1	0.7223	0.09989	83	7.23	<0.0001	0.6731	0.02198
2	-1.7482	0.1543	83	-11.33	<0.0001	0.1483	0.01949
3	-2.0178	0.2214	83	-9.11	<0.0001	0.1174	0.02294
4	-1.9358	0.1873	83	-10.34	<0.0001	0.1261	0.02064
5	-1.7887	0.2173	83	-8.23	<0.0001	0.1432	0.02667
6	-1.5360	0.1665	83	-9.23	<0.0001	0.1771	0.02427

Table 6.48 Mean comparison (LSD method)

T grouping of t least squares means ($\alpha = 0.05$)		
LS means with the same letter are not significantly different		
t	Estimate	
1	0.7223	A
6	-1.5360	B
		B
2	-1.7482	B
		B
5	-1.7887	B
		B
4	-1.9358	B
		B
3	-2.0178	B

all proposed treatments in this study reduce the proportion of diseased leaves compared to the control treatment ($t = 1$).

The mean comparison (LSD) obtained with the option “lines” indicates that the proportion of diseased leaves in treatment one is statistically different from the rest of the treatments (Table 6.48).

6.8 Exercises

Exercise 6.8.1 Seeds of a particular crop were stored at four different temperatures (T_1, T_2, T_3 , and T_4) under four different chemical concentrations (0, 0.1, 1.0, and 10). To study the effects of temperature and chemical concentration, a completely randomized experiment was conducted with a factorial treatment structure 4×4 and four replicates. For each of the 64 experimental units, 50 seeds were placed in a dish and the number of seeds that germinated under standard conditions was recorded. Germination data were obtained from Mead et al. (1993, p. 325) (Table 6.49).

Table 6.49 Seed germination experiment results

Temperature	Chemical concentration			
	0	0.1	1.0	
T_1	9, 9, 3, 7	13, 12, 14, 15	21, 23, 24, 27	40, 32, 43, 34
T_2	19, 30, 21, 29	33, 32, 30, 26	43, 40, 37, 41	48, 48, 49, 48
T_3	7, 7, 2, 5	1, 2, 4, 4	8, 10, 6, 7	3, 4, 8, 5
T_4	4, 9, 3, 7	13, 6, 15, 7	16, 13, 18, 19	13, 18, 11, 16

Table 6.50 Results of the apple sprouts experiment

Density of inoculum	Cultivate	Block 1	Block 2	Block 3	Block 4
200	Jonagold	5/1	5/2	5/1	5/0
200	Golden delicious	5/1	5/0	5/0	5/0
200	Jonathan	5/2	5/2	5/2	5/0
1000	Jonagold	5/0	5/2	5/2	5/4
1000	Golden delicious	5/0	5/0	5/2	5/0
1000	Jonathan	5/4	5/4	5/4	5/0
5000	Jonagold	5/5	5/5	5/4	5/5
5000	Golden delicious	5/5	5/4	5/3	5/5
5000	Jonathan	5/5	5/0	5/3	5/5

The first number refers to the number of inoculations (n) and the second to the number of inoculations that developed the gangrenous sore (Y)

- Write down an ANOVA table (sources of variation, degrees of freedom) for this experiment.
- List all the components of the GLMM in (a).
- Analyze this dataset and summarize the relevant results.

Exercise 6.8.2 Data were obtained from an experiment in which separate sprouts of apple trees were inoculated with macroconidia of the fungus *Nectria galligena*, which causes apple cancer (canker gangrene). The experimental factors were inoculum density (three levels: 200, 1000, and 5000 macroconidia per ml) and variety (three levels: Jonagold, Golden Delicious, and Jonathan). The experiment was carried out in 4 randomized blocks with 12 plots. Each plot consisted of one sprout on which five inoculations were made. The numbers of successful inoculations per plot on day 17 after inoculation are shown in the table below (Table 6.50).

- Write down an ANOVA table (sources of variation, degrees of freedom) for this experiment.
- List all the components of the GLMM from part (a).
- Analyze this dataset and summarize the relevant results.
- Is there is an extra-variation in the dataset? What alternative distribution do you propose? Reanalyze the data and compare the results.

Exercise 6.8.3 This experiment concerns the germination efficiencies of protoplasts obtained from plants of seven species of the genera *Lycopersicon* (tomato) and

Table 6.51 Protoplast germination experiment results

Species	Isolation	1	2	3	4	5	6	7	8	9	10
1	1	8.9	6.3	10.5							
1	2	3.1	2.7	4.1							
1	3	2.1	1.9	1.4	1.5						
1	4	2.5	2.9	2.6	2.6	2.6	2.6	2.8	2.7	2.8	2.7
2	1	0.2	0.9	0.5	0.6	1.2	0.4				
2	2	1.8	1.6	1.6							
2	3	6.6	7.5	5.4	5.3	5	6.5	6.3	5.8	5.9	5.6
3	1	1.8	1.5	1.9	1.7	1.3	1.5				
3	2	1.5	3.2	1.1	1.3	1.8	1.2	1.6	1.4	1.2	1.8
3	3	2	2.3	2.8	2.6	3.2	2.2	2.5	2.4	2.8	2.4
4	1	11.4	11.3	14.4	13.7						
4	2	2.9	3.8	4.7	5.1	2.7	3.2				
4	3	2.3	4.4	4.8	4.9	5.8	4.7	5.6	4.2	3.3	4.5
5	1	21.5	25.5	18.1	22.2						
5	2	18.7	20								
5	3	11.5	13.1	11.5	16.2	10.1	17.2	16	10.5		
6	1	4.6	3.4	2.7	3	4.1	3.1				
6	2	2.4	2.4	2	2.5	3.6	3.2	2.6	1.4	2.5	2.7
6	3	1.6	1.1	1.6	1.3	1.6	1	0.8	1.3	0.8	2.2
7	1	3	4	4.1	4.4	2.8	3.3	4.5	3.3	3	3.2
7	2	2.5	2.5	2.5	2.7	2.3	2.6				
7	3	2.6	2.7	2.9	2.7	2.7	2.6				
7	4	2.9	3	3	3.1						

Solanum (potato). For each species, three or four protoplast isolates were used and, depending on the availability of the protoplasts, a variable number of plates was carried out. Per plate, approximately 105 protoplasts were placed in a Petri dish, and, after 4 weeks, the proportion of dividing protoplasts was recorded. The results in percentages are listed below (Table 6.51).

- (a) Write down an ANOVA table (sources of variation, degrees of freedom) for the experimental design of this study.
- (b) Write down a generalized linear mixed model base in (a), assuming a beta distribution on the response variable.
- (c) Implement an analysis of these data according to the linear predictor and model in part (b). Summarize the relevant results.

Exercise 6.8.4 The data in this example are the results of a triangle test for 12 raters tasting 10 pairs of coffee varieties (Table 6.52). The triangle test consisted of each rater drinking three cups, one of one variety and two of the other. Each rater had 12 triangles for each pair of varieties, 2 for each of the following sequences: AAB, ABA, BAA, ABB, BAB, and BBA. The answer is the correct variety identification number appearing once. The experiment was conducted in two groups of six

Table 6.52 Triangle test (G = group, Eval = panelist, PdV = variety pair, V_A = variety A; V_B = variety B; Y = number of correct discriminations, n = number of trials)

G	Eval	PdV	V_A	V_B	Y	n	G	Eval	PdV	V_A	V_B	Y	n
1	1	1	8	9	2	12	2	7	1	8	9	4	12
1	1	2	5	9	11	12	2	7	2	5	9	12	12
1	1	3	9	6	9	12	2	7	3	9	6	7	12
1	1	4	6	5	6	12	2	7	4	6	5	9	12
1	1	5	6	8	8	12	2	7	5	6	8	10	12
1	1	6	5	8	9	12	2	7	6	5	8	5	12
1	1	7	7	8	6	12	2	7	7	7	8	9	12
1	1	8	7	9	8	12	2	7	8	7	9	9	12
1	1	9	7	5	11	12	2	7	9	7	5	7	12
1	1	10	7	6	5	12	2	7	10	7	6	5	12
1	2	1	8	9	5	12	2	8	1	8	9	2	12
1	2	2	5	9	8	12	2	8	2	5	9	10	12
1	2	3	9	6	8	12	2	8	3	9	6	8	12
1	2	4	6	5	9	12	2	8	4	6	5	9	12
1	2	5	6	8	10	12	2	8	5	6	8	8	12
1	2	6	5	8	11	12	2	8	6	5	8	9	12
1	2	7	7	8	8	12	2	8	7	7	8	4	12
1	2	8	7	9	9	12	2	8	8	7	9	6	12
1	2	9	7	5	8	12	2	8	9	7	5	10	12
1	2	10	7	6	7	12	2	8	10	7	6	7	12
1	3	1	8	9	4	12	2	9	1	8	9	3	12
1	3	2	5	9	9	12	2	9	2	5	9	11	12
1	3	3	9	6	9	12	2	9	3	9	6	11	12
1	3	4	6	5	11	12	2	9	4	6	5	8	12
1	3	5	6	8	8	12	2	9	5	6	8	8	12
1	3	6	5	8	10	12	2	9	6	5	8	11	12
1	3	7	7	8	3	12	2	9	7	7	8	5	12
1	3	8	7	9	7	12	2	9	8	7	9	4	12
1	3	9	7	5	10	12	2	9	9	7	5	11	12
1	3	10	7	6	9	12	2	9	10	7	6	8	12
1	4	1	8	9	7	12	2	10	1	8	9	7	12
1	4	2	5	9	10	12	2	10	2	5	9	9	12
1	4	3	9	6	7	12	2	10	3	9	6	5	12
1	4	4	6	5	8	12	2	10	4	6	5	11	12
1	4	5	6	8	7	12	2	10	5	6	8	5	12
1	4	6	5	8	8	12	2	10	6	5	8	10	12
1	4	7	7	8	7	12	2	10	7	7	8	7	12
1	4	8	7	9	6	12	2	10	8	7	9	8	12
1	4	9	7	5	10	12	2	10	9	7	5	6	12
1	4	10	7	6	7	12	2	10	10	7	6	9	12
1	5	1	8	9	6	12	2	11	1	8	9	7	12
1	5	2	5	9	10	12	2	11	2	5	9	9	12

(continued)

Table 6.52 (continued)

G	Eval	PdV	V_A	V_B	Y	n	G	Eval	PdV	V_A	V_B	Y	n
1	5	3	9	6	4	12	2	11	3	9	6	6	12
1	5	4	6	5	8	12	2	11	4	6	5	10	12
1	5	5	6	8	6	12	2	11	5	6	8	5	12
1	5	6	5	8	7	12	2	11	6	5	8	10	12
1	5	7	7	8	8	12	2	11	7	7	8	8	12
1	5	8	7	9	9	12	2	11	8	7	9	6	12
1	5	9	7	5	9	12	2	11	9	7	5	9	12
1	5	10	7	6	8	12	2	11	10	7	6	9	12
1	6	1	8	9	3	12	2	12	1	8	9	6	12
1	6	2	5	9	9	12	2	12	2	5	9	7	12
1	6	3	9	6	6	12	2	12	3	9	6	7	12
1	6	4	6	5	9	12	2	12	4	6	5	7	12
1	6	5	6	8	7	12	2	12	5	6	8	8	12
1	6	6	5	8	10	12	2	12	6	5	8	11	12
1	6	7	7	8	7	12	2	12	7	7	8	9	12
1	6	8	7	9	7	12	2	12	8	7	9	9	12
1	6	9	7	5	8	12	2	12	9	7	5	10	12
1	6	10	7	6	9	12	2	12	10	7	6	9	12

evaluators, each with the aim of discriminating the abilities of the panelists for future evaluations. The data for this example are shown below:

- (a) Write down an ANOVA table (sources of variation, degrees of freedom) for this experiment.
- (b) List all the components of the GLMM according to part (a).
- (c) Analyze this dataset and summarize the relevant results.
- (d) Is there an extra-variation in the dataset? If so, what alternative distribution do you propose? Reanalyze the data and compare the results.

Exercise 6.8.5 Several brewing techniques are used in the production of espresso coffee. Among them, the most widespread are bar machines and single-dose pods, designed in large numbers due to their commercial popularity. This experiment tries to compare the foaming rate (Y , in percentage) effects of three different brewing techniques on espresso quality (method 1 = bar machine (BM), method 2 = hyper-espresso method (HIP), and method 3 = I-espresso system (IT)). Nine replicates per method were carried out (Table 6.53).

- (a) Write down an ANOVA table (sources of variation, degrees of freedom) for the experimental design of this study.
- (b) Describe the generalized linear mixed model in (a), assuming a beta distribution.
- (c) Implement the analysis of these data according to the predictor and model in (b). Summarize the relevant results.

Table 6.53 Experimental results of espresso coffee

Method	Index	Method	Index	Method	Index
1	36.64	2	70.84	3	56.19
1	39.65	2	46.68	3	36.67
1	37.74	2	73.19	3	35.35
1	35.96	2	57.78	3	40.11
1	38.52	2	48.61	3	33.52
1	21.02	2	72.77	3	37.12
1	24.81	2	65.04	3	37.33
1	34.18	2	62.53	3	32.68
1	23.08	2	54.26	3	48.33

Table 6.54 Results of wheat germination experiment in pots. Number of seeds that did not germinate out of 50

	Treatments						
	1	2	3	4	5	6	7
A	10	11	8	9	7	6	9
B	8	10	3	7	9	3	11
C	5	11	2	8	10	7	11
D	1	6	4	13	7	10	10

Exercise 6.8.6 The decision to adopt a particular scale for data involving small integers is not an easy one because any analysis must be – to some extent – as adequate as possible to obtain estimates with as little uncertainty as possible. As a simple example of this type of data, consider the following results from a potted wheat germination experiment (Table 6.54).

- (a) Write down an ANOVA table (sources of variation, degrees of freedom) for this experiment.
- (b) List all components of the GLMM in (a), assuming a binomial response variable.
- (c) Analyze this dataset and summarize the relevant results.
- (d) Is there an extra-variation in the dataset? If so, reanalyze the data with an alternative distribution. Summarize and compare your findings.

Exercise 6.8.7 A greenhouse experiment was carried out to investigate how a disease spreads in two varieties of (agurkesyge) cucumber, which is supposed to depend on the climate and the amount of fertilizers used for the two varieties. The following data come from the Department of Plant Pathology. Two climates were used: (1) change to day temperature 3 hours before sunrise and (2) normal change to day temperature. Three amounts of fertilizer were applied, normal (2.0 units), high (3.5 units), and very high (4.0 units). The two varieties were Aminex and Dalibor. To have a better controlled experiment, the plants were “standardized” to equally have as many leaves, and, then (on day 0, for example), the plants were contaminated with the disease. Subsequently, 8 days after the plants were contaminated, the amount of infection (in percentage) was recorded. From the resulting infection curve, two measures were calculated (in a manner not specified here), namely, the rate of spread of the disease (%) and the level of infection at the

end of the disease period. The experiment was implemented in three blocks, each of which consisted of two sections. Each section consisted of three plots, which were divided into two subplots, each of which had six to eight plants. Thus, there were a total of 36 subplots. The results were recorded for each subplot. The experimental factors were randomly assigned to the different units as follows: two climates to the two sections within each block, three amounts of fertilizer to the three plots within each section, and, finally, the two varieties to the two subplots within each plot. The data are shown below (Table 6.55).

- (a) Write down a statistical model of this experiment.
- (b) List all the components of the GLMM in (a).
- (c) Write down the null and alternative hypotheses associated with this experiment.
- (d) Construct an ANOVA table indicating the sources of variation and degrees of freedom.
- (e) Analyze the rate of disease spread to investigate the effect of different factors.
- (f) Comment on the results obtained.

Exercise 6.8.8 This example is an experiment to identify damage to the uterus in laboratory rodents after exposure to boric acid, a compound widely used in pesticides, pharmaceuticals, and other household products (Heindel et al. 1992). The study design included four doses of boric acid. The compound was administered to pregnant female mice during the first 17 days of gestation, and, then, the females were sacrificed and their litters examined. The table below presents the resulting trials for litters dying in utero (Y) of the total number of trials conducted (N) at each of the four doses tested: $d_1 = 0$ {control}, $d_2 = 0.1$, $d_3 = 0.2$, and $d_3 = 0.4$ (as percentage of boric acid in the diet) (Table 6.56).

- (a) Write down an ANOVA table (sources of variation, degrees of freedom) for this experiment.
- (b) List all the components of the GLMM in (a).
- (c) Analyze this dataset and summarize the relevant results.
- (d) Is there an extra-variation in the dataset? If so, what alternative distribution do you propose? Reanalyze the data and compare your findings.

Table 6.55 Greenhouse experiment results of cucumber varieties

Block	Section	Plot	Weather	Fertilizer	Variety	Proportion (%)	Level
1	1	1	2	2	Aminex	48.8981	0.06915
1	1	1	2	2	Dalibor	42.2463	0.06595
1	1	2	2	3.5	Aminex	48.2108	0.04679
1	1	2	2	3.5	Dalibor	41.6767	0.04881
1	1	3	2	4	Aminex	55.4369	0.04025
1	1	3	2	4	Dalibor	40.9562	0.04859
1	2	4	1	2	Aminex	51.5573	0.09353
1	2	4	1	2	Dalibor	36.7739	0.10353
1	2	5	1	3.5	Aminex	47.9937	0.05327
1	2	5	1	3.5	Dalibor	47.8723	0.04397
1	2	6	1	4	Aminex	57.9171	0.05225
1	2	6	1	4	Dalibor	37.7185	0.09324
1	3	7	2	2	Aminex	60.1747	0.04182
2	3	7	2	2	Dalibor	45.6937	0.06983
2	3	8	2	3.5	Aminex	51.0017	0.08863
2	3	8	2	3.5	Dalibor	52.2796	0.03622
2	3	9	2	4	Aminex	51.1251	0.05875
2	3	9	2	4	Dalibor	48.7217	0.08169
2	4	10	1	2	Aminex	51.6001	0.07001
2	4	10	1	2	Dalibor	50.4463	0.09907
2	4	11	1	3.5	Aminex	48.3387	0.05788
2	4	11	1	3.5	Dalibor	38.6538	0.06834
2	4	12	1	4	Aminex	51.3147	0.05695
2	4	12	1	4	Dalibor	38.2488	0.07908
3	5	13	1	2	Aminex	49.6958	0.07218
3	5	13	1	2	Dalibor	29.6786	0.11351
3	5	14	1	3.5	Aminex	46.6692	0.08825
3	5	14	1	3.5	Dalibor	36.5892	0.09107
3	5	15	1	4	Aminex	56.032	0.04532
3	5	15	1	4	Dalibor	36.0955	0.08712
3	6	16	2	2	Aminex	45.979	0.08882
3	6	16	2	2	Dalibor	37.2489	0.12796
3	6	17	2	3.5	Aminex	40.7277	0.06418
3	6	17	2	3.5	Dalibor	38.4831	0.0854
3	6	18	2	4	Aminex	44.5242	0.06215
3	6	18	2	4	Dalibor	34.3907	0.09651

Table 6.56 Rodent experiment results

Dose	Y	N	Dose	Y	N	Dose	Y	N	Dose	Y	N
0	0	15	0.1	0	6	0.2	1	12	0.4	12	12
0	0	3	0.1	1	14	0.2	0	12	0.4	1	12
0	1	9	0.1	1	12	0.2	0	11	0.4	0	13
0	1	12	0.1	0	10	0.2	0	13	0.4	2	8
0	1	13	0.1	2	14	0.2	0	12	0.4	2	12
0	2	13	0.1	0	12	0.2	0	14	0.4	4	13
0	0	16	0.1	0	14	0.2	4	15	0.4	0	13
0	0	11	0.1	3	14	0.2	0	14	0.4	1	13
0	1	11	0.1	0	10	0.2	0	12	0.4	0	12
0	2	8	0.1	2	12	0.2	1	6	0.4	1	9
0	0	14	0.1	3	13	0.2	2	13	0.4	3	9
0	0	13	0.1	1	11	0.2	0	10	0.4	0	11
0	3	14	0.1	1	11	0.2	1	14	0.4	1	14
0	1	13	0.1	0	11	0.2	1	12	0.4	0	10
0	0	8	0.1	0	13	0.2	0	10	0.4	3	12
0	0	13	0.1	0	10	0.2	0	9	0.4	2	21
0	2	14	0.1	1	12	0.2	1	12	0.4	3	10
0	3	14	0.1	0	11	0.2	0	13	0.4	3	11
0	0	11	0.1	2	10	0.2	1	14	0.4	1	11
0	2	12	0.1	2	12	0.2	0	13	0.4	1	11
0	0	15	0.1	2	15	0.2	0	14	0.4	8	14
0	0	15	0.1	3	12	0.2	1	13	0.4	0	15
0	2	14	0.1	1	12	0.2	2	12	0.4	2	13
0	1	11	0.1	0	12	0.2	1	14	0.4	8	11
0	1	16	0.1	1	12	0.2	0	13	0.4	4	12
0	0	12	0.1	1	13	0.2	0	12	0.4	2	12
0	0	14	0.1	1	15	0.2	1	7			

Appendix

Data: Fleas

Bioen	SP	Treat	Rep	Overvi	Dead
B1	Daphnia	T1	1	10	0
B1	Daphnia	T1	2	10	0
B1	Daphnia	T1	3	10	0
B1	Daphnia	T2	1	10	0
B1	Daphnia	T2	2	10	0
B1	Daphnia	T2	3	10	0
B1	Daphnia	T3	1	9	1
B1	Daphnia	T3	2	9	1
B1	Daphnia	T3	3	8	2

(continued)

Data: Fleas					
Bioen	SP	Treat	Rep	Overvi	Dead
B1	Daphnia	T4	1	2	8
B1	Daphnia	T4	2	2	8
B1	Daphnia	T4	3	3	7
B1	Daphnia	T5	1	0	10
B1	Daphnia	T5	2	0	10
B1	Daphnia	T5	3	0	10
B1	Daphnia	T6	1	0	10
B1	Daphnia	T6	2	0	10
B1	Daphnia	T6	3	0	10
B2	Daphnia	T1	1	10	0
B2	Daphnia	T1	2	10	0
B2	Daphnia	T1	3	10	0
B2	Daphnia	T2	1	10	0
B2	Daphnia	T2	2	10	0
B2	Daphnia	T2	3	10	0
B2	Daphnia	T3	1	9	1
B2	Daphnia	T3	2	9	1
B2	Daphnia	T3	3	9	1
B2	Daphnia	T4	1	2	8
B2	Daphnia	T4	2	2	8
B2	Daphnia	T4	3	2	8
B2	Daphnia	T5	1	0	10
B2	Daphnia	T5	2	0	10
B2	Daphnia	T5	3	0	10
B2	Daphnia	T6	1	0	10
B2	Daphnia	T6	2	0	10
B2	Daphnia	T6	3	0	10
B3	Daphnia	T1	1	10	0
B3	Daphnia	T1	2	10	0
B3	Daphnia	T1	3	10	0
B3	Daphnia	T2	1	10	0
B3	Daphnia	T2	2	10	0
B3	Daphnia	T2	3	10	0
B3	Daphnia	T3	1	8	2
B3	Daphnia	T3	2	9	1
B3	Daphnia	T3	3	9	1
B3	Daphnia	T4	1	3	7
B3	Daphnia	T4	2	2	8
B3	Daphnia	T4	3	2	8
B3	Daphnia	T5	1	0	10
B3	Daphnia	T5	2	0	10
B3	Daphnia	T5	3	0	10

(continued)

Data: Fleas					
Bioen	SP	Treat	Rep	Overvi	Dead
B3	Daphnia	T6	1	0	10
B3	Daphnia	T6	2	0	10
B3	Daphnia	T6	3	0	10
B1	Dubia	T1	1	10	0
B1	Dubia	T1	2	10	0
B1	Dubia	T1	3	10	0
B1	Dubia	T2	1	5	5
B1	Dubia	T2	2	6	4
B1	Dubia	T2	3	6	4
B1	Dubia	T3	1	5	5
B1	Dubia	T3	2	5	5
B1	Dubia	T3	3	5	5
B1	Dubia	T4	1	2	8
B1	Dubia	T4	2	3	7
B1	Dubia	T4	3	3	7
B1	Dubia	T5	1	2	8
B1	Dubia	T5	2	2	8
B1	Dubia	T5	3	2	8
B1	Dubia	T6	1	0	10
B1	Dubia	T6	2	0	10
B1	Dubia	T6	3	0	10
B2	Dubia	T1	1	10	0
B2	Dubia	T1	2	10	0
B2	Dubia	T1	3	10	0
B2	Dubia	T2	1	7	3
B2	Dubia	T2	2	5	5
B2	Dubia	T2	3	6	4
B2	Dubia	T3	1	5	5
B2	Dubia	T3	2	5	5
B2	Dubia	T3	3	5	5
B2	Dubia	T4	1	4	6
B2	Dubia	T4	2	4	6
B2	Dubia	T4	3	4	6
B2	Dubia	T5	1	2	8
B2	Dubia	T5	2	2	8
B2	Dubia	T5	3	2	8
B2	Dubia	T6	1	0	10
B2	Dubia	T6	2	0	10
B2	Dubia	T6	3	0	10
B3	Dubia	T1	1	10	0
B3	Dubia	T1	2	10	0
B3	Dubia	T1	3	10	0

(continued)

Data: Fleas

Bioen	SP	Treat	Rep	Overvi	Dead
B3	Dubia	T2	1	8	2
B3	Dubia	T2	2	8	2
B3	Dubia	T2	3	7	3
B3	Dubia	T3	1	5	5
B3	Dubia	T3	2	5	5
B3	Dubia	T3	3	6	4
B3	Dubia	T4	1	2	8
B3	Dubia	T4	2	3	7
B3	Dubia	T4	3	2	8
B3	Dubia	T5	1	3	7
B3	Dubia	T5	2	2	8
B3	Dubia	T5	3	2	8
B3	Dubia	T6	1	0	10
B3	Dubia	T6	2	0	10
B3	Dubia	T6	3	0	10

Data: Commercial crop explant detachment

Block	A	B	C	y	N
1	1	1	1	15	73
2	1	1	1	10	86
1	1	1	2	17	69
2	1	1	2	19	32
1	1	2	1	26	125
2	1	2	1	21	62
1	1	2	2	14	81
2	1	2	2	12	21
1	2	1	1	10	92
2	2	1	1	12	108
1	2	1	2	30	44
2	2	1	2	32	33
1	2	2	1	37	91
2	2	2	1	30	42
1	2	2	2	32	98
2	2	2	2	37	44
1	3	1	1	18	52
2	3	1	1	18	73
1	3	1	2	23	108
2	3	1	2	21	55
1	3	2	1	24	106
2	3	2	1	27	92
1	3	2	2	37	64
2	3	2	2	37	97

Data: Cockroaches (E1 = np, E2 = ng, E3 = adult)			
Bioassay	Isolation	Age	Dead
1	Bb1	np	7
2	Bb1	np	6
1	Bb1	ng	2
2	Bb1	ng	2
1	Bb1	a	9
2	Bb1	a	6
1	Bb2	np	6
2	Bb2	np	7
1	Bb2	ng	7
2	Bb2	ng	3
1	Bb2	a	10
2	Bb2	a	8
1	Bb3	np	3
2	Bb3	np	2
1	Bb3	ng	2
2	Bb3	ng	3
1	Bb3	a	8
2	Bb3	a	9
1	Bb4	np	6
2	Bb4	np	4
1	Bb4	ng	5
2	Bb4	ng	3
1	Bb4	a	10
2	Bb4	a	8
1	Bb5	np	7
2	Bb5	np	7
1	Bb5	ng	3
2	Bb5	ng	1
1	Bb5	a	7
2	Bb5	a	3
1	Bb6	np	7
2	Bb6	np	9
1	Bb6	ng	6
2	Bb6	ng	2
1	Bb6	a	10
2	Bb6	a	7
1	Bb8	np	9
2	Bb8	np	9
1	Bb8	ng	4
2	Bb8	ng	2
1	Bb8	a	9
2	Bb8	a	10

(continued)

Data: Cockroaches (E1 = np, E2 = ng, E3 = adult)			
Bioassay	Isolation	Age	Dead
1	Bb9	np	5
2	Bb9	np	8
1	Bb9	ng	6
2	Bb9	ng	2
1	Bb9	a	7
2	Bb9	a	5
1	Bb10	np	8
2	Bb10	np	6
1	Bb10	ng	1
2	Bb10	ng	4
1	Bb10	a	3
2	Bb10	a	4
1	Bb11	np	8
2	Bb11	np	7
1	Bb11	ng	1
2	Bb11	ng	3
1	Bb11	a	6
2	Bb11	a	8
1	Bb12	np	8
2	Bb12	np	9
1	Bb12	ng	8
2	Bb12	ng	9
1	Bb12	a	7
2	Bb12	a	6
1	Bb13	np	6
2	Bb13	np	3
1	Bb13	ng	0
2	Bb13	ng	1
1	Bb13	a	5
2	Bb13	a	6
1	Bb14	np	10
2	Bb14	np	5
1	Bb14	ng	4
2	Bb14	ng	2
1	Bb14	a	6
2	Bb14	a	6
1	Bb15	np	5
2	Bb15	np	10
1	Bb15	ng	6
2	Bb15	ng	1
1	Bb15	a	4
2	Bb15	a	5

(continued)

Data: Cockroaches (E1 = np, E2 = ng, E3 = adult)

Bioassay	Isolation	Age	Dead
1	Bb16	np	5
2	Bb16	np	7
1	Bb16	ng	3
2	Bb16	ng	4
1	Bb16	a	8
2	Bb16	a	6
1	Control	np	1
2	Control	np	0
1	Control	ng	0
2	Control	ng	0
1	Control	a	0
2	Control	a	1

Data: Disease incidence in grapevine plants (b = block, v = plant, r = shoot, t = treatment, m = number of diseased leaves per shoot, and n = total number of leaves per shoot).

b	v	r	t	M	n
1	1	1	1	1	14
1	1	1	2	2	12
1	1	1	3	0	12
1	1	1	4	0	13
1	1	1	5	3	8
1	1	1	6	0	9
1	1	2	1	7	8
1	1	2	2	0	10
1	1	2	3	1	14
1	1	2	4	0	10
1	1	2	5	0	17
1	1	2	6	0	10
1	1	3	1	9	14
1	1	3	2	1	11
1	1	3	3	0	10
1	1	3	4	1	14
1	1	3	5	0	10
1	1	3	6	0	21
1	1	4	1	10	17
1	1	4	2	0	9
1	1	4	3	1	12
1	1	4	4	0	11
1	1	4	5	0	12
1	1	4	6	0	10
1	1	5	1	8	11
1	1	5	2	1	10

(continued)

Data: Disease incidence in grapevine plants (b = block, v = plant, r = shoot, t = treatment, m = number of diseased leaves per shoot, and n = total number of leaves per shoot).

b	v	r	t	M	n
1	1	5	3	0	9
1	1	5	4	2	12
1	1	5	5	0	10
1	1	5	6	1	11
1	2	1	1	7	9
1	2	1	2	2	10
1	2	1	3	0	10
1	2	1	4	0	14
1	2	1	5	1	12
1	2	1	6	0	13
1	2	2	1	6	12
1	2	2	2	0	11
1	2	2	3	1	13
1	2	2	4	0	9
1	2	2	5	2	11
1	2	2	6	0	10
1	2	3	1	6	7
1	2	3	2	1	12
1	2	3	3	0	9
1	2	3	4	1	10
1	2	3	5	0	14
1	2	3	6	2	12
1	2	4	1	7	13
1	2	4	2	0	10
1	2	4	3	0	10
1	2	4	4	1	12
1	2	4	5	0	9
1	2	4	6	1	8
1	2	5	1	11	15
1	2	5	2	1	13
1	2	5	3	0	14
1	2	5	4	1	14
1	2	5	5	0	11
1	2	5	6	0	11
1	3	1	1	5	11
1	3	1	2	5	11
1	3	1	3	0	15
1	3	1	4	1	15
1	3	1	5	0	8
1	3	1	6	1	10
1	3	2	1	4	9
1	3	2	2	1	15

(continued)

Data: Disease incidence in grapevine plants (b = block, v = plant, r = shoot, t = treatment, m = number of diseased leaves per shoot, and n = total number of leaves per shoot).

b	v	r	t	M	n
1	3	2	3	0	11
1	3	2	4	0	13
1	3	2	5	1	12
1	3	2	6	0	12
1	3	3	1	9	12
1	3	3	2	2	14
1	3	3	3	0	12
1	3	3	4	0	12
1	3	3	5	0	10
1	3	3	6	0	13
1	3	4	1	10	10
1	3	4	2	0	10
1	3	4	3	0	8
1	3	4	4	0	10
1	3	4	5	1	14
1	3	4	6	3	11
1	3	5	1	9	11
1	3	5	2	0	11
1	3	5	3	1	11
1	3	5	4	1	14
1	3	5	5	0	9
1	3	5	6	0	9
2	1	1	1	0	12
2	1	1	2	0	12
2	1	1	3	0	14
2	1	1	4	0	12
2	1	1	5	0	10
2	1	1	6	1	13
2	1	2	1	8	9
2	1	2	2	1	9
2	1	2	3	0	12
2	1	2	4	0	10
2	1	2	5	0	12
2	1	2	6	1	10
2	1	3	1	11	14
2	1	3	2	1	12
2	1	3	3	1	11
2	1	3	4	0	10
2	1	3	5	0	9
2	1	3	6	3	11
2	1	4	1	12	15
2	1	4	2	0	15

(continued)

Data: Disease incidence in grapevine plants (b = block, v = plant, r = shoot, t = treatment, m = number of diseased leaves per shoot, and n = total number of leaves per shoot).

b	v	r	t	M	n
2	1	4	3	0	10
2	1	4	4	1	9
2	1	4	5	1	10
2	1	4	6	0	16
2	1	5	1	10	14
2	1	5	2	1	9
2	1	5	3	0	11
2	1	5	4	0	11
2	1	5	5	0	11
2	1	5	6	0	11
2	2	1	1	1	9
2	2	1	2	0	9
2	2	1	3	0	12
2	2	1	4	1	10
2	2	1	5	1	12
2	2	1	6	0	17
2	2	2	1	9	12
2	2	2	2	0	12
2	2	2	3	0	11
2	2	2	4	2	14
2	2	2	5	0	11
2	2	2	6	0	10
2	2	3	1	7	13
2	2	3	2	0	16
2	2	3	3	1	12
2	2	3	4	0	10
2	2	3	5	0	10
2	2	3	6	0	11
2	2	4	1	7	13
2	2	4	2	1	18
2	2	4	3	0	10
2	2	4	4	0	11
2	2	4	5	0	11
2	2	4	6	3	13
2	2	5	1	5	10
2	2	5	2	0	10
2	2	5	3	0	10
2	2	5	4	0	10
2	2	5	5	0	9
2	2	5	6	1	12
2	3	1	1	6	13
2	3	1	2	0	10

(continued)

Data: Disease incidence in grapevine plants (b = block, v = plant, r = shoot, t = treatment, m = number of diseased leaves per shoot, and n = total number of leaves per shoot).

b	v	r	t	M	n
2	3	1	3	1	11
2	3	1	4	3	11
2	3	1	5	0	12
2	3	1	6	1	19
2	3	2	1	12	13
2	3	2	2	0	11
2	3	2	3	0	8
2	3	2	4	0	9
2	3	2	5	0	17
2	3	2	6	0	12
2	3	3	1	8	11
2	3	3	2	4	12
2	3	3	3	0	11
2	3	3	4	0	10
2	3	3	5	0	15
2	3	3	6	3	13
2	3	4	1	5	14
2	3	4	2	1	9
2	3	4	3	0	12
2	3	4	4	1	12
2	3	4	5	0	10
2	3	4	6	2	14
2	3	5	1	10	14
2	3	5	2	0	14
2	3	5	3	1	10
2	3	5	4	1	13
2	3	5	5	1	15
2	3	5	6	4	10
3	1	1	1	8	12
3	1	1	2	1	14
3	1	1	3	0	12
3	1	1	4	0	20
3	1	1	5	1	18
3	1	1	6	7	15
3	1	2	1	9	16
3	1	2	2	1	12
3	1	2	3	0	13
3	1	2	4	0	15
3	1	2	5	0	17
3	1	2	6	1	18
3	1	3	1	7	12
3	1	3	2	0	14

(continued)

Data: Disease incidence in grapevine plants (b = block, v = plant, r = shoot, t = treatment, m = number of diseased leaves per shoot, and n = total number of leaves per shoot).

b	v	r	t	M	n
3	1	3	3	1	13
3	1	3	4	0	18
3	1	3	5	0	14
3	1	3	6	0	14
3	1	4	1	10	14
3	1	4	2	2	17
3	1	4	3	0	10
3	1	4	4	1	19
3	1	4	5	0	17
3	1	4	6	0	16
3	1	5	1	9	10
3	1	5	2	1	14
3	1	5	3	1	11
3	1	5	4	0	18
3	1	5	5	0	15
3	1	5	6	1	11
3	2	1	1	10	10
3	2	1	2	1	11
3	2	1	3	0	12
3	2	1	4	1	15
3	2	1	5	4	20
3	2	1	6	0	14
3	2	2	1	9	12
3	2	2	2	1	10
3	2	2	3	1	12
3	2	2	4	3	18
3	2	2	5	0	16
3	2	2	6	0	12
3	2	3	1	10	11
3	2	3	2	1	16
3	2	3	3	1	14
3	2	3	4	1	17
3	2	3	5	2	15
3	2	3	6	1	16
3	2	4	1	9	11
3	2	4	2	2	14
3	2	4	3	0	10
3	2	4	4	0	18
3	2	4	5	0	17
3	2	4	6	0	12
3	2	5	1	11	12
3	2	5	2	2	12

(continued)

Data: Disease incidence in grapevine plants (b = block, v = plant, r = shoot, t = treatment, m = number of diseased leaves per shoot, and n = total number of leaves per shoot).

b	v	r	t	M	n
3	2	5	3	0	11
3	2	5	4	0	13
3	2	5	5	0	18
3	2	5	6	0	12
3	3	1	1	7	9
3	3	1	2	0	13
3	3	1	3	0	9
3	3	1	4	0	18
3	3	1	5	0	18
3	3	1	6	0	13
3	3	2	1	6	14
3	3	2	2	3	16
3	3	2	3	1	15
3	3	2	4	0	17
3	3	2	5	1	17
3	3	2	6	3	14
3	3	3	1	10	11
3	3	3	2	0	10
3	3	3	3	1	16
3	3	3	4	1	18
3	3	3	5	0	16
3	3	3	6	0	11
3	3	4	1	10	10
3	3	4	2	1	14
3	3	4	3	0	10
3	3	4	4	1	19
3	3	4	5	2	19
3	3	4	6	2	14
3	3	5	1	8	10
3	3	5	2	0	12
3	3	5	3	0	12
3	3	5	4	0	18
3	3	5	5	0	14
3	3	5	6	0	12

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



Chapter 7

Time of Occurrence of an Event of Interest



7.1 Introduction

In studies such as biological sciences, animal science, and agronomy, a common outcome of interest is the time at which an event of interest occurs. The main characteristic of these data is that the subjects/experimental units are usually observed for different periods of time until the event of interest occurs. These events of interest may be adverse events such as the death of an experimental unit and the cessation of lactation, or positive events such as the conception of a female's offspring from a particular treatment and the onset of estrus in a female undergoing hormone treatment, among others. Because of the characteristics of these response variables, a "normal" distribution is often a poor choice for modeling the time at which the event of interest occurs. Exponential, log-normal, gamma, Weibull, and other more complex distributions that tend to be more common and are better choices for modeling these phenomena.

Fitting a generalized linear mixed model (GLMM) is a good option for analyzing these phenomena because the conditional response distribution of the random effects of this model has desirable properties. In this vein, it is conventional to speak of survival data and survival analysis, regardless of the nature of the event. Similar data also arise when measuring the time to complete a task, such as walking 50 meters, passing an agronomy exam, performing a sensory evaluation of coffee, and so on. The purpose of this chapter is to provide the reader with the essential language of linear models and the connection between GLMMs and survival analysis.

7.2 Generalized Linear Mixed Models with a Gamma Response

The gamma family of distributions encompasses continuous, nonnegative, right-skewed values. A gamma distribution has two nonnegative parameters α and β –the probability density function of which is given by:

$$f(y; \alpha, \beta) = \frac{1}{\Gamma(\alpha)\beta^\alpha} y^{\alpha-1} e^{-y/\beta}, y \geq 0$$

where $\Gamma(\alpha) = \int_0^\infty t^{\alpha-1} e^{-t} dt$ is the gamma function (Casella and Berger 2002). The mean and variance of a random gamma variable are $E[Y] = \alpha\beta = \mu$ and $\text{Var}[Y] = \alpha\beta^2 = \mu^2/\alpha$, respectively. This density function can be rewritten in terms of the mean μ and the scale parameter $\phi = 1/\alpha$.

$$f(y; \alpha, \beta) = \frac{1}{\Gamma\left(\frac{1}{\phi}\right) (\mu\phi)^{1/\phi} y^{1/\phi-1} e^{-y/\mu\phi}}, y \geq 0.$$

7.2.1 CRD: Estrus Induction in Pelibuey Ewes

Estrus induction in ewes is a very common practice carried out in livestock farms or at research centers. For this, an animal researcher uses gonadotropin-releasing hormone (GnRH), equine chorionic gonadotropin (eCG), and P4 in a controlled internal drug-releasing (CIDR) intravaginal device in female Pelibuey ewes ($n = 78$) with single, double, and triple lambing as treatments. In order to ensure that all animals were in good condition during the experiment, ewes received the same zootechnical management and feeding. For this experiment, the ewes were synchronized on the same day under a synchronization protocol. Table 7.1 presents the analysis of variance (ANOVA).

The variables evaluated in this experiment were the time of onset and duration of estrus (y_{ij}) in hours according to the type of calving. The variability among female sheep on weight, age, and body condition must be taken into account in the

Table 7.1 Sources of variation and degrees of freedom

Sources of variation	Degrees of freedom
Treatment	$t - 1 = 3 - 1 = 2$
Error	$\sum_{i=1}^3 r_i - t = 75$
Total	$\sum_{i=1}^3 r_i - 1 = 78 - 1 = 77$

Table 7.2 Results of the analysis of variance

(a) Covariance parameter estimates						
Cov Parm	Start of estrus		Duration of estrus			
	Estimate	Standard error	Estimate	Standard error		
Parto (animal) ($\hat{\sigma}_{\text{birthtype}(\text{animal})}^2$)	-0.01572	.	0.08370	0.09692		
Residual ($\hat{\phi}$)	0.06668	0.01232	0.2073	0.08938		
(b) Type III tests of fixed effects						
Effect	Num DF	Den DF	Inicio estro		Duración estro	
			F-value	Pr > F	F-value	Pr > F
Birth type	2	75	5.12	0.0082	22.61	<0.0001

analysis. The data from this experiment can be found in the Appendix 1 of this book (Data: Pelibuey Sheep). Thus, the components of a gamma GLMM are as follows:

$$\text{Distributions: } y_{ij} \mid \tau(r)_{ij} \sim \text{Gamma}(\mu_{ij}, \phi); i = 1, 2, 3; j = 1, \dots, r_i.$$

$$r(\tau)_{ij} \sim N(0, \sigma_{\tau(\text{animal})}^2)$$

$$\text{Linear predictor: } \eta_{ij} = \mu + \tau_i + \tau(r)_{ij}$$

$$\text{Link function: } \log(\mu_{ij}) = \eta_{ij}$$

where η_{ij} is the i th link function for treatment i (type of birth angle, double or triple) in ewes j , μ is the overall mean, τ_i is the fixed effect due to type of birth (treatment), $r(\tau)_{ij}$ is the random effect due to type of birth (treatment) in ewes j with $\tau(r)_j \sim N(0, \sigma_{\tau(\text{animal})}^2)$.

The following GLIMMIX program fits the model

```
proc glimmix nobound method=laplace;
class animal birthtype;
model Inestro = birthtype/dist=gamma;
random birthtype (animal);
lsmeans birthtype/lines ilink;
run;
```

Part of the results is reported in Table 7.2.

Subsection (a) shows the estimated variance components due to the type of parturition used in females ($\hat{\sigma}_{\text{birthtype}(\text{animal})}^2 = -0.0157(\pm 0.0837)$) as well as the scale parameter ($\hat{\phi} = 0.06668$).

Table 7.2 (b) shows the results of the hypothesis tests for type III fixed effects, which indicate that there is a statistically significant effect of treatment (type of birth) on the time of onset and duration of ewe estrus.

Table 7.3 Means and standard errors on the model scale (“Estimate” column) and the data scale (“Mean” column) for the onset and duration of estrus in Pelibuey ewe lambs

Parto least squares means							
Birth_type	Estimate	Standard error	DF	t-value	Pr > t	Mean	Standard error mean
Start of estrus							
1	3.2913	0.06631	75	49.63	<0.0001	26.8787	1.7824
2	3.0622	0.04606	75	66.48	<0.0001	21.3735	0.9845
3	3.0496	0.04542	75	67.14	<0.0001	21.1059	0.9586
Duration of estrus							
1	1.6826	0.1518	75	11.09	<0.0001	5.3795	0.8164
2	2.6716	0.1171	75	22.81	<0.0001	14.4637	1.6938
3	2.8075	0.09846	75	28.51	<0.0001	16.5684	1.6313

The last two columns of Table 7.3, labeled “Mean” and “Standard error,” correspond to the means (μ_{ij}) on the data scale for the ewes’ mean onset and duration of estrus with their respective standard errors. For example, the mean time to onset of estrus in single-birth ewes was 26.87 ± 1.78 hours, whereas for double- and triple-birth ewes, it was 21.37 ± 0.98 and 21.1 ± 0.95 , respectively. On the other hand, the average time (in hours) of estrus duration was longer in double- and triple-birth ewes (14.46 ± 1.69 and 16.56 ± 1.63 , respectively) compared to single-birth ewes (5.38 ± 0.81).

7.2.2 *Randomized Complete Block Design (RCBD): Itch Relief Drugs*

A total of 10 male volunteer patients between 20 and 30 years of age participated as a study group to compare 7 treatments (Trts) (5 drugs, 1 placebo, and 1 no drug) to relieve their itching. Since each subject responded differently to each drug, and, in addition, each subject received a different treatment in the 7 days of study, each of the subjects can be considered a block. Treatment assignment was randomized across days. Except for the drug-free day, subjects were administered the treatment intravenously, and, then, their forearms were induced to itch using an effective itch stimulus called cowage. The duration of itching, in seconds, was recorded. The data are shown in Table 7.4.

From left to right, the drugs used were papaverine = Papv, morphine = Morp, aminophylline = Amino, pentobarbital = Pent, and tripelethamine pentobarbital = Tripel.

The analysis of variance table (Table 7.5) shows the sources of variation and degrees of freedom for this experiment.

Table 7.4 Time taken to get rid of the itch

Patient	No drug	Placebo	Papv	Morp	Amino	Pent	Tripel
1	174	263	105	199	141	108	141
2	224	213	103	143	168	341	184
3	260	231	145	113	78	159	125
4	255	291	103	225	164	135	227
5	165	168	144	176	127	239	194
6	237	121	94	144	114	136	155
7	191	137	35	87	96	140	121
8	100	102	133	120	222	134	129
9	115	89	83	100	165	185	79
10	189	433	237	173	168	188	317

Table 7.5 Sources of variation and degrees of freedom

Sources of variation	Degrees of freedom
Blocks	$r - 1 = 10 - 1 = 9$
Treatment	$t - 1 = 7 - 1 = 6$
Error	$(t - 1)(r - 1) = 6 \times 9 = 54$
Total	$r \times t - 1 = 10 \times 7 - 1 = 69$

The components of the GLMM with a gamma response are as follows:

Distributions : $y_{ij} \mid r(\alpha\beta)_{ijk} \sim \text{Gamma}(\mu_{ij}, \phi); i = 1, \dots, 7; j = 1, \dots, 10.$

$$r_j \sim N\left(0, \sigma_{\text{patient}}^2\right)$$

Linear predictor: $\eta_{ij} = \mu + r_j + \tau_i$

Link function: $\log(\mu_{ijk}) = \eta_{ijk}$

where η_{ij} is the predictor with treatment i and block j , μ is the overall mean, r_j is the random effect of the patient with $r_j \sim N\left(0, \sigma_{\text{patient}}^2\right)$, and τ_i is the fixed effect due to treatment.

Note, although the exponential and gamma distributions have a canonical link equal to the inverse of the mean, the gamma and exponential GLMMs most often use a computationally more stable link (link = log), which was used in this and in the previous analysis.

The following GLIMMIX syntax adjusts a GLMM into complete blocks.

```
proc glimmix nobound method=laplace;
class Patient Trt;
model y = Trt/dist=gamma;
random Patient;
lsmeans Trt/lines ilink;
run;
```

Table 7.6 Results of the analysis of variance

(a) Fit statistics for conditional distribution				
-2 Log L (y r. effects)			728.62	
Pearson's chi-square			5.69	
Pearson's chi-square/DF			0.08	
(b) Covariance parameter estimates				
Cov Parm	Estimate	Standard error		
Patient	0.03964	0.02375		
Residual ($\hat{\phi}$)	0.09132	0.01640		
(c) Type III tests of fixed effects				
Effect	Num DF	Den DF	F-value	Pr > F
Trt	6	54	3.82	0.0030

The statistics of the conditional model (Pearson's chi - square/DF = 0.08) as well as the variance components (Patient) and the scale parameter ($\hat{\phi}$) of the model indicate that the gamma model adequately describes the dataset (Table 7.6 parts (a) and (b)). The analysis of variance (Table 7.6 part (c)) indicates that there is a highly significant difference of treatments in the mean time of itch duration ($P = 0.0030$).

The dispersion observed in the following plot (top left) of the residuals versus the linear predictor value suggests that the variance is constant and homogeneous (Fig. 7.1). The histogram (upper right) shows a nearly symmetrical pattern with little bias. Furthermore, the residuals versus quantile plot (bottom left) shows no marked deviations, indicating that the fit is adequate. Finally, the bottom right plot shows that the average residuals are zero and vary between -0.5 and 0.75 .

The "lsmeans" on the data scale, for each of the five treatments, placebo, and the control treatment, are shown under the "Mean" column with their respective "Standard error" in Table 7.7. Each of the five drugs appear to have a significant effect compared to the placebo and control. Papaverine (Papv) is the most effective drug. Both the placebo and control treatment have statistically similar means. The relatively large difference in the placebo group suggests that some patients responded negatively to the placebo compared to the control, whereas others responded positively.

Figure 7.2 shows that the drug papaverine significantly reduced the itching time, followed by the drugs aminophylline and morphine, whereas the efficacies of the drugs pentobarbital and tripeleennamine were highly similar to each other in eliminating itching.

7.2.3 Factorial Design: Insect Survival Time

This experiment consisted of studying the effectiveness of four different types of insecticides (Insec1, Insec2, Insec3, and Insec4) at three different concentration levels (low, medium, and high) in the survival time (in hours) of a particular species

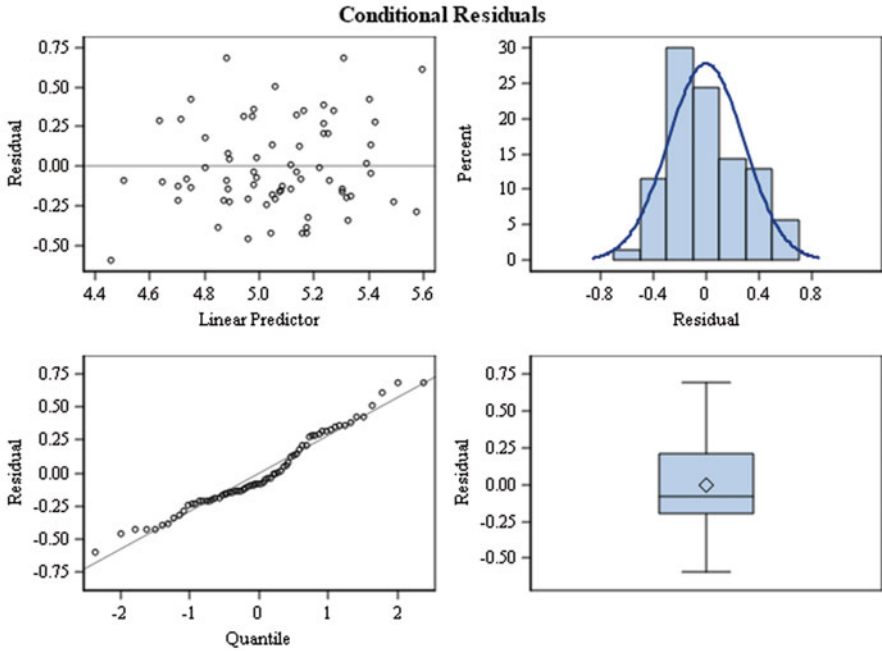


Fig. 7.1 Conditional residuals

Table 7.7 Means and standard errors on the model scale (“Estimate” column) and the data scale (“Mean” column) for the average duration time of the itch

Trt least squares means							
Trt	Estimate	Standard error	DF	t-value	Pr > t	Mean	Standard error of mean
Amino	4.9795	0.1149		43.32	<0.0001	145.41	16.7129
Morp	4.9797	0.1146		43.44	<0.0001	145.43	16.6733
Papv	4.7356	0.1149		41.20	<0.0001	113.93	13.0956
Pento	5.1703	0.1149		44.99	<0.0001	175.97	20.2211
Placebo	5.2704	0.1151		45.79	<0.0001	194.49	22.3867
No drug	5.2542	0.1148		45.76	<0.0001	191.36	21.9723
Tripel	5.0802	0.1147		44.28	<0.0001	160.80	18.4487

of beetles (Appendix 1: Data: Beetles). The interaction between both factors (insecticide * dose) yielded a total of 12 combinations (treatments). The objective of this study was to compare the insecticides, dose, and interaction with beetle survival time. Due to the intrinsic characteristics of each of the insects, these must be considered as a source of variation in the experiment, since they respond differently

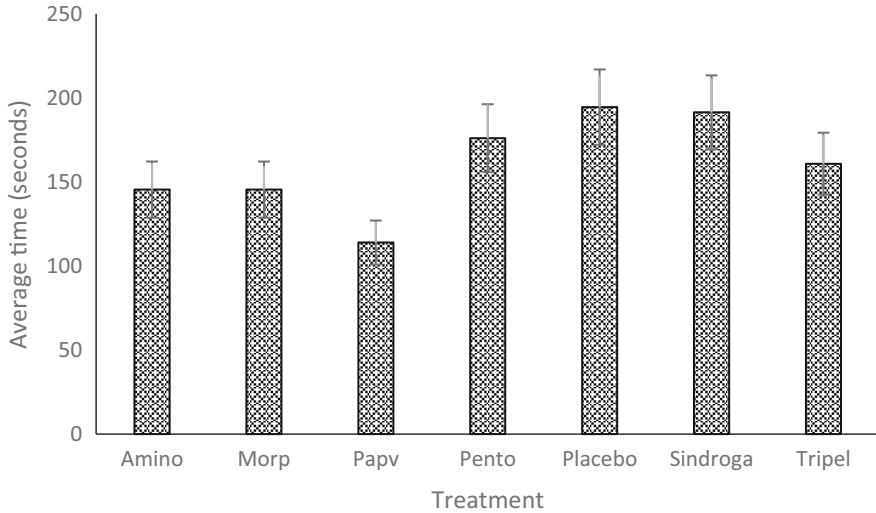


Fig. 7.2 Average time taken to eliminate itching

Table 7.8 Sources of variation and degrees of freedom

Sources of variation	Degrees of freedom
Blocks	$r - 1 = 4 - 1 = 3$
Insecticide	$a - 1 = 4 - 1 = 3$
Dose	$b - 1 = 3 - 1 = 2$
Insecticide * dosage	$(a - 1)(b - 1) = 3 \times 2 = 6$
Error	$ab(r - 1) = 4 \times 3 \times 3 = 33$
Total	$r \times a \times b - 1 = 4 \times 4 \times 3 - 1 = 47$

to certain stimuli. Assuming that 48 beetles are available, they were randomly assigned equally to 4 groups (blocks) with 12 treatment combinations. That is, four beetles were randomly assigned to each treatment.

The sources of variation and degrees of freedom for this experiment are shown in the following analysis of variance table (Table 7.8).

The components of the gamma-response GLMM are as follows:

$$\text{Distributions : } y_{ijk} \mid r_k \sim \text{Gamma}(\mu_{ijk}, \phi); i = 1, \dots, 4; j = 1, 2, 3; k = 1, \dots, 4.$$

$$r_k \sim N(0, \sigma_{\text{block}}^2)$$

$$\text{Linear predictor: } \eta_{ijk} = \mu + r_k + \alpha_i + \beta_j + (\alpha\beta)_{ij}$$

$$\text{Link function: } \log(\mu_{ijk}) = \eta_{ijk}$$

The following GLIMMIX command adjusts a GLMM with a gamma response.

Table 7.9 Results of the analysis of variance

(a) Fit statistics for conditional distribution				
-2 Log L (tiempo r. effects)				121.05
Pearson's chi-square				1.91
Pearson's chi-square/DF				0.04
(b) Covariance parameter estimates				
Cov Parm	Estimate	Standard error		
block	-0.00173	.		
Residual	0.04155	0.008818		
(c) Type III tests of fixed effects				
Effect	Num DF	Den DF	F-value	Pr > F
Dose	2	33	69.61	<0.0001
Insecticide	3	33	31.36	<0.0001
Dose*insecticide	6	33	2.05	0.0868

```
proc glimmix nobound method=laplace;
class dose insecticide insect;
model time = dose|insecticide/dist=gamma;
random insect;
lsmeans dose|insecticide/lines ilink;
run;
```

Part of the Statistical Analysis Software (SAS) output is shown in Table 7.9. The value of the conditional model's Pearson's chi - square/DF = 0.04 indicates that the gamma distribution adequately models the data. The estimated variance component for blocks and the scaling parameter given by the "residual" value are shown below (in part (b)) ($\hat{\sigma}_{\text{block}}^2 = -0.00173$, and $\hat{\sigma}^2 = 0.04155$, respectively).

The analysis of variance in (c) of Table 7.9 indicates that the insecticides and dose ($P = 0.0001$) have different significant effectiveness (toxicity) on beetle survival time. However, the interaction between both factors is close to significance ($P = 0.0868$). The "lsmeans" values on the data scale for dose $\hat{\mu}_{i..}$ (part (a)) and insecticide $\hat{\mu}_{.j}$ (part (b)) with their respective standard errors for both factors are listed under the columns titled "Mean" and "Standard error mean" of Table 7.10, respectively.

The combination of levels of both factors affected the average survival time of the beetles (Table 7.11). For insecticides 1 and 3 at a high dose, the survival time was lower with average times of 2.1 ± 0.209 and 2.35 ± 0.334 hours, respectively. In general, low values of survival times were observed for insecticides 1 and 3 compared to insecticides 2 and 4.

7.2.4 A Split Plot with a Factorial Structure on a Large Plot in a Completely Randomized Design (CRD)

Four samples were obtained from each of two batches (Reps) of unprocessed gum from *Acacia sp.* Trees, with eight samples in total. Within each batch, the four

Table 7.10 Means and standard errors on the model scale (“Estimate”) and the data scale (“Mean”) for the factor dose and type of insecticide

(a) Dose least squares means							
Dose	Estimate	Standard error	DF	<i>t</i> -value	Pr > <i>t</i>	Mean	Standard error mean
High	0.9960	0.04984	33	19.98	<0.0001	2.7075	0.1349
Low	1.7840	0.04984	33	35.79	<0.0001	5.9538	0.2967
Medium	1.6203	0.04984	33	32.51	<0.0001	5.0548	0.2519
(b) Insecticide least squares means							
Insecticide	Estimate	Standard error	DF	<i>t</i> -value	Pr > <i>t</i>	Mean	Standard error mean
Insec1	1.1074	0.05755	33	19.24	<0.0001	3.0265	0.1742
Insec2	1.8272	0.05755	33	31.75	<0.0001	6.2166	0.3578
Insec3	1.3041	0.05755	33	22.66	<0.0001	3.6845	0.2121
Insec4	1.6284	0.05755	33	28.29	<0.0001	5.0960	0.2933

Table 7.11 Means and standard errors on the model scale and the data scale for the interaction between dose and type of insecticide

Dose*insecticide least squares means								
Dose	Insecticide	Estimate	Standard error	DF	<i>t</i> -value	Pr > <i>t</i>	Mean	Standard error mean
High	Insec1	0.7419	0.09968	33	7.44	<0.0001	2.1000	0.2093
High	Insec2	1.2089	0.09968	33	12.13	<0.0001	3.3499	0.3339
High	Insec3	0.8545	0.09969	33	8.57	<0.0001	2.3501	0.2343
High	Insec4	1.1788	0.09969	33	11.82	<0.0001	3.2503	0.3240
Low	Insec1	1.4171	0.09968	33	14.22	<0.0001	4.1250	0.4112
Low	Insec2	2.1747	0.09968	33	21.82	<0.0001	8.7998	0.8772
Low	Insec3	1.7361	0.09969	33	17.42	<0.0001	5.6754	0.5658
Low	Insec4	1.8082	0.09968	33	18.14	<0.0001	6.0994	0.6080
Medium	Insec1	1.1632	0.09969	33	11.67	<0.0001	3.2000	0.3190
Medium	Insec2	2.0980	0.09968	33	21.05	<0.0001	8.1499	0.8124
Medium	Insec3	1.3218	0.09969	33	13.26	<0.0001	3.7501	0.3738
Medium	Insec4	1.8984	0.09969	33	19.04	<0.0001	6.6753	0.6654

samples were randomly assigned to combinations of two factors with two levels each. The first factor refers to whether the gum was demineralized or not, and the second factor refers to whether the gum was pasteurized or not. An emulsion made from each gum sample was divided into three smaller parts, which were randomly assigned to the levels of a third factor, the PH, and pH was adjusted to 2.5, 4.5, or 5.5 using citric acid (Appendix 1: Data: Gum Breakdown Times).

This is a split-plot design, with whole plots and rubber samples in a block arrangement. The combined levels of demineralization and pasteurization of the paste are large (whole) plot factors. The split plots are the smaller parts, with a specific pH, which is the only split-plot factor. The response measured (*y*) was the

Table 7.12 Sources of variation and degrees of freedom

Sources of variation	Degrees of freedom
Demineralization (Des)	$a - 1 = 2 - 1 = 1$
Pasteurization (Pasteu)	$b - 1 = 2 - 1 = 1$
Demineralization*pasteurization	$(a - 1)(b - 1) = 1$
Des*Pasteu (rep)	$ab(r - 1) = 2 \times 2 \times 1 = 4$
pH	$(c - 1) = 3 - 1 = 2$
Demineralization*pH	$(a - 1)(c - 1) = 2$
Pasteurization*pH	$(b - 1)(c - 1) = 2$
Des*Pasteu*pH	$(a - 1)(b - 1)(c - 1) = 2$
Error	$ab(c - 1)(r - 1) = 2 \times 2 \times 2 \times 1 = 8$
Total	$r \times a \times b \times c - 1 = 2 \times 2 \times 2 \times 3 - 1 = 23$

time to break, i.e., the time (in hours) until the emulsion failed. The sources of variation and degrees of freedom for this experiment are shown in Table 7.12.

The components of the GLMM with a Gamma response are as follows:

Distributions: $y_{ijkl} \mid r_l, \alpha\beta(r)_{ijl} \sim \text{Gamma}(\mu_{ijkl}, \phi); i = 1, 2; j = 1, 2; k = 1, 2, 3; l = 1, 2.$

$$r_l \sim N(0, \sigma_r^2), \alpha\beta(r)_{ijl} \sim N(0, \sigma_{r\alpha\beta}^2)$$

Linear predictor: $\eta_{ijkl} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + r(\alpha\beta)_{ijl} + \gamma_k + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk} + (\alpha\beta\gamma)_{ijk};$

where $\alpha_i, \beta_j,$ and γ_k are the fixed effects due to the factors demineralization, pasteurization, and pH, respectively; the effects $(\alpha\beta)_{ij}, (\alpha\gamma)_{ik}, (\beta\gamma)_{jk},$ and $(\alpha\beta\gamma)_{ijk}$ are the two- and three-way interactions of the factors under study; and $\alpha\beta(r)_{ijl}$ are random effects due to the demineralization x pasteurization x rep interaction, assuming that $\alpha\beta(r)_{ijl} \sim N(0, \sigma_{r\alpha\beta}^2).$

Link function: $\log(\mu_{ijk}) = \eta_{ijk}$

The GLIMMIX commands for setting this GLMM are as follows:

```
proc glimmix nobound method=laplace;
class Batch Demineralization Pasteurization pH;
model y = Demineralization|Pasteurization|pH/dist=gamma;
random batch (Demineralization*Pasteurization);
lsmeans Demineralization|Pasteurization|pH/lines ilink;
run;
```


Table 7.13 Results of the analysis of variance

(a) Fit statistics for conditional distribution				
-2 Log L (y r. effects)				192.24
Pearson's chi-square				0.12
Pearson's chi-square/DF				0.01
(b) Covariance parameter estimates				
Cov Parm	Estimate	Standard error		
Rep (Desmin*Pasteur)	0.001428	0.001864		
Residual ($\hat{\phi}$)	0.006011	0.002126		
(c) Type III tests of fixed effects				
Effect	Num DF	Den DF	F-value	Pr > F
Demineralization (Des)	1	4	35.48	0.0040
Pasteurization (Pasteu)	1	4	19.49	0.0116
Demineralization*pasteurization	1	4	35.67	0.0039
pH	2	8	5.27	0.0346
Demineralization*pH	2	8	3.84	0.0676
Pasteurization*pH	2	8	0.57	0.5889
Des*Pasteu*pH	2	8	4.32	0.0535

The relevant results from the SAS output are shown in Table 7.13. The value of the conditional model $\frac{\chi^2}{DF} = 0.01$ indicates that the gamma distribution does not cause overdispersion. The variance component due to blocks \times demineralization \times pasteurization $\hat{\sigma}_{r(\alpha\beta)}^2$ and the scale parameter $\hat{\phi}$ are shown in (b).

The hypothesis tests for type III fixed effects are presented in part (c) of Table 7.13, where a significant effect of the factors demineralization, pasteurization, and pH as well as the interaction between demineralization with pasteurization are observed on the gum. However, the interactions demineralization*pH ($P = 0.0676$) and demineralization*pasteurization*pH are close to significance ($P = 0.0535$). The emulsion breaking time is strongly affected by no demineralization (demineralization = 1) and no pasteurization (pasteurization = 1) of the gum and, to a lesser extent, by the pH adjusted to the gum (Table 7.14).

Analyzing the simple effects of the factors, we can observe that when the gum has not been pasteurized ($B = 1$), the average emulsion break time is very similar in the demineralized paste than in the non-demineralized paste at the three pH levels. However, when the gum has been pasteurized, demineralization has a significant impact on the emulsion breakup time; for example, for a paste that is not demineralized and pasteurized (A1B2), the emulsion breakup time is much lower than when the gum has been demineralized and pasteurized (A2B2) at all three pH levels. Finally, with a demineralized, pasteurized gum at pH = 4.5, a gum with higher breaking stability is obtained (Table 7.15).

Table 7.14 Means and standard errors of the main effects on the model scale (Estimate) and the data scale (Mean)

(a) Demineralization least squares means							
Demineralization	Estimate	Standard error	DF	t-value	Pr > t	Mean	Standard error mean
1	5.0911	0.02930	4	173.77	<0.0001	162.57	4.7628
2	5.3379	0.02930	4	182.18	<0.0001	208.07	6.0964
(b) Pasteurization least squares means							
Pasteurization	Estimate	Standard error	DF	t-value	Pr > t	Mean	Standard error mean
1	5.1230	0.02930	4	174.87	<0.0001	167.84	4.9171
2	5.3059	0.02930	4	181.08	<0.0001	201.53	5.9051
(c) pH least squares means							
pH	Estimate	Standard error	DF	t-value	Pr > t	Mean	Standard error mean
1	5.1610	0.03050	8	169.22	<0.0001	174.33	5.3171
2	5.2839	0.03050	8	173.24	<0.0001	197.13	6.0124
3	5.1986	0.03052	8	170.32	<0.0001	181.02	5.5255

Table 7.15 Means and standard errors of the simple effects on the model scale (Estimate) and the data scale (Mean)

Demineralization*pasteurization*pH least squares means									
A	B	C	Estimate	Standard error	DF	t-value	Pr > t	Mean	Standard error mean
1	1	1	5.0696	0.06099	8	83.13	<0.0001	159.11	9.7035
1	1	2	5.1695	0.06105	8	84.68	<0.0001	175.83	10.7339
1	1	3	5.1311	0.06100	8	84.12	<0.0001	169.20	10.3204
1	2	1	5.1137	0.06099	8	83.84	<0.0001	166.28	10.1419
1	2	2	5.0445	0.06098	8	82.72	<0.0001	155.17	9.4623
1	2	3	5.0183	0.06098	8	82.29	<0.0001	151.15	9.2170
2	1	1	5.0811	0.06103	8	83.26	<0.0001	160.95	9.8225
2	1	2	5.1694	0.06099	8	84.76	<0.0001	175.81	10.7225
2	1	3	5.1175	0.06110	8	83.76	<0.0001	166.91	10.1978
2	2	1	5.3796	0.06100	8	88.19	<0.0001	216.93	13.2320
2	2	2	5.7520	0.06100	8	94.30	<0.0001	314.81	19.2031
2	2	3	5.5277	0.06106	8	90.53	<0.0001	251.57	15.3607

A = demineralization (1 = no, 2 = yes), B = pasteurization (1 = no, 2 = yes), and C = pH (1 = 2.5, 2 = 4.5, and 3 = 5.5)

7.3 Survival Analysis

When a research focuses on the time of occurrence of a specific event, we usually refer to survival times, and, hence, the statistical analysis of these times, as mentioned above, is known as survival analysis. A very characteristic feature of survival

times is the presence of censored times, that is, when there are individuals whose actual survival time is not known.

For a set of survival times (including censored ones) of a sample of individuals, it is possible to estimate the proportion of the population that will survive a time interval under the same circumstances. The methods used to make this estimate are based on the proposal of Kaplan and Meier (1958). This method allows – through different statistical tests (log rank, Breslow, Tarone–Ware, etc.) – the comparison of the survival of two or more groups of individuals who differ with respect to certain factors.

Survival analysis focuses its interest on a group or several groups of individuals for whom an event is defined, which occurs after a time interval. To determine the time of interest, there are three requirements: an initial time, a scale to measure the passage of time (minutes, hours, days, etc.), and clarity about what is meant by the event of interest.

Survival of an individual is conceptually the probability of being alive in a given time " t " from diagnosis, i.e., initiation of treatment or complete remission for a group of individuals. In clinical studies, survival times often refer to time till death, development of a particular symptom, or relapse after complete remission of a disease. Failure is defined as death, relapse, or the occurrence of a new disease. In many survival analyses, when the end of the observation period previously set by the investigator is reached, there are individuals to whom the event has not occurred and we do not know when it will occur. Therefore, the actual survival time for them is unknown, and only the survival time to the end of the study is known. Such survival times are called censored times. It also happens, in some cases, that some individuals do not continue the study until the end of the analysis period for reasons unrelated to the research, e.g., death from other causes; these times are also censored. These censored data contribute valuable information and, therefore, should not be omitted from the analysis.

The pharmaceutical and food industries are legally required to label the shelf life of their product on the packaging. For pharmaceuticals, the requirements for how to determine shelf life are highly regulated. However, the regulatory standards do not specifically define shelf life. Instead, the definition is implicit through the estimation procedure. The interest is in the situation where multiple batches are used to determine a shelf life of a product that applies to all future batches. Consequently, both shelf life and label life are of great importance because of the variability within and between batches. Product development must be very well thought out before a company can have confidence in shelf life estimates. The company must be able to reliably produce a homogeneous product from batch to batch of ingredients, as physical and chemical factors impact the ability of bacteria to grow, such as pH, water activity, and uniformity of the mix (moisture distribution, salt, preservative or food acid) and, consequently, the shelf life of the product. Therefore, products should be inspected at appropriate times and samples should be tested for critical stability of physical and chemical characteristics. These tests also provide an opportunity to begin microbiological testing for spoilage organisms. Testing should continue beyond the intended shelf life unless the product fails earlier. Testing

should lead to an understanding of target levels and ranges of ingredients for evaluation of the critical physical and chemical characteristics of the product over the intended shelf life.

Survival analysis is the name for a collection of statistical techniques used to describe and quantify the time in which the event of interest occurs. The term “survival time” specifies the amount of time taken to occur. Situations in which survival analyses have been used in epidemiology include:

- (a) Survival of insects after having received an insecticide.
- (b) The time taken by cows or ewes to conceive after calving.
- (c) The time taken for a farm to experience its first case of an exotic disease.

7.3.1 Concepts and Definitions

To clearly understand and interpret a rate of change calculated from the event data of interest, a more extensive approach is needed. The definition of a rate of change begins with the mathematical description of a changing pattern over time, represented by the symbol $S(t)$. A version of a ratio is created by dividing the change in function $S(t)$ [$S(t)$ to $S(t + \Delta t)$] by the corresponding change over time t (t to $t + \Delta t$) producing the rate of change

$$\text{rate of change} = \frac{\text{change on } S(t)}{\text{change on time}} = \frac{S(t) - S(t + \Delta t)}{(t + \Delta t) - t} = \frac{S(t) - S(t + \Delta t)}{\Delta t}$$

Rates of change, with respect to time, apply to a variety of situations, but one specific function, traditionally denoted by $S(t)$, is fundamental to the analysis of survival data. This is called the survival function and is defined as the probability of surviving (probability of survival) beyond a specific point in time (denoted by t). That is;

$$\begin{aligned} S(t) &= P(\text{survival time} = 0 \text{ at time} = t) \\ &= P(\text{survival in the interval } [0, t]) \end{aligned}$$

Equivalent to

$$S(t) = P(\text{surviving beyond time } t) = P(T \geq t) = 1 - F(t)$$

where $F(t)$ is the cumulative distribution function with $F(t) = P(T \leq t)$. Another important concept in survival analysis is the hazard function $h(t)$. The hazard function that depends on T is defined as

$$h(t) = \lim_{\Delta t \rightarrow 0} \left\{ \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t} \right\}$$

such that the following expression can be expressed as

$$h(t) = \lim_{\Delta t \rightarrow 0} \left\{ \frac{F(t + \Delta t) - F(t)}{\Delta t} \right\} \times \frac{1}{P(T \geq t)}$$

$$h(t) = \frac{f(t)}{S(t)}$$

where $f(t)$ is the probability density function. Any distribution defined by $t \in [0, t)$ can serve as a survival distribution. Consequently,

$$h(t) = - \frac{\partial}{\partial t} \{ \log S(t) \}.$$

It then follows that

$$S(t) = \exp\{ -H(t) \}$$

where $H(t)$ the cumulative hazard function

$$H(t) = \int_0^t h(u) du$$

Another useful relationship is

$$H(t) = - \log S(t).$$

For the simplest model, the exponential model with $h(t) = \lambda$ (λ is a constant), the survival function is given by

$$S(t) = \exp \left\{ - \int_0^t h(u) du \right\} = \exp - \int_0^t \lambda du = e^{-\lambda t}$$

with the probability density function given by

$$f(t) = \frac{\partial}{\partial t} S(t) = \lambda e^{-\lambda t}.$$

Thus, the survival function, hazard function, and cumulative risk for the exponential model is given by:

Survival function: $S(t) = e^{-\lambda t}$

Risk function: $h(t) = \frac{f(t)}{S(t)} = \frac{\lambda e^{-\lambda t}}{e^{-\lambda t}} = \lambda$

Cumulative risk function: $H(t) = \int_0^t h(u)du = \int_0^t \lambda du = \lambda t.$

7.3.2 CRD: *Aedes aegypti*

The objective of this experiment was to test the vulnerability of *Aedes aegypti* mosquitoes to different fungal treatments (four treatments). A bioassay was conducted to determine the survival time of each of the mosquitoes. Three-day-old mosquitoes were maintained after hatching in 45-cm rearing cages with access to water but not food. The mosquitoes were kept in rearing cages with water and fed warm pig blood (37 °C) through a natural membrane (sausage casing) approximately every 3 days and allowed to oviposit freely during the waiting period. A total of 10 mosquitoes were placed in a chamber to which one of the treatments (four) plus a control was applied. Here, we present part of the data from a bioassay with four replicates. The complete data from this trial can be found in the Appendix 1 (Data: *Aedes aegypti*).

Treatment	Rep	Y
C	1	8
C	1	11
⋮	⋮	⋮
C	4	20
Mam	1	2
Mam	1	2
⋮	⋮	⋮
MaS	1	3
MaS	1	3
⋮	⋮	⋮
MaC	1	2
MaC	1	2
MaC	1	2
⋮	⋮	⋮
Ma1	1	2
Ma1	1	2
⋮	⋮	⋮
Ma1	4	11

Table 7.16 Results of the analysis of variance

(a) Fit statistics for conditional distribution				
-2 Log L (T r. effects)				716.70
Pearson's chi-square				35.33
Pearson's chi-square/DF				0.18
(b) Type III tests of fixed effects				
Effect	Num DF	Den DF	F-value	Pr > F
Trt	4	192	186.42	<0.0001

The components of this GLMM are as follows:

$$\text{Distributions: } y_{ij} \mid \text{rep}_j \sim \text{Gamma}(\mu_{ij}, \phi)$$

$$\text{rep}_j \sim N\left(0, \sigma_{\text{rep}}^2\right)$$

$$\text{Linear predictor: } \eta_{ij} = \eta + \tau_i + \text{rep}_j$$

$$\text{Link function: } \eta_{ij} = \log(\mu_{ij})$$

where η is the intercept, τ_i is the treatment effect, and rep_j is the random effect due to the mosquito chamber assuming $\text{rep}_j \sim N\left(0, \sigma_{\text{rep}}^2\right)$.

The following GLIMMIX commands adjust a GLMM with a gamma response:

```
proc glimmix data=mosquitos method=laplace;
class bio trt rep;
model y = trt/dist=gamma;
random rep;
lsmeans trt/lines ilink;
run;
```

Part of the output is shown in Table 7.16. The statistic in (a) above indicates that there is no over-dispersion in the fit of the data, as indicated by Pearson's chi – square/DF = 0.18. The analysis of variance (type III tests of fixed effects) indicates that there is a highly significant effect ($P = 0.0001$) of the fungal treatments on the mean mosquito survival time.

The relevant information in Table 7.17 “lsmeans” comes from the columns labeled “Estimate” and “Mean”: these are the estimates on the model scale and the data scale, and the average survival time in each of the treatments is represented by $\hat{\mu}_i$ (\pm standard error).

The estimated risk function for each treatment combination is $\hat{\lambda}_i = 1/\hat{\mu}_i$. For example, for treatment Ma1, the estimated hazard function is $\hat{\lambda}_{Ma1} = 1/3.4223 = 0.2922$. We can manually calculate these values from the Mean column or we can automate the process by adding the command “ods output lsmeans = mu” in the GLIMMIX

Table 7.17 Means and standard errors of the main effects on the model scale (Estimate) and the data scale (Mean)

Trt least squares means							
Trt	Estimate	Standard error	DF	t-value	Pr > t	Mean	Standard error mean
Ma1	1.2303	0.06354	192	19.36	<0.0001	3.4223	0.2174
MaC	0.9562	0.06350	192	15.06	<0.0001	2.6017	0.1652
MaS	1.5798	0.06357	192	24.85	<0.0001	4.8542	0.3086
Mam	0.6946	0.06350	192	10.94	<0.0001	2.0029	0.1272
Control	2.7155	0.06362	192	42.68	<0.0001	15.1126	0.9615

program above. Once we have saved the treatment means, we can ask SAS to estimate the estimated hazard function for the treatments. The commands are as follows:

```
data hazard;
set mu;
hazard=1/mu;
proc print data=hazard;
run;
```

The results are listed below in Table 7.18. The hazard column contains the estimated hazard functions for each treatment $\hat{h}_i(t) = \hat{\lambda}_i$.

From the values $\hat{\lambda}_i$, we can calculate the estimated survival function $S_i(t) = e^{-\hat{\lambda}_i t}$ for each of the treatments. Figure 7.3 shows the probability of survival over time obtained with $S_i(t) = e^{-\hat{\lambda}_i t}$ of each of the proposed treatments and the control. Clearly, the treatments MaS, Ma1, MaC, and Mam showed a greater efficacy in the biological control of these mosquitoes.

7.3.3 RCBD: *Aedes aegypti*

Similar to the previous example, this experiment consisted of testing the vulnerability of *Aedes aegypti* mosquitoes to different fungal treatments (four treatments). For this, two bioassays were conducted to determine the survival time of each of the mosquitoes. Three-day-old mosquitoes were maintained after hatching in 45-cm rearing cages with access to water but not food. Mosquitoes were maintained in rearing cages with water and were fed warm pig blood (37 °C) through a natural membrane (sausage casing) approximately every 3 days. They were allowed to freely oviposit during the waiting period. A total of 10 mosquitoes were placed in a chamber to which one of the treatments (four) plus a control was applied. The data can be found in the Appendix 1 (Data: *Aedes aegypti*).

Table 7.18 Means and standard errors of the main effects on the model scale (Estimate) and the data scale (Mean) and the hazard function $\hat{\lambda}_i$

Effect	TRT	Estimate	Standard error	DF	t-value	Probt	Mean	Standard error mean	Hazard $\hat{\lambda}_i$
TRT	MaI	1.2303	0.06354	192	19.36	<0.0001	3.4223	0.2174	0.29220
TRT	MaC	0.9562	0.06350	192	15.06	<0.0001	2.6017	0.1652	0.38437
TRT	MaS	1.5798	0.06357	192	24.85	<0.0001	4.8542	0.3086	0.20601
TRT	Mam	0.6946	0.06350	192	10.94	<0.0001	2.0029	0.1272	0.49927
TRT	Control	2.7155	0.06362	192	42.68	<0.0001	15.1126	0.9615	0.06617

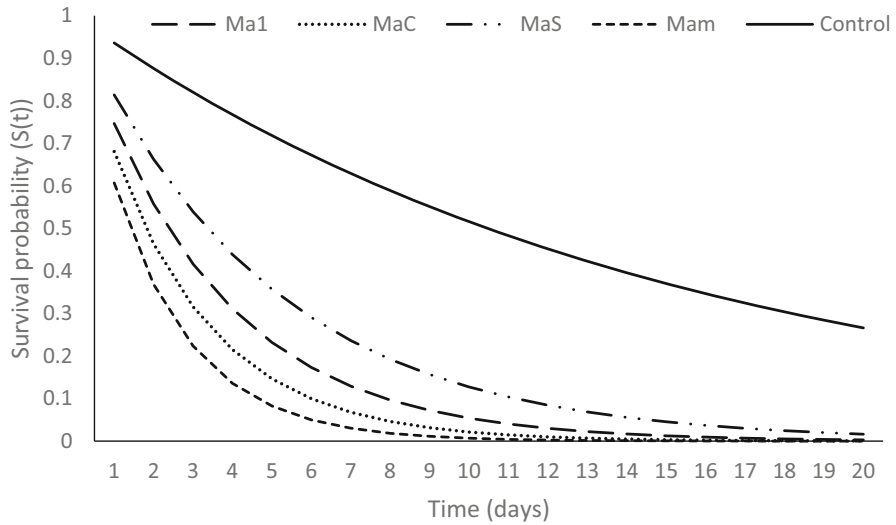


Fig. 7.3 Estimated survival probability for each treatment

The components of this GLMM are as follows:

$$\text{Distributions: } y_{ijk} \mid \text{bio}_j, \text{rep}(\text{bio})_{k(j)} \sim \text{Gamma}(\mu_{ijk}, \phi)$$

$$\text{bio}_j \sim N(0, \sigma_{\text{bio}}^2), \text{rep}(\text{bio})_{k(j)} \sim N(0, \sigma_{\text{rep}(\text{bio})}^2)$$

$$\text{Linear predictor: } \eta_{ij} = \eta + \tau_i + \text{bio}_j + \text{rep}(\text{bio})_{k(j)}$$

where η is the intercept, τ_i is the treatment effect, bio_j and $\text{rep}(\text{bio})_{k(j)}$ are the random effects of the bioassay and the mosquito chamber within the bioassay, respectively, assuming $\text{bio}_j \sim N(0, \sigma_{\text{bio}}^2)$ and $\text{rep}(\text{bio})_{k(j)} \sim N(0, \sigma_{\text{rep}(\text{bio})}^2)$.

$$\text{Link function: } \eta_{ij} = \log(\mu_{ij})$$

The following GLIMMIX program fits a block GLMM with a gamma response.

```
proc glimmix method=laplace nobound;
class bio trt ind rep;
model y = trt/dist=gamma;
random bio rep(bio);
ods output lsmeans=mu;
lsmeans trt/lines ilink;
run;quit;
```

The results obtained are shown below. Part of the statistics and variance components are listed in Table 7.19. In part (a), the value of the statistic of

Table 7.19 Results of the analysis of variance

(a) Fit statistics for conditional distribution				
-2 log L (Y r. effects)				3303.50
Pearson's chi-square				202.30
Pearson's chi-square/DF				0.34
(b) Cov Parm		Estimate	Standard error	
BIO		0.1859	0.1936	
REP(BIO)		0.02562	0.01673	
Residual		0.2822	0.01568	
(c) Type III tests of fixed effects				
Effect	Num DF	Den DF	F-value	Pr > F
TRAT	4	588	115.36	<0.0001

Table 7.20 Means and standard errors of the main effects on the model scale (Estimate) and the data scale (Mean)

TRT least squares means							
TRT	Estimate	Standard error	DF	t-value	Pr > t	Mean	Standard error mean
Ma1	1.6344	0.3140	588	5.21	<0.0001	5.1266	1.6097
MaC	1.4903	0.3140	588	4.75	<0.0001	4.4386	1.3939
MaS	1.8788	0.3140	588	5.98	<0.0001	6.5455	2.0550
Mam	1.8053	0.3143	588	5.74	<0.0001	6.0820	1.9115
Control	2.8293	0.3139	588	9.01	<0.0001	16.9329	5.3153

Table 7.21 Means and standard errors of the main effects on the model scale (Estimate), the data scale (Mean), and the hazard function $\hat{\lambda}_i$

TRT	Estimate	Standard error	DF	t-value	Probt	Mean	Standard error mean	Hazard $\hat{\lambda}_i$
Ma1	1.6344	0.3140	588	5.21	<0.0001	5.1266	1.6097	0.19506
MaC	1.4903	0.3140	588	4.75	<0.0001	4.4386	1.3939	0.22529
MaS	1.8788	0.3140	588	5.98	<0.0001	6.5455	2.0550	0.15278
Mam	1.8053	0.3143	588	5.74	<0.0001	6.0820	1.9115	0.16442
Control	2.8293	0.3139	588	9.01	<0.0001	16.9329	5.3153	0.05906

Pearson's chi - square/DF = 0.34 and in part (b), the estimated variance components due to blocks, within-block replicates, and experimental error are $\hat{\sigma}_{\text{bio}}^2 = 0.1859$, $\hat{\sigma}_{\text{rep}(\text{bio})}^2 = 0.02562$, and $\hat{\sigma}^2 = 0.2822$, respectively. The type III effect hypothesis tests (part (c)) indicate that there is a highly significant difference between treatments on the mean survival time, as indicated by $P = 0.0001$.

Tables 7.20 and 7.21 show the estimates on the model scale and the data scale, linear predictors ($\hat{\eta}_i$), means ($\hat{\mu}_i$) with their respective standard errors, and the estimated hazard function. The results indicate that the MaC treatment has a greater lethal effect than *A. aegypti* mosquito control.

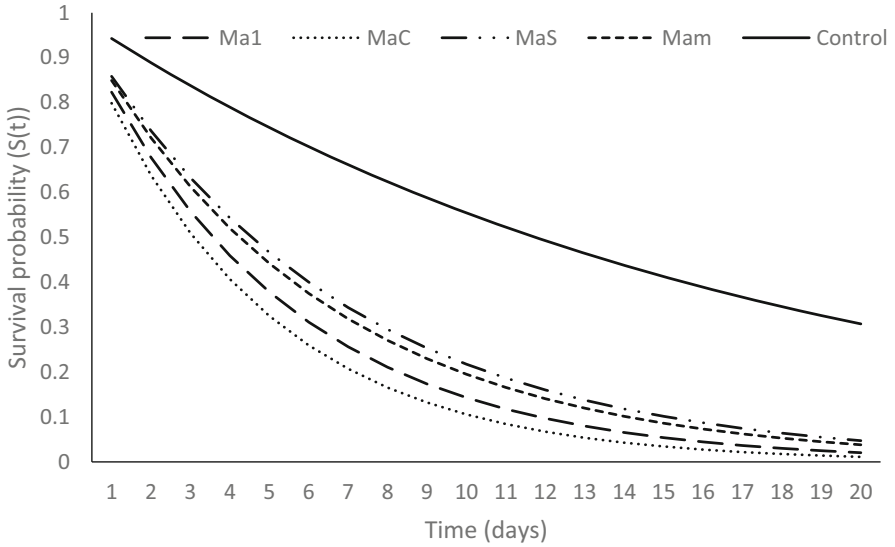


Fig. 7.4 Estimated survival probability for each treatment

Figure 7.4 shows the survival times for the different treatments tested. These curves were obtained with $S_i(t) = e^{(-\hat{\lambda}_i * t)}$.

7.4 Exercises

Exercise 7.4.1 The investigation of this experiment focused on studying the times of animal incapacitation experienced after being exposed to the burning of eight types of aircraft interior materials (M1–M9) and performances in milligram/gram combustion of seven gases (CO, HCN, H₂S, HCl, HBr, NO₂, SO₂) (Spurgeon 1978). The recorded incapacitation time of the animal when exposed to different combustion materials (under the column “Material”) is found under the column “Time in minutes” and in the third column the value of (1000/Time); these data are shown below (Table 7.22):

- (a) Write down a statistical model of this experiment.
- (b) List all the components of the GLMM in (a).
- (c) Write down the null and alternative hypotheses associated with this experiment.
- (d) Construct an ANOVA table indicating the sources of variation and degrees of freedom.
- (e) Analyze the time of inability of the animal to be exposed to the gases of the different types of materials.
- (f) Comment on the results obtained.

Table 7.22 Time of incapacity of the animal when exposed to different combustion gases

Material	Time	1000/Time	CO	HCN	H ₂ S	HCl	HBr	NO ₂	SO ₂
M1	2.36	423.7	164	6.4	0	0	0	0.26	0
M1	2.38	420.2	174	7.5	0	0	5	1.07	0
M1	2.61	383.1	96	4.7	0	33	5	0.08	0
M1	3.07	325.7	101	7.5	0	0	7.1	0.43	0
M1	3.07	325.7	142	6.8	0	27.6	0	0.25	0
M1	3.19	313.5	143	8.2	0	0	5.5	0.33	0
M1	3.7	270.3	147	5.2	0	11.3	0	0.37	0
M1	3.9	256.4	156	4.7	0	12	2.6	0.39	0
M1	4.18	239.2	124	3.2	0	23.3	0	0.2	0
M1	4.7	212.8	101	8.9	0.9	5.4	8	0.63	0
M1	4.86	205.8	142	4.6	0	19.4	4.1	0.19	0
M1	5.58	179.2	104	3.4	0	80	0	0.15	0.4
M1	5.85	170.9	90	2.3	0	34.4	0	0.09	1.2
M2	3.22	310.6	159	16.4	0	0	5.3	2	0
M2	3.89	257.1	153	2.9	0	0	6.6	0.15	0
M2	4.79	208.8	161	0.6	0	0	0	0.62	0
M2	5.07	197.2	159	0	0	4.6	1.7	0.04	0
M2	5.22	191.6	162	0	0	22	0	0.04	0
M2	5.82	171.8	106	3.2	0	45.2	15.6	0.08	0
M2	6.09	164.2	124	1.5	0	0	0	0.85	0
M2	8.36	119.6	89	0.7	0	0	5.3	0.29	0
M2	13.02	76.8	88	0	0	0	0	0.02	0
M3	4.29	233.1	129	6	0	4.2	0	0.02	0.7
M3	4.8	208.3	105	5.8	0	0	0	0.03	0
M3	5.04	198.4	108	7.8	0	7.3	0	0.04	0
M3	5.06	197.6	120	11.6	0	23	0	0.02	0
M3	5.25	190.5	149	0	0	8.6	0	0	0
M3	5.5	181.8	28	9.1	0.4	56.2	0	0	2.2
M3	5.55	180.2	83	5	0	0	0	0.02	0
M3	7.55	132.5	68	5.5	0	27.3	0	0.01	0.9
M3	9.58	104.4	28	2.4	2	137	0	0	16.6
M4	1.15	869.6	88	62.4	0	182	0	0.52	2.1
M4	2	500	89	41.7	13.4	0	0	0	0.3
M4	2.15	465.1	63	14.9	0	0	9.6	1.6	8.5
M4	2.22	450.5	112	37.2	14.2	0	20.5	0	1.5
M4	2.23	448.4	96	7	0	43.1	0	0.53	11.2
M4	2.72	367.6	78	33.8	13.9	0	0	0	0
M4	2.93	341.3	348	1.9	0	28	7.1	1	1.8
M4	3.07	325.7	255	1.9	0	0	0	0.57	0
M4	3.47	288.2	112	19.5	10.7	88	0	0.03	4.8
M4	4.18	239.2	144	3.8	0	14.5	5.1	0.39	0.9
M4	4.64	215.5	70	11.2	6.2	205	0	0.04	4.9

(continued)

Table 7.22 (continued)

Material	Time	1000/Time	CO	HCN	H ₂ S	HCl	HBr	NO ₂	SO ₂
M4	7.57	132.1	92	0	0.3	536	0	0.01	3
M5	6.97	143.5	114	0	0	114	0	0	0
M5	7.47	133.9	103	0	0	221	0	0	0
M5	10.7	93.5	70	0	0	259	0	0.02	1.4
M5	13.71	72.9	56	0	0	220	0	0.01	0.9
M6	4.94	202.4	94	6.7	0	0	0	0.32	0
M6	5.26	190.1	55	14.9	5.3	21.9	0	0	2.2
M6	5.53	180.8	46	13.5	6.1	24.9	0	0	2.5
M6	7.46	134	77	3.1	0	158	0	0.04	0
M6	9.84	101.6	52	4.1	0.7	19	0	0.01	1.4
M6	10.9	91.7	41	2.4	0	82	0	0	0
M7	3.7	270.3	398	0	0	0	21	0	0
M7	3.8	263.2	345	0	0	0	15.5	0.01	0
M7	3.83	261.1	406	0	0	0	47	0	0
M7	4.04	247.5	342	0	0	23	10.3	0.04	0
M7	5.19	192.7	196	0	0	0	0	0	0
M7	6.01	166.4	148	0	0.2	387	0	0.01	1.9
M7	7.56	132.3	86	0	0	0	47	0	0
M7	9.41	106.3	54	2.2	0	197	0	0	2.6
M7	9.59	104.3	55	1.7	0	321	0	0	1.1
M7	10.79	92.7	55	4.1	0	162	0	0.02	2.9
M8	3.7	270.3	0	15	0	0	0	0.34	0
M8	3.99	250.6	90	8.6	0	88	0	0.59	0
M8	6.56	152.4	37	3.1	0	27.7	0	0.01	0
M8	7.68	130.2	66	0	0	105	0	0	0
M8	9.16	109.2	45	0	0	0	0	0.01	0
M8	10.33	96.8	62	0	0	61	0	0.01	0
M8	12.26	81.6	31	2.7	0	0	0	0.22	0
M8	14.96	66.8	9	0	0	0	0	0.01	0

Exercise 7.4.2 Cockroaches are responsible for 80% of infestations in spaces used by humans. They associate with humans and have the ability to contaminate food with their feces and secretions, having both medical and economic implications. Different insecticides have been formulated, mainly synthetic, and, in some cases, have led to the development of cockroaches' resistance. This example deals with the study of survival in days (y) of this insect when exposed to two promising fungi in the biological control of this insect plus an already known control. The data for this example are shown below (Table 7.23):

Table 7.23 Results of the cockroach biological control experiment

Insect	Strain	Age	Time	Insect	Strain	Age	Time	Insect	Strain	Age	Time
1	Bb1	1	2	1	Bb2	1	2	1	Test	1	2
2	Bb1	1	3	2	Bb2	1	2	2	Test	1	20
3	Bb1	1	3	3	Bb2	1	2	3	Test	1	20
4	Bb1	1	3	4	Bb2	1	3	4	Test	1	20
5	Bb1	1	4	5	Bb2	1	3	5	Test	1	20
6	Bb1	1	7	6	Bb2	1	3	6	Test	1	20
7	Bb1	1	8	7	Bb2	1	3	7	Test	1	20
8	Bb1	1	9	8	Bb2	1	4	8	Test	1	20
9	Bb1	1	10	9	Bb2	1	5	9	Test	1	20
10	Bb1	1	10	10	Bb2	1	8	10	Test	1	20
11	Bb1	1	17	11	Bb2	1	9	11	Test	1	20
12	Bb1	1	19	12	Bb2	1	10	12	Test	1	20
13	Bb1	1	19	13	Bb2	1	11	13	Test	1	20
14	Bb1	1	20	14	Bb2	1	20	14	Test	1	20
15	Bb1	1	20	15	Bb2	1	20	15	Test	1	20
16	Bb1	1	20	16	Bb2	1	20	16	Test	1	20
17	Bb1	1	20	17	Bb2	1	20	17	Test	1	20
18	Bb1	1	20	18	Bb2	1	20	18	Test	1	20
19	Bb1	1	20	19	Bb2	1	20	19	Test	1	20
20	Bb1	1	20	20	Bb2	1	20	20	Test	1	20
21	Bb1	2	4	21	Bb2	2	3	21	Test	2	20
22	Bb1	2	13	22	Bb2	2	3	22	Test	2	20
23	Bb1	2	13	23	Bb2	2	4	23	Test	2	20
24	Bb1	2	13	24	Bb2	2	6	24	Test	2	20
25	Bb1	2	20	25	Bb2	2	8	25	Test	2	20
26	Bb1	2	20	26	Bb2	2	13	26	Test	2	20
27	Bb1	2	20	27	Bb2	2	17	27	Test	2	20
28	Bb1	2	20	28	Bb2	2	17	28	Test	2	20
29	Bb1	2	20	29	Bb2	2	19	29	Test	2	20
30	Bb1	2	20	30	Bb2	2	19	30	Test	2	20
31	Bb1	2	20	31	Bb2	2	20	31	Test	2	20
32	Bb1	2	20	32	Bb2	2	20	32	Test	2	20
33	Bb1	2	20	33	Bb2	2	20	33	Test	2	20
34	Bb1	2	20	34	Bb2	2	20	34	Test	2	20
35	Bb1	2	20	35	Bb2	2	20	35	Test	2	20
36	Bb1	2	20	36	Bb2	2	20	36	Test	2	20
37	Bb1	2	20	37	Bb2	2	20	37	Test	2	20
38	Bb1	2	20	38	Bb2	2	20	38	Test	2	20
39	Bb1	2	20	39	Bb2	2	20	39	Test	2	20
40	Bb1	2	20	40	Bb2	2	20	40	Test	2	20
41	Bb1	3	3	41	Bb2	3	2	41	Test	3	11
42	Bb1	3	3	42	Bb2	3	3	42	Test	3	20

(continued)

Table 7.23 (continued)

Insect	Strain	Age	Time	Insect	Strain	Age	Time	Insect	Strain	Age	Time
43	Bb1	3	4	43	Bb2	3	4	43	Test	3	20
44	Bb1	3	4	44	Bb2	3	5	44	Test	3	20
45	Bb1	3	6	45	Bb2	3	5	45	Test	3	20
46	Bb1	3	7	46	Bb2	3	10	46	Test	3	20
47	Bb1	3	8	47	Bb2	3	10	47	Test	3	20
48	Bb1	3	8	48	Bb2	3	10	48	Test	3	20
49	Bb1	3	9	49	Bb2	3	11	49	Test	3	20
50	Bb1	3	10	50	Bb2	3	13	50	Test	3	20
51	Bb1	3	13	51	Bb2	3	13	51	Test	3	20
52	Bb1	3	14	52	Bb2	3	13	52	Test	3	20
53	Bb1	3	16	53	Bb2	3	14	53	Test	3	20
54	Bb1	3	17	54	Bb2	3	15	54	Test	3	20
55	Bb1	3	17	55	Bb2	3	15	55	Test	3	20
56	Bb1	3	20	56	Bb2	3	15	56	Test	3	20
57	Bb1	3	20	57	Bb2	3	15	57	Test	3	20
58	Bb1	3	20	58	Bb2	3	19	58	Test	3	20
59	Bb1	3	20	59	Bb2	3	20	59	Test	3	20
60	Bb1	3	20	60	Bb2	3	20	60	Test	3	20

- (a) Write down a statistical model of this experiment.
- (b) List all components of the GLMM from (a).
- (c) Write down the null and alternative hypotheses associated with this experiment.
- (d) Analyze the survival time of the insect when infected with the different types of fungi.
- (e) Comment on the results obtained.

Exercise 7.4.3 Consider a study on the effect of analgesic treatments (Trt) in elderly patients with neuralgia. Two test treatments (A and B) and a placebo (P) are compared. The response variable is whether the patient reported pain or not (yes = 1, $n = 0$). The investigators recorded the age (E) and sex (S) of 60 patients and the duration (time = T) in which the pain disappeared after starting the treatment. The data are presented in the Table 7.24 below.

- (a) List all components of the GLMM for this exercise.
- (b) Write down the null and alternative hypotheses associated with this experiment.
- (c) Construct an ANOVA table indicating the sources of variation and degrees of freedom.
- (d) Analyze the average time during which the patient experiences pain after starting the treatment. Are there any significant differences?
- (e) Comment on the results obtained.

Table 7.24 Results with neuralgia patients (Trt = Treatment, *S* = Sex, *E* = Age, *T* = Time, *D* = Pain with yes = 1 and no = 0)

Trt	<i>S</i>	<i>E</i>	<i>T</i>	<i>D</i>	Tr	<i>S</i>	<i>E</i>	<i>T</i>	<i>D</i>	Tr	<i>S</i>	<i>E</i>	<i>T</i>	<i>D</i>
P	F	68	1	0	B	M	74	16	0	P	F	67	30	0
P	M	66	26	1	B	F	67	28	0	B	F	77	16	0
A	F	71	12	0	B	F	72	50	0	B	F	76	9	1
A	M	71	17	1	A	F	63	27	0	A	F	69	18	1
B	F	66	12	0	A	M	62	42	0	P	F	64	1	1
A	F	64	17	0	P	M	74	4	0	A	F	72	25	0
P	M	70	1	1	B	M	66	19	0	B	M	59	29	0
A	F	64	30	0	A	M	70	28	0	A	M	69	1	0
B	F	78	1	0	P	M	83	1	1	B	F	69	42	0
B	M	75	30	1	P	M	77	29	1	P	F	79	20	1
A	M	70	12	0	A	F	69	12	0	B	F	65	14	0
B	M	70	1	0	B	M	67	23	0	A	M	76	25	1
P	M	78	12	1	B	M	77	1	1	B	F	69	24	0
P	M	66	4	1	P	F	65	29	0	P	M	60	26	1
A	M	78	15	1	B	M	75	21	1	A	F	67	11	0
P	F	72	27	0	P	F	70	13	1	A	M	75	6	1
B	F	65	7	0	P	F	68	27	1	P	M	68	11	1
P	M	67	17	1	B	M	70	22	0	A	M	65	15	0
P	F	67	1	1	A	M	67	10	0	P	F	72	11	1
A	F	74	1	0	B	M	80	21	1	A	F	69	3	0

Exercise 7.4.4 Refer to the previous exercise and perform an analysis of covariance.

- (a) List the linear predictor of this experiment.
- (b) Analyze the average time during which the patient experiences pain after starting the treatment using an analysis of covariance. Are there any significant differences?
- (c) Comment on the results obtained. Your results differ from those obtained in the previous year.

Appendix 1

Data: Onset and duration of estrus in Pelibuey ewes (age in weeks, weight in kilograms, Inestro = number of days from the onset of estrus, Durestro = number of days in the duration of estrus)

Animal	Birth type	Age	Weight	CC	Inestro	Durestro
1	1	18.5096	52.5	4	28	4
2	1	18.4438	47.4	4	28	4
3	1	19.3973	50.2	4	16	20
4	1	19.3973	53.6	4	28	16

(continued)

Animal	Birth type	Age	Weight	CC	Inestro	Durestro
5	1	9.4356	47.5	4	28	4
6	1	18.674	41.3	3	28	4
7	1	20.0877	60.5	5	28	4
8	1	19.5616	49.4	4	28	4
9	1	19.5288	53.4	4	28	4
10	1	19.7589	52.6	5	28	4
11	1	29.9507	35.5	3	28	4
12	1	19.3644	50.5	4	28	4
13	1	19.0027	62.2	5	28	4
14	1	18.3452	54.7	4	28	4
15	1	20.0877	48.7	4	28	4
1	2	40.7671	40.5	3	28	4
2	2	51.189	49.3	4	12	8
3	2	40.0767	38.1	3	20	20
4	2	54.3123	41.9	3	24	8
5	2	52.274	58	4	28	4
6	2	53.6219	34.8	3	28	4
7	2	40.2082	40.3	2	24	8
8	2	36.4932	34.6	2	28	4
9	2	50.6301	42.1	2	28	4
10	2	51.0247	52.6	4	28	4
11	2	46.389	32.1	2	20	12
12	2	50.7945	40	2	16	16
13	2	30.411	37.9	2	24	8
14	2	30.5096	42.2	3	20	20
15	2	50.6959	33.2	2	24	16
16	2	36.6247	34.2	3	20	20
17	2	30.5425	39	2	12	32
18	2	36.6247	33.7	2	24	16
19	2	29.9507	32.9	2	24	16
20	2	47.211	39.5	2	32	12
21	2	40.2082	57.5	5	12	32
22	2	52.2411	53.3	4	12	28
23	2	53.4247	43.4	3	12	32
24	2	55.5616	46	3	24	16
25	2	30.5425	31.6	2	24	16
26	2	29.0959	47.8	3	20	20
27	2	40.1425	36	2	20	20
28	2	50.7945	42.2	3	24	16
29	2	37.6767	44.3	3	24	16
30	2	36.4274	43.1	2	20	20
31	2	30.5425	38	2	20	20

(continued)

Animal	Birth type	Age	Weight	CC	Inestro	Durestro
1	3	68.9753	42.9	2	24	8
2	3	63.1233	44	4	24	8
3	3	68.7781	38.5	3	20	12
4	3	64.3068	48	4	24	8
5	3	68.6795	40.1	2	20	12
6	3	62.6301	46.3	3	32	4
7	3	69.8959	32.5	2	20	12
8	3	69.6	42.8	3	20	12
9	3	63.4849	51.3	4	24	8
10	3	64.274	47.7	3	24	16
11	3	63.5178	44.5	3	12	28
12	3	78.7397	38	2	12	28
13	3	64.537	52.5	4	12	28
14	3	62.4329	41.2	2	12	28
15	3	67.6603	50.8	4	20	20
16	3	63.7151	48.2	3	20	24
17	3	74.4986	33.3	2	32	8
18	3	63.6493	45.1	3	24	16
19	3	72.9205	33	3	24	20
20	3	69.4027	40.4	3	24	16
21	3	69.9616	43.3	3	12	28
22	3	69.6	43.2	2	24	16
23	3	63.4849	51	4	24	16
24	3	63.6164	57.4	4	24	16
25	3	67.8575	43	3	24	16
26	3	63.6822	49.7	4	24	16
27	3	65.7534	40.1	3	24	16
28	3	67.989	33.4	1	20	20
29	3	61.1836	51.6	4	20	20
30	3	63.3534	43.3	3	20	20
31	3	79.8904	44.7	3	24	16
32	3	63.7151	37.9	3	20	20

Data: Beetles

Dose	Insecticide	Rep	Frac	Time
Low	Insec1	1	0.31	3.1
Low	Insec2	1	0.82	8.2
Low	Insec3	1	0.43	4.3
Low	Insec4	1	0.45	4.5
Medium	Insec1	1	0.36	3.6
Medium	Insec2	1	0.92	9.2
Medium	Insec3	1	0.44	4.4
Medium	Insec4	1	0.56	5.6
High	Insec1	1	0.22	2.2

(continued)

Dose	Insecticide	Rep	Frac	Time
High	Insec2	1	0.3	3
High	Insec3	1	0.23	2.3
High	Insec4	1	0.3	3
Low	Insec1	2	0.45	4.5
Low	Insec2	2	1.1	11
Low	Insec3	2	0.45	4.5
Low	Insec4	2	0.71	7.1
Medium	Insec1	2	0.29	2.9
Medium	Insec2	2	0.61	6.1
Medium	Insec3	2	0.35	3.5
Medium	Insec4	2	1.02	10.2
High	Insec1	2	0.21	2.1
High	Insec2	2	0.37	3.7
High	Insec3	2	0.25	2.5
High	Insec4	2	0.36	3.6
Low	Insec1	3	0.46	4.6
Low	Insec2	3	0.88	8.8
Low	Insec3	3	0.63	6.3
Low	Insec4	3	0.66	6.6
Medium	Insec1	3	0.4	4
Medium	Insec2	3	0.49	4.9
Medium	Insec3	3	0.31	3.1
Medium	Insec4	3	0.71	7.1
High	Insec1	3	0.18	1.8
High	Insec2	3	0.38	3.8
High	Insec3	3	0.24	2.4
High	Insec4	3	0.31	3.1
Low	Insec1	4	0.43	4.3
Low	Insec2	4	0.72	7.2
Low	Insec3	4	0.76	7.6
Low	Insec4	4	0.62	6.2
Medium	Insec1	4	0.23	2.3
Medium	Insec2	4	1.24	12.4
Medium	Insec3	4	0.4	4
Medium	Insec4	4	0.38	3.8
High	Insec1	4	0.23	2.3
High	Insec2	4	0.29	2.9
High	Insec3	4	0.22	2.2
High	Insec4	4	0.33	3.3

Data: Rubber break time

Block	Demineralization	Pasteurization	pH	y
1	2	2	1	198.5
1	2	2	2	299
1	2	2	3	223.1
1	1	1	1	166.6
1	1	1	2	196.5
1	1	1	3	178.9
1	1	2	1	160.7
1	1	2	2	151.1
1	1	2	3	146.5
1	2	1	1	146.3
1	2	1	2	169.3
1	2	1	3	198.1
2	2	2	1	236.3
2	2	2	2	330.7
2	2	2	3	281.2
2	1	1	1	151.8
2	1	1	2	156
2	1	1	3	159.7
2	1	2	1	171.8
2	1	2	2	159.3
2	1	2	3	155.9
2	2	1	1	175.2
2	2	1	2	182.2
2	2	1	3	136.2

Data: *Aedes aegypti* (Trt = treatment, Rep = repetition, Y = survival time)

Trt	Rep	Y	Trt	Rep	Y	Trt	Rep	Y	Trt	Rep	Y	Trt	Rep	Y
Control	1	8	Mam	1	2	MaS	1	3	MaC	1	2	Ma1	1	2
Control	1	11	Mam	1	2	MaS	1	3	MaC	1	2	Ma1	1	2
Control	1	11	Mam	1	2	MaS	1	3	MaC	1	2	Ma1	1	2
Control	1	11	Mam	1	2	MaS	1	3	MaC	1	2	Ma1	1	2
Control	1	11	Mam	1	2	MaS	1	4	MaC	1	3	Ma1	1	3
Control	1	11	Mam	1	2	MaS	1	5	MaC	1	3	Ma1	1	3
Control	1	13	Mam	1	2	MaS	1	6	MaC	1	3	Ma1	1	3
Control	1	13	Mam	1	2	MaS	1	6	MaC	1	3	Ma1	1	3
Control	1	14	Mam	1	2	MaS	1	9	MaC	1	3	Ma1	1	6
Control	1	20	Mam	1	2	MaS	1	12	MaC	1	4	Ma1	1	12
Control	2	8	Mam	2	2	MaS	2	3	MaC	2	2	Ma1	2	2
Control	2	11	Mam	2	2	MaS	2	3	MaC	2	2	Ma1	2	2
Control	2	11	Mam	2	2	MaS	2	3	MaC	2	2	Ma1	2	2
Control	2	11	Mam	2	2	MaS	2	3	MaC	2	2	Ma1	2	3

(continued)

Trt	Rep	Y	Trt	Rep	Y	Trt	Rep	Y	Trt	Rep	Y	Trt	Rep	Y
Control	2	11	Mam	2	2	MaS	2	3	MaC	2	2	Ma1	2	3
Control	2	11	Mam	2	2	MaS	2	3	MaC	2	2	Ma1	2	3
Control	2	15	Mam	2	2	MaS	2	3	MaC	2	3	Ma1	2	3
Control	2	15	Mam	2	2	MaS	2	4	MaC	2	3	Ma1	2	4
Control	2	15	Mam	2	2	MaS	2	5	MaC	2	3	Ma1	2	4
Control	2	16	Mam	2	2	MaS	2	6	MaC	2	4	Ma1	2	4
Control	3	11	Mam	3	2	MaS	3	3	MaC	3	2	Ma1	3	2
Control	3	11	Mam	3	2	MaS	3	3	MaC	3	2	Ma1	3	2
Control	3	11	Mam	3	2	MaS	3	3	MaC	3	2	Ma1	3	2
Control	3	11	Mam	3	2	MaS	3	2	MaC	3	2	Ma1	3	2
Control	3	23	Mam	3	2	MaS	3	2	MaC	3	3	Ma1	3	3
Control	3	25	Mam	3	2	MaS	3	5	MaC	3	3	Ma1	3	3
Control	3	26	Mam	3	2	MaS	3	5	MaC	3	3	Ma1	3	3
Control	3	27	Mam	3	2	MaS	3	6	MaC	3	3	Ma1	3	3
Control	3	30	Mam	3	2	MaS	3	10	MaC	3	4	Ma1	3	4
Control	3	30	Mam	3	2	MaS	3	12	MaC	3	4	Ma1	3	4
Control	4	8	Mam	4	2	MaS	4	3	MaC	4	2	Ma1	4	2
Control	4	8	Mam	4	2	MaS	4	3	MaC	4	2	Ma1	4	2
Control	4	11	Mam	4	2	MaS	4	3	MaC	4	2	Ma1	4	2
Control	4	13	Mam	4	2	MaS	4	4	MaC	4	2	Ma1	4	3
Control	4	14	Mam	4	2	MaS	4	4	MaC	4	2	Ma1	4	3
Control	4	19	Mam	4	2	MaS	4	5	MaC	4	2	Ma1	4	3
Control	4	20	Mam	4	2	MaS	4	5	MaC	4	3	Ma1	4	4
Control	4	20	Mam	4	2	MaS	4	6	MaC	4	3	Ma1	4	5
Control	4	20	Mam	4	2	MaS	4	9	MaC	4	3	Ma1	4	6
Control	4	22	Mam	4	2	MaS	4	12	MaC	4	3	Ma1	4	11

Data: *Aedes aegypti* (Bio = bioassay, Trt = treatment, Rep = repetition, Y = survival time)

Bio	Trt	Rep	Y	Bio	Trt	Rep	Y	Bio	Trt	Rep	Y
B1	C	1	8	B1	MaS	3	3	B2	C	1	5
B1	C	1	11	B1	MaS	3	3	B2	C	1	7
B1	C	1	11	B1	MaS	3	3	B2	C	1	8
B1	C	1	11	B1	MaS	3	2	B2	C	1	8
B1	C	1	11	B1	MaS	3	2	B2	C	1	10
B1	C	1	11	B1	MaS	3	5	B2	C	1	13
B1	C	1	13	B1	MaS	3	5	B2	C	1	14
B1	C	1	13	B1	MaS	3	6	B2	C	1	16
B1	C	1	14	B1	MaS	3	10	B2	C	1	20
B1	C	1	20	B1	MaS	3	12	B2	C	1	22
B1	C	2	8	B1	MaS	4	3	B2	C	1	22
B1	C	2	11	B1	MaS	4	3	B2	C	1	23
B1	C	2	11	B1	MaS	4	3	B2	C	1	23
B1	C	2	11	B1	MaS	4	4	B2	C	1	23

(continued)

Bio	Trt	Rep	Y	Bio	Trt	Rep	Y	Bio	Trt	Rep	Y
B1	C	2	11	B1	MaS	4	4	B2	C	1	24
B1	C	2	11	B1	MaS	4	5	B2	C	1	24
B1	C	2	15	B1	MaS	4	5	B2	C	1	24
B1	C	2	15	B1	MaS	4	6	B2	C	1	24
B1	C	2	15	B1	MaS	4	9	B2	C	1	28
B1	C	2	16	B1	MaS	4	12	B2	C	1	28
B1	C	3	11	B1	MaC	1	2	B2	C	2	10
B1	C	3	11	B1	MaC	1	2	B2	C	2	11
B1	C	3	11	B1	MaC	1	2	B2	C	2	11
B1	C	3	11	B1	MaC	1	2	B2	C	2	12
B1	C	3	23	B1	MaC	1	3	B2	C	2	12
B1	C	3	25	B1	MaC	1	3	B2	C	2	15
B1	C	3	26	B1	MaC	1	3	B2	C	2	15
B1	C	3	27	B1	MaC	1	3	B2	C	2	16
B1	C	3	30	B1	MaC	1	3	B2	C	2	16
B1	C	3	30	B1	MaC	1	4	B2	C	2	16
B1	C	4	8	B1	MaC	2	2	B2	C	2	16
B1	C	4	8	B1	MaC	2	2	B2	C	2	18
B1	C	4	11	B1	MaC	2	2	B2	C	2	19
B1	C	4	13	B1	MaC	2	2	B2	C	2	27
B1	C	4	14	B1	MaC	2	2	B2	C	2	27
B1	C	4	19	B1	MaC	2	2	B2	C	2	27
B1	C	4	20	B1	MaC	2	3	B2	C	2	27
B1	C	4	20	B1	MaC	2	3	B2	C	2	27
B1	C	4	20	B1	MaC	2	3	B2	C	2	28
B1	C	4	22	B1	MaC	2	4	B2	C	2	28
B1	Mam	1	2	B1	MaC	3	2	B2	C	3	16
B1	Mam	1	2	B1	MaC	3	2	B2	C	3	19
B1	Mam	1	2	B1	MaC	3	2	B2	C	3	19
B1	Mam	1	2	B1	MaC	3	2	B2	C	3	19
B1	Mam	1	2	B1	MaC	3	3	B2	C	3	19
B1	Mam	1	2	B1	MaC	3	3	B2	C	3	19
B1	Mam	1	2	B1	MaC	3	3	B2	C	3	25
B1	Mam	1	2	B1	MaC	3	3	B2	C	3	25
B1	Mam	1	2	B1	MaC	3	4	B2	C	3	26
B1	Mam	1	2	B1	MaC	3	4	B2	C	3	28
B1	Mam	2	2	B1	MaC	4	2	B2	C	3	28
B1	Mam	2	2	B1	MaC	4	2	B2	C	3	28
B1	Mam	2	2	B1	MaC	4	2	B2	C	3	28
B1	Mam	2	2	B1	MaC	4	2	B2	C	3	28
B1	Mam	2	2	B1	MaC	4	2	B2	C	3	28
B1	Mam	2	2	B1	MaC	4	2	B2	C	3	28
B1	Mam	2	2	B1	MaC	4	3	B2	C	3	28

(continued)

Bio	Trt	Rep	Y	Bio	Trt	Rep	Y	Bio	Trt	Rep	Y
B1	Mam	2	2	B1	MaC	4	3	B2	C	3	28
B1	Mam	2	2	B1	MaC	4	3	B2	C	3	28
B1	Mam	2	2	B1	MaC	4	3	B2	C	3	28
B1	Mam	3	2	B1	Ma1	1	2	B2	C	4	16
B1	Mam	3	2	B1	Ma1	1	2	B2	C	4	17
B1	Mam	3	2	B1	Ma1	1	2	B2	C	4	17
B1	Mam	3	2	B1	Ma1	1	2	B2	C	4	17
B1	Mam	3	2	B1	Ma1	1	3	B2	C	4	17
B1	Mam	3	2	B1	Ma1	1	3	B2	C	4	17
B1	Mam	3	2	B1	Ma1	1	3	B2	C	4	17
B1	Mam	3	2	B1	Ma1	1	3	B2	C	4	19
B1	Mam	3	2	B1	Ma1	1	6	B2	C	4	28
B1	Mam	3	2	B1	Ma1	1	12	B2	C	4	28
B1	Mam	4	2	B1	Ma1	2	2	B2	C	4	28
B1	Mam	4	2	B1	Ma1	2	2	B2	C	4	28
B1	Mam	4	2	B1	Ma1	2	3	B2	C	4	28
B1	Mam	4	2	B1	Ma1	2	3	B2	C	4	28
B1	Mam	4	2	B1	Ma1	2	3	B2	C	4	28
B1	Mam	4	2	B1	Ma1	2	3	B2	C	4	28
B1	Mam	4	2	B1	Ma1	2	4	B2	C	4	28
B1	Mam	4	2	B1	Ma1	2	4	B2	C	4	28
B1	Mam	4	2	B1	Ma1	2	4	B2	C	4	28
B1	MaS	1	3	B1	Ma1	3	2	B2	Mam	1	2
B1	MaS	1	3	B1	Ma1	3	2	B2	Mam	1	3
B1	MaS	1	3	B1	Ma1	3	2	B2	Mam	1	3
B1	MaS	1	3	B1	Ma1	3	2	B2	Mam	1	4
B1	MaS	1	4	B1	Ma1	3	3	B2	Mam	1	4
B1	MaS	1	5	B1	Ma1	3	3	B2	Mam	1	4
B1	MaS	1	6	B1	Ma1	3	3	B2	Mam	1	5
B1	MaS	1	6	B1	Ma1	3	3	B2	Mam	1	5
B1	MaS	1	9	B1	Ma1	3	4	B2	Mam	1	5
B1	MaS	1	12	B1	Ma1	3	4	B2	Mam	1	6
B1	MaS	2	3	B1	Ma1	4	2	B2	Mam	1	6
B1	MaS	2	3	B1	Ma1	4	2	B2	Mam	1	6
B1	MaS	2	3	B1	Ma1	4	2	B2	Mam	1	7
B1	MaS	2	3	B1	Ma1	4	3	B2	Mam	1	15
B1	MaS	2	3	B1	Ma1	4	3	B2	Mam	1	17
B1	MaS	2	3	B1	Ma1	4	3	B2	Mam	1	21
B1	MaS	2	3	B1	Ma1	4	4	B2	Mam	1	25
B1	MaS	2	4	B1	Ma1	4	5	B2	Mam	1	28
B1	MaS	2	5	B1	Ma1	4	6	B2	Mam	1	28
B1	MaS	2	6	B1	Ma1	4	11	B2	Mam	1	28

(continued)

Bio	Trt	Rep	Y	Bio	Trt	Rep	Y	Bio	Trt	Rep	Y
B2	Mam	2	2	B2	MaS	2	18	B2	MaC	3	13
B2	Mam	2	2	B2	MaS	3	5	B2	MaC	3	23
B2	Mam	2	2	B2	MaS	3	5	B2	MaC	4	2
B2	Mam	2	3	B2	MaS	3	5	B2	MaC	4	2
B2	Mam	2	3	B2	MaS	3	9	B2	MaC	4	3
B2	Mam	2	3	B2	MaS	3	10	B2	MaC	4	6
B2	Mam	2	4	B2	MaS	3	10	B2	MaC	4	6
B2	Mam	2	7	B2	MaS	3	10	B2	MaC	4	6
B2	Mam	2	11	B2	MaS	3	12	B2	MaC	4	8
B2	Mam	2	11	B2	MaS	3	12	B2	MaC	4	8
B2	Mam	2	11	B2	MaS	3	12	B2	MaC	4	8
B2	Mam	2	13	B2	MaS	3	12	B2	MaC	4	8
B2	Mam	2	13	B2	MaS	3	12	B2	MaC	4	9
B2	Mam	2	14	B2	MaS	3	14	B2	MaC	4	9
B2	Mam	2	14	B2	MaS	3	15	B2	MaC	4	9
B2	Mam	2	14	B2	MaS	3	18	B2	MaC	4	9
B2	Mam	2	15	B2	MaS	3	18	B2	MaC	4	10
B2	Mam	2	16	B2	MaS	3	23	B2	MaC	4	10
B2	Mam	2	16	B2	MaS	3	25	B2	MaC	4	12
B2	Mam	2	23	B2	MaS	3	25	B2	MaC	4	19
B2	Mam	3	3	B2	MaS	3	25	B2	MaC	4	20
B2	Mam	3	3	B2	MaS	4	5	B2	MaC	4	24
B2	Mam	3	5	B2	MaS	4	5	B2	Ma1	1	2
B2	Mam	3	6	B2	MaS	4	6	B2	Ma1	1	2
B2	Mam	3	8	B2	MaS	4	6	B2	Ma1	1	3
B2	Mam	3	8	B2	MaS	4	6	B2	Ma1	1	3
B2	Mam	3	10	B2	MaS	4	6	B2	Ma1	1	3
B2	Mam	3	10	B2	MaS	4	7	B2	Ma1	1	4
B2	Mam	3	11	B2	MaS	4	8	B2	Ma1	1	4
B2	Mam	3	11	B2	MaS	4	8	B2	Ma1	1	5
B2	Mam	3	11	B2	MaS	4	9	B2	Ma1	1	5
B2	Mam	3	12	B2	MaS	4	10	B2	Ma1	1	5
B2	Mam	3	17	B2	MaS	4	10	B2	Ma1	1	6
B2	Mam	3	17	B2	MaS	4	10	B2	Ma1	1	6
B2	Mam	3	17	B2	MaS	4	11	B2	Ma1	1	6
B2	Mam	3	17	B2	MaS	4	11	B2	Ma1	1	7
B2	Mam	3	17	B2	MaS	4	11	B2	Ma1	1	8
B2	Mam	3	17	B2	MaS	4	11	B2	Ma1	1	8
B2	Mam	3	18	B2	MaS	4	11	B2	Ma1	1	8
B2	Mam	3	25	B2	MaS	4	11	B2	Ma1	1	10
B2	Mam	4	4	B2	MaS	4	24	B2	Ma1	1	17
B2	Mam	4	4	B2	MaC	1	2	B2	Ma1	1	21
B2	Mam	4	5	B2	MaC	1	2	B2	Ma1	2	2

(continued)

Bio	Trt	Rep	Y	Bio	Trt	Rep	Y	Bio	Trt	Rep	Y
B2	Mam	4	7	B2	MaC	1	2	B2	Ma1	2	2
B2	Mam	4	9	B2	MaC	1	2	B2	Ma1	2	2
B2	Mam	4	10	B2	MaC	1	2	B2	Ma1	2	3
B2	Mam	4	12	B2	MaC	1	2	B2	Ma1	2	3
B2	Mam	4	12	B2	MaC	1	3	B2	Ma1	2	3
B2	Mam	4	12	B2	MaC	1	3	B2	Ma1	2	4
B2	Mam	4	12	B2	MaC	1	3	B2	Ma1	2	5
B2	Mam	4	12	B2	MaC	1	3	B2	Ma1	2	6
B2	Mam	4	12	B2	MaC	1	3	B2	Ma1	2	7
B2	Mam	4	13	B2	MaC	1	4	B2	Ma1	2	7
B2	Mam	4	13	B2	MaC	1	4	B2	Ma1	2	7
B2	Mam	4	13	B2	MaC	1	5	B2	Ma1	2	8
B2	Mam	4	18	B2	MaC	1	5	B2	Ma1	2	9
B2	Mam	4	27	B2	MaC	1	7	B2	Ma1	2	9
B2	Mam	4	27	B2	MaC	1	7	B2	Ma1	2	10
B2	Mam	4	27	B2	MaC	1	9	B2	Ma1	2	10
B2	Mam	4	27	B2	MaC	1	9	B2	Ma1	2	10
B2	MaS	1	2	B2	MaC	1	10	B2	Ma1	2	10
B2	MaS	1	2	B2	MaC	2	2	B2	Ma1	2	10
B2	MaS	1	2	B2	MaC	2	2	B2	Ma1	3	2
B2	MaS	1	2	B2	MaC	2	2	B2	Ma1	3	3
B2	MaS	1	3	B2	MaC	2	3	B2	Ma1	3	3
B2	MaS	1	3	B2	MaC	2	3	B2	Ma1	3	3
B2	MaS	1	3	B2	MaC	2	3	B2	Ma1	3	5
B2	MaS	1	3	B2	MaC	2	3	B2	Ma1	3	7
B2	MaS	1	4	B2	MaC	2	3	B2	Ma1	3	7
B2	MaS	1	5	B2	MaC	2	5	B2	Ma1	3	7
B2	MaS	1	5	B2	MaC	2	6	B2	Ma1	3	8
B2	MaS	1	5	B2	MaC	2	6	B2	Ma1	3	8
B2	MaS	1	6	B2	MaC	2	6	B2	Ma1	3	8
B2	MaS	1	6	B2	MaC	2	7	B2	Ma1	3	9
B2	MaS	1	8	B2	MaC	2	7	B2	Ma1	3	10
B2	MaS	1	8	B2	MaC	2	7	B2	Ma1	3	10
B2	MaS	1	9	B2	MaC	2	9	B2	Ma1	3	10
B2	MaS	1	11	B2	MaC	2	10	B2	Ma1	3	10
B2	MaS	1	13	B2	MaC	2	10	B2	Ma1	3	10
B2	MaS	1	21	B2	MaC	2	18	B2	Ma1	3	10
B2	MaS	2	3	B2	MaC	2	19	B2	Ma1	3	11
B2	MaS	2	3	B2	MaC	3	2	B2	Ma1	3	17
B2	MaS	2	4	B2	MaC	3	3	B2	Ma1	4	4
B2	MaS	2	6	B2	MaC	3	6	B2	Ma1	4	5
B2	MaS	2	6	B2	MaC	3	6	B2	Ma1	4	8
B2	MaS	2	8	B2	MaC	3	8	B2	Ma1	4	8

(continued)

Bio	Trt	Rep	Y	Bio	Trt	Rep	Y	Bio	Trt	Rep	Y
B2	MaS	2	8	B2	MaC	3	8	B2	Ma1	4	8
B2	MaS	2	9	B2	MaC	3	9	B2	Ma1	4	9
B2	MaS	2	9	B2	MaC	3	9	B2	Ma1	4	9
B2	MaS	2	10	B2	MaC	3	9	B2	Ma1	4	11
B2	MaS	2	10	B2	MaC	3	9	B2	Ma1	4	11
B2	MaS	2	11	B2	MaC	3	9	B2	Ma1	4	11
B2	MaS	2	11	B2	MaC	3	10	B2	Ma1	4	11
B2	MaS	2	11	B2	MaC	3	10	B2	Ma1	4	11
B2	MaS	2	11	B2	MaC	3	10	B2	Ma1	4	12
B2	MaS	2	11	B2	MaC	3	10	B2	Ma1	4	12
B2	MaS	2	12	B2	MaC	3	10	B2	Ma1	4	13
B2	MaS	2	12	B2	MaC	3	11	B2	Ma1	4	13
B2	MaS	2	12	B2	MaC	3	13	B2	Ma1	4	13
								B2	Ma1	4	13
								B2	Ma1	4	14
								B2	Ma1	4	18

Data: Pelibuey Sheep

Animal	Birthtype	Age	Weight	Inestro	Durestro
1	1	18.509589	52.5	28	4
2	1	18.4438356	47.4	28	4
3	1	19.3972603	50.2	16	20
4	1	19.3972603	53.6	28	16
5	1	9.43561644	47.5	28	4
6	1	18.6739726	41.3	28	4
7	1	20.0876712	60.5	28	4
8	1	19.5616438	49.4	28	4
9	1	19.5287671	53.4	28	4
10	1	19.7589041	52.6	28	4
11	1	29.9506849	35.5	28	4
12	1	19.3643836	50.5	28	4
13	1	19.0027397	62.2	28	4
14	1	18.3452055	54.7	28	4
15	1	20.0876712	48.7	28	4
1	2	40.7671233	40.5	28	4
2	2	51.1890411	49.3	12	8
3	2	40.0767123	38.1	20	20
4	2	54.3123288	41.9	24	8
5	2	52.2739726	58	28	4
6	2	53.6219178	34.8	28	4
7	2	40.2082192	40.3	24	8
8	2	36.4931507	34.6	28	4
9	2	50.630137	42.1	28	4

(continued)

Animal	Birthtype	Age	Weight	Inestro	Durestro
10	2	51.0246575	52.6	28	4
11	2	46.3890411	32.1	20	12
12	2	50.7945206	40	16	16
13	2	30.4109589	37.9	24	8
14	2	30.509589	42.2	20	20
15	2	50.6958904	33.2	24	16
16	2	36.6246575	34.2	20	20
17	2	30.5424658	39	12	32
18	2	36.6246575	33.7	24	16
19	2	29.9506849	32.9	24	16
20	2	47.2109589	39.5	32	12
21	2	40.2082192	57.5	12	32
22	2	52.2410959	53.3	12	28
23	2	53.4246575	43.4	12	32
24	2	55.5616438	46	24	16
25	2	30.5424658	31.6	24	16
26	2	29.0958904	47.8	20	20
27	2	40.1424658	36	20	20
28	2	50.7945206	42.2	24	16
29	2	37.6767123	44.3	24	16
30	2	36.4273973	43.1	20	20
31	2	30.5424658	38	20	20
1	3	68.9753425	42.9	24	8
2	3	63.1232877	44	24	8
3	3	68.7780822	38.5	20	12
4	3	64.3068493	48	24	8
5	3	68.6794521	40.1	20	12
6	3	62.630137	46.3	32	4
7	3	69.8958904	32.5	20	12
8	3	69.6	42.8	20	12
9	3	63.4849315	51.3	24	8
10	3	64.2739726	47.7	24	16
11	3	63.5178082	44.5	12	28
12	3	78.739726	38	12	28
13	3	64.5369863	52.5	12	28
14	3	62.4328767	41.2	12	28
15	3	67.660274	50.8	20	20
16	3	63.7150685	48.2	20	24
17	3	74.4986301	33.3	32	8
18	3	63.6493151	45.1	24	16
19	3	72.920548	33	24	20
20	3	69.4027397	40.4	24	16
21	3	69.9616438	43.3	12	28

(continued)

Animal	Birthtype	Age	Weight	Inestro	Durestro
22	3	69.6	43.2	24	16
23	3	63.4849315	51	24	16
24	3	63.6164384	57.4	24	16
25	3	67.8575343	43	24	16
26	3	63.6821918	49.7	24	16
27	3	65.7534247	40.1	24	16
28	3	67.9890411	33.4	20	20
29	3	61.1835616	51.6	20	20
30	3	63.3534247	43.3	20	20
31	3	79.890411	44.7	24	16
32	3	63.7150685	37.9	20	20

Data: Gum Breakdown Times

Batch	Demineralization	Pasteurization	pH	Time
1	2	2	1	198.5
1	2	2	2	299
1	2	2	3	223.1
1	1	1	1	166.6
1	1	1	2	196.5
1	1	1	3	178.9
1	1	2	1	160.7
1	1	2	2	151.1
1	1	2	3	146.5
1	2	1	1	146.3
1	2	1	2	169.3
1	2	1	3	198.1
2	2	2	1	236.3
2	2	2	2	330.7
2	2	2	3	281.2
2	1	1	1	151.8
2	1	1	2	156
2	1	1	3	159.7
2	1	2	1	171.8
2	1	2	2	159.3
2	1	2	3	155.9
2	2	1	1	175.2
2	2	1	2	182.2
2	2	1	3	136.2

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



Chapter 8

Generalized Linear Mixed Models for Categorical and Ordinal Responses



8.1 Introduction

According to Agresti (2013), a multinomial distribution is a generalization of a binomial distribution in cases with more than two possible ordered (ordinal) or unordered (nominal) outcomes. Given a response with more than two possible outcomes and independent trials with probabilities of similar category for each trial, the distribution of counts across categories follows a multinomial distribution. Quinn and Keough (2002) believe that several methods exist for multinomial data analysis. The most common form of categorical data analysis in biological sciences, which results in frequency counts, is creating cross-tabulations or contingency tables and chi-squared tests to examine associations between two or more categorical variables. However, such an approach is ill suited for a study aimed at estimating the response when there is a change in the explanatory variable(s), as contingency tables are used to analyze the association between variables without considering a predictor or response variable. In this analysis, the results are valid as long as less than 20% of the cells have an expected count less than five and none are less than one (Logan 2010). Fisher's exact test extends the chi-squared test in studies involving small sample sizes.

There are several methods for modeling multinomial data; traditional methods of multinomial data analysis include frequency analysis (counts), which uses the chi-squared test and the log-linear model for contingency tables. This chapter focuses on describing multinomial logit and probit models in detail.

8.2 Concepts and Definitions

For the multinomial distribution each observation drawn from a total of N observations belongs to exactly one of the mutually and exclusive $c = 1, \dots, C$ categories and each category has a probability π_c ($c = 1, \dots, C$) of belonging to the category c . A multinomial distribution refers to the probability that exactly one randomly sampled observation from the population belongs to category y_1 , that is, it belongs to category 1, y_2 observations belong to category 2, and so forth up to category C , where $\sum_{c=1}^C y_c = N$ and $\sum_{c=1}^C \pi_c = 1$. The density function of this distribution is equal to

$$f(y_1, y_2, \dots, y_C) = \frac{N!}{y_1! y_2! \dots y_C!} \pi_1^{y_1} \pi_2^{y_2} \dots \pi_C^{y_C}$$

Multinomial models are applied in data analysis where the categorical response variable has more than two possible outcomes while the independent variables can be continuous, categorical, or both (Hosmer and Lemeshow 2000). The categorical response variable can be either ordinal (ordered) or nominal (unordered). Ordinal response variables are single values that represent a rank order on some dimension, but there are not enough values to be treated as a continuous variable. Nominal (unordered) response variables are those whose values provide a rank but do not provide an indication of order. Models for multinomial data are constructed in a similar way as for binomial data. The link functions used in these types of models are similar to the logit and probit functions used for binomial data. Cumulative logit and cumulative probit models define the link function such that when properly fitted to the data, they allow for parsimonious modeling of ordinal or multinomial data. Generalized logit and probit models do not require ordered categories and are therefore suitable for multinomial nominal data.

In terms of generalized linear models (GLMs) and generalized linear mixed models (GLMMs), a multinomial distribution with C categories requires $C - 1$ link functions to fully specify a model that relates the response probabilities $(\pi_1, \pi_2, \dots, \pi_C)$ to the linear predictor. The commonly used models are the cumulative logit model, also known as the proportional odds model proposed by McCullagh (1980), and the cumulative probit model, also known as the threshold model. Throughout this chapter, we will use either of these two link functions interchangeably.

The link functions for a cumulative logit model with C categories are

$$\begin{aligned}
 \eta_1 &= \log\left(\frac{\pi_1}{1 - \pi_1}\right) = \eta_1 + \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b} \\
 \eta_2 &= \log\left(\frac{\pi_1 + \pi_2}{1 - (\pi_1 + \pi_2)}\right) = \eta_2 + \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b} \\
 &\quad \vdots \\
 \eta_{C-1} &= \log\left(\frac{\pi_1 + \pi_2 + \dots + \pi_{C-1}}{1 - (\pi_1 + \pi_2 + \dots + \pi_{C-1})}\right) = \eta_{C-1} + \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b}
 \end{aligned}$$

where \mathbf{X} and \mathbf{Z} are the design matrices, whereas $\boldsymbol{\beta}$ and \mathbf{b} are the vectors of fixed and random effects parameters, respectively. The inverse links of each of the functions are as follows:

$$\begin{aligned}
 \pi_1 &= \frac{1}{1 + e^{-\eta_1}} = h(\eta_1) \\
 \pi_1 + \pi_2 &= \frac{1}{1 + e^{-\eta_2}} = h(\eta_2) \\
 &\quad \vdots \\
 \pi_1 + \pi_2 + \dots + \pi_{C-1} &= \frac{1}{1 + e^{-\eta_{C-1}}} = h(\eta_{C-1}).
 \end{aligned}$$

Once $h(\eta_1), h(\eta_2), \dots, h(\eta_{C-1})$ have been estimated, we can then estimate the probabilities $\hat{\pi}_1, \hat{\pi}_2, \dots, \hat{\pi}_c$.

8.3 Cumulative Logit Models (Proportional Odds Models)

Multinomial logit models are used to model the relationships between a polytomous response variable and a set of predictor variables. These polytomous response models can be classified – as mentioned above – into two different types, depending on whether the response variable has an ordered or an unordered structure.

In a proportional odds model, the covariates (linear predictor $\boldsymbol{\eta}$) have the same effect on the probabilities that the response variable has in any category when considering different values of the covariates, thus shifting the response distribution to the right (or left) without changing the shape of the distribution. In a proportional odds model, the cumulative logits model the effect of the covariates on the response probabilities below or equal to the category cutoff.

A multinomial logit model assumes independence of categories, which implies that the probabilities of choosing a category c relative to a category c' are independent of the category characteristics of c and c' for $c \neq c'$. The assumption requires that if a new category is available, then the prior probabilities are precisely adjusted to preserve the original probabilities between all pairs of outcomes. The proportional odds model employs a strict assumption that the odds ratio does not depend on the category, and, therefore, we need to test the proportional odds assumption, which is also called the “parallel regression assumption.”

8.3.1 Complete Randomize Design (CRD) with a Multinomial Response: Ordinal

Data are obtained from an experiment related to red core disease in strawberries, which is caused by the fungus *Phytophthora fragariae*. In this example, 12 strawberry populations were evaluated in a completely randomized experiment with 4 replications (Table 8.1). Plots generally consisted of 10 plants; in some cases, only 9 plants were observed. At the end of the experiment, each plant was assigned to one of three ordered categories representing fungal damage (1 = no damage, 2 = moderate damage, and 3 = severe damage).

A total of 12 populations were obtained by crossing 3 genotypes of male parents with 4 genotypes of female parents. The variation between and within plots is considered minimal, whereas the genetic and nongenetic effects are more significant, as plants from the same cross are not genetically identical.

The model that fits these data for the cumulative probabilities is a GLMM, which exhibit a classification effect on the treatment variable (population resulting from crossing genotypes). Thus, the GLMM for multinomial ordered outcomes with C categories requires $C - 1$ link function equations to fully specify the model that relates the response probabilities $(\pi_1, \pi_2, \dots, \pi_C)$ to the linear predictor η_{ij} (Stroup 2013). The $C - 1$ multinomial logit equations are tested against each of the remaining categories 1, 2, . . . , $C - 1$.

Table 8.1 Evaluation of red core disease in strawberry plants

Repetition		1			2			3			4		
		Disease category											
Parent plant male/female		1	2	3	1	2	3	1	2	3	1	2	3
1	1	0	3	6	2	2	6	2	3	5	2	5	3
1	2	2	3	5	0	3	7	4	6	0	2	3	5
1	3	3	4	3	7	2	1	1	1	7	2	3	5
1	4	0	5	5	5	4	1	2	8	0	1	4	5
2	1	1	4	4	2	2	6	1	2	7	1	5	4
2	2	1	4	5	3	4	2	1	6	3	4	2	4
2	3	4	3	3	5	1	4	3	3	4	4	2	4
2	4	1	4	5	1	2	6	8	5	0	2	5	3
3	1	0	0	9	3	5	2	2	5	3	0	0	10
3	2	5	3	2	3	2	5	3	6	1	2	1	7
3	3	0	3	6	2	5	3	1	3	6	0	3	7
3	4	3	0	7	5	2	3	7	3	0	3	4	3

Table 8.3 Results of the multinomial analysis of variance for injury level in strawberry plants

(a) Covariance parameter estimates				
Cov Parm	Subject	Estimate	Standard error	
Intercept	Rep	0.1453	0.1437	
(b) Type III tests of fixed effects				
Effect	Num degree of freedom (DF)	Den DF	F-value	Pr > F
Trt	11	457	2.60	0.0032

Table 8.4 Fixed effects solution for injury categories

Solutions for fixed effects							
Effect	Cat	Trt	Estimate	Standard error	DF	t-value	Pr > t
Intercept	Without ($\hat{\eta}_1$)		-0.4571	0.3526	3	-1.30	0.2855
Intercept	Moderate ($\hat{\eta}_2$)		1.0631	0.3558	3	2.99	0.0582
Trt	($\hat{\tau}_1$)	M1H1	-1.1456	0.4264	457	-2.69	0.0075
Trt	($\hat{\tau}_2$)	M1H2	-0.8355	0.4179	457	-2.00	0.0462
Trt	($\hat{\tau}_3$)	M1H3	-0.4621	0.4171	457	-1.11	0.2685
Trt	($\hat{\tau}_4$)	M1H4	-0.4716	0.4145	457	-1.14	0.2558
Trt	($\hat{\tau}_5$)	M2H1	-1.2644	0.4295	457	-2.94	0.0034
Trt	($\hat{\tau}_6$)	M2H2	-0.6060	0.4181	457	-1.45	0.1479
Trt	($\hat{\tau}_7$)	M2H3	-0.2332	0.4140	457	-0.56	0.5735
Trt	($\hat{\tau}_8$)	M2H4	-0.3912	0.4168	457	-0.94	0.3484
Trt	($\hat{\tau}_9$)	M3H1	-1.5563	0.4393	457	-3.54	0.0004
Trt	($\hat{\tau}_{10}$)	M3H2	-0.4508	0.4144	457	-1.09	0.2772
Trt	($\hat{\tau}_{11}$)	M3H3	-1.4426	0.4350	457	-3.32	0.0010
Trt	($\hat{\tau}_{12}$)	M3H4	0

population1 (M1H1 = trt) “Without” damage and “Moderate” damage when exposed to the fungus (*Phytophthora fragariae*). Part of the output is presented in Table 8.3.

The estimated variance component (part (a)) due to plants is $\hat{\sigma}_r^2 = 0.1453$, whereas the hypothesis tests for type III effects (part (b)) (“Type III tests of fixed effects”) indicate that the crosses have different significant tolerance levels to fungal attacks ($Pr > F = P = 0.0032$). The results of the fixed effects solution, obtained by specifying the “solution” option in the model, are shown in Table 8.4.

From the fixed effects solution, we can estimate the linear predictors for the two categories of each treatment, which are in terms of the model scale. For example, for treatment 1, the first category of injury $\hat{\eta}_{11} = \hat{\eta}_1 + \hat{\tau}_1 = -0.4571 + (-1.1456) = -1.6027$, where $\hat{\eta}_1$ defines the boundary between the categories “Without” damage and “Moderate” damage and $\hat{\eta}_2$ defines the boundary between the categories “Moderate” damage and “Severe” damage, and the linear predictor is $\hat{\eta}_{11} = \hat{\eta}_2 + \hat{\tau}_1 = 1.0631 + (-1.1456) = -0.0825$. Note that for the proportional odds, the τ_i values are not category-specific; treatment effects move the boundaries as a group.

Table 8.5 Estimated odds ratio

Odds ratio estimates					
Trt	_Trt	Estimate	DF	95% Confidence limits	
M1H1	M3H4	0.318	457	0.138	0.735
M1H2	M3H4	0.434	457	0.191	0.986
M1H3	M3H4	0.630	457	0.278	1.430
M1H4	M3H4	0.624	457	0.276	1.409
M2H1	M3H4	0.282	457	0.121	0.657
M2H2	M3H4	0.546	457	0.240	1.241
M2H3	M3H4	0.792	457	0.351	1.787
M2H4	M3H4	0.676	457	0.298	1.534
M3H1	M3H4	0.211	457	0.089	0.500
M3H2	M3H4	0.637	457	0.282	1.438
M3H3	M3H4	0.236	457	0.101	0.556

The odds ratio (Table 8.5) is the result of taking $e^{\hat{\tau}_i}$ for crosses 1–12. Since odds ratios are not specific to a particular category, this value is the same for all three categories and hence the name odds ratio.

In Table 8.6, we show the maximum likelihood estimates of the linear predictors $\hat{\eta}_{ci} = \hat{\eta}_C + \hat{\tau}_i$ in the “Estimate” column, in terms of the model scale, as well as the means on the data scale for each of the categories of the treatments tested (“Mean”).

Thus, for $c = 1, t = 1$ (response category “Without” damage and treatment 1), the estimator is $\hat{\eta}_{11} = -1.6027$ and for $c = 2, t = 1$ (“Moderate” damage and treatment 1), the linear predictor is $\hat{\eta}_{21} = -0.0825$. Taking the inverse of the link function yields the probability of $\hat{\pi}_{11} = 1/(1 + e^{1.6027}) = 0.1676$. This is the estimated probability for which the cross (treatment) M1H1 has a response score of “Without damage.” This inverse value is presented under the “Mean” column (Table 8.6).

Now, for $c = 2, t = 1$, the inverse of the link yields the following probability: $\hat{\pi}_{11} + \hat{\pi}_{21} = 1/(1 + e^{0.0825}) = 0.4794$ (cumulative probability). From this value, we deduce the probability of observing a “Moderate” damage and a “Severe” damage in the plant of the cross M1H1. For “Moderate” damage, the probability is $\hat{\pi}_{21} = 0.4794 - \hat{\pi}_{11} = 0.4794 - 0.1676 = 0.3118$, and, for “Severe” damage, it is $\hat{\pi}_{31} = 1 - \hat{\pi}_{11} + \hat{\pi}_{21} = 1 - 0.4794 = 0.5206$. Similarly, the rest of the probabilities in the different crosses are estimated.

8.3.2 Randomized Complete Block Design (RCBD) with a Multinomial Response: Ordinal

In recent years, poultry production has become conscious of animal welfare, which is associated with bird mortality, behavior, and health, among others (Stanley 1981; Martrenchar et al. 2002). One of the diseases related to animal welfare is footpad dermatitis, and, among many repercussions, it affects a bird’s ability to walk (Bilgili et al. 2009). Pododermatitis is known as contact dermatitis or footpad dermatitis and

Table 8.6 Estimates on the model scale (Estimate) and on the data scale (Mean) for the damage categories in strawberry plants

Estimates							
Label	Estimate	Standard error	DF	t-value	Pr > t	Mean	Standard error mean
$c = 1, t = 1$	-1.6027	0.3706	457	-4.33	<0.0001	0.1676	0.05170
$c = 2, t = 1$	-0.08254	0.3625	457	-0.23	0.8200	0.4794	0.09047
$c = 1, t = 2$	-1.2926	0.3597	457	-3.59	0.0004	0.2154	0.06080
$c = 2, t = 2$	0.2276	0.3542	457	0.64	0.5208	0.5567	0.08741
$c = 1, t = 3$	-0.9191	0.3572	457	-2.57	0.0104	0.2851	0.07281
$c = 2, t = 3$	0.6010	0.3555	457	1.69	0.0916	0.6459	0.08131
$c = 1, t = 4$	-0.9286	0.3542	457	-2.62	0.0090	0.2832	0.07190
$c = 2, t = 4$	0.5915	0.3524	457	1.68	0.0939	0.6437	0.08081
$c = 1, t = 5$	-1.7214	0.3744	457	-4.60	<0.0001	0.1517	0.04818
$c = 2, t = 5$	-0.2013	0.3656	457	-0.55	0.5822	0.4499	0.09047
$c = 1, t = 6$	-1.0631	0.3590	457	-2.96	0.0032	0.2567	0.06850
$c = 2, t = 6$	0.4571	0.3557	457	1.28	0.1995	0.6123	0.08444
$c = 1, t = 7$	-0.6903	0.3526	457	-1.96	0.0509	0.3340	0.07842
$c = 2, t = 7$	0.8299	0.3533	457	2.35	0.0193	0.6963	0.07471
$c = 1, t = 8$	-0.8483	0.3566	457	-2.38	0.0178	0.2998	0.07485
$c = 2, t = 8$	0.6719	0.3556	457	1.89	0.0595	0.6619	0.07958
$c = 1, t = 9$	-2.0133	0.3864	457	-5.21	<0.0001	0.1178	0.04016
$c = 2, t = 9$	-0.4932	0.3759	457	-1.31	0.1902	0.3791	0.08849
$c = 1, t = 10$	-0.9079	0.3540	457	-2.56	0.0106	0.2874	0.07250
$c = 2, t = 10$	0.6123	0.3524	457	1.74	0.0830	0.6485	0.08033
$c = 1, t = 11$	-1.8997	0.3813	457	-4.98	<0.0001	0.1301	0.04317
$c = 2, t = 11$	-0.3795	0.3714	457	-1.02	0.3074	0.4062	0.08958
$c = 1, t = 12$	-0.4571	0.3526	457	-1.30	0.1955	0.3877	0.08369
$c = 2, t = 12$	1.0631	0.3558	457	2.99	0.0030	0.7433	0.06789

is characterized by inflammation and necrotic lesions from the plantar surface to deep within the footpads of chicken. Deep ulcers may result in abscesses and in the thickening of the underlying tissues and structures (Greene et al. 1985).

Chicken feet have great economic importance because they are in high demand in the foreign market, mainly in Southeast Asia and China; however, due to diseases or alterations such as pododermatitis, there are significant economic losses since diseased feet are not suitable for human consumption and this, subsequently, reflects in market prices (Taira et al. 2014). Due to the economic importance of this product, Garcia et al. (2010) have focused on studying the factors that cause this disease and

Table 8.7 Treatment design

Treatment	Features
Trt1	Traditional program +1 kg m ⁻² of rice husks
Trt2	Traditional program +2 kg m ⁻² of rice husks
Trt3	Traditional program + podal health program +1 kg m ⁻² of rice husks
Trt4	Traditional program + podal health program +2 kg m ⁻² of rice husks

on finding strategies to reduce leg and carcass lesions in poultry. Important factors in broiler fattening are the type of litter, litter height, nutrition and feeding programs, and bird health, among others.

The objective of this study was to evaluate the effect of litter density and organic minerals (Availa Zn and Availa Mn), with an extract of *Yucca schidigera* (Micro-Aid) as a supplement to a traditional fattening program, on the development of footpad dermatitis in broilers. The genetic material used in this experiment was mainly male Ross line chickens. The traditional broiler fattening program by the poultry farm consists of three phases: a starter diet (1–18 days), a grower diet (19–35 days), and a finisher diet (36–50 days), applied for a period of 50 days, where rice husk is used as bedding material at a density of 1 kg m⁻². In this research, a foot health program was implemented in addition to the traditional fattening program, which included the addition of 125 ppm of Micro-Aid (*Yucca schidigera* extract), 40 ppm of Availa Zn, and 40 ppm of Availa Mn to the fattening diet.

Based on the above information, four treatments were evaluated at two poultry farms, as described below:

- Treatment 1 involved the application of the company's traditional fattening program (Trt1).
- Treatment 2 was the company's traditional fattening program plus an increase in litter density from 1 to 2 kg m⁻² (Trt2).
- Treatment 3 was the traditional fattening program plus the implementation of the foot health program during the fattening period until completion (Trt3).
- Treatment 4 consisted of the traditional fattening program plus the implementation of the foot health program and an increase in litter density from 1 to 2 kg m⁻² (Trt4). The following table lists the treatments studied (Table 8.7):

The response variable evaluated was the degree of foot lesion (pododermatitis) at the end of the fattening period (50 days). The response variable was evaluated on 1250 chickens per treatment. The degree of a footpad lesion was determined according to a visual guide for lesions in chickens based on the method of De Jong and Guémené (2012). This method entails defining three grades: grade 0 is attributed to legs with no lesions, grade one is if lesions exist in some areas of the footpad (<50%), and grade two is if the leg has extensive lesions in areas of the footpad (50–100%). Table 8.8 shows the dataset indicating the block, treatment, level of lesion, and the number of birds observed with a given lesion (frequency).

Table 8.8 Pododermatitis in broilers

Block	Trt	Category	Frequency	Block	Trt	Category	Frequency
1	1	Without	26	1	3	Without	54
1	1	Slight	58	1	3	Slight	43
1	1	Severe	17	1	3	Severe	3
2	1	Without	37	2	3	Without	25
2	1	Slight	56	2	3	Slight	69
2	1	Severe	6	2	3	Severe	7
1	2	Without	40	1	4	Without	65
1	2	Slight	57	1	4	Slight	34
1	2	Severe	3	1	4	Severe	1
2	2	Without	77	2	4	Without	63
2	2	Slight	23	2	4	Slight	36
2	2	Severe	0	2	4	Severe	0

Note: Without stands for no lesion, slight stands for moderate lesion, and severe stands for severe lesion

The GLMM for multinomial ordered results with C categories requires $C - 1$ link function equations instead of one to fully specify a model that relates the response probabilities $(\pi_1, \pi_2, \dots, \pi_C)$ to the linear predictor η_{ij} (Stroup 2013). The $C - 1$ multinomial logit equations are tested against each of the categories 1, 2, ..., $C - 1$.

The link functions for the cumulative logit model to describe the response variable with C categories are as follows:

$$\begin{aligned} \eta_{(1)ij} &= \log\left(\frac{\pi_{1ij}}{1 - \pi_{1ij}}\right) = \eta_1 + \tau_i + b_j \\ \eta_{(2)ij} &= \log\left(\frac{\pi_{1ij} + \pi_{2ij}}{1 - (\pi_{1ij} + \pi_{2ij})}\right) = \eta_2 + \tau_i + b_j \\ &\vdots \\ \eta_{(C-1)ij} &= \log\left(\frac{\pi_{1ij} + \pi_{2ij} + \dots + \pi_{(C-1)ij}}{1 - (\pi_{1ij} + \pi_{2ij} + \dots + \pi_{(C-1)ij})}\right) = \eta_{C-1} + \tau_i + b_j \end{aligned}$$

The components of the GLMM with an ordinal multinomial response variable are as follows:

Distributions: $y_{oij}, y_{1ij}, y_{2ij} | b_j \sim \text{Multinomial}(N_{ij}, \pi_{0ij}, \pi_{1ij}, \pi_{2ij})$, where y_{oij}, y_{1ij} , and y_{2ij} are the observed frequencies of the responses (paw injury) in each category (none, mild, and severe) and b_j is the random effect due to block assuming $b_j \sim N(0, \sigma_b^2)$.

Linear predictor: $\eta_{(c)ij} = \eta_c + \tau_i + b_j$, where $\eta_{(c)ij}$ is c th link ($c = 0, 1$) for processing i and block j , η_c is the intercept for the c th link, τ_i is the fixed effect due to the i th treatment, and b_j is the random effect due to the j th block ($b_j \sim N(0, \sigma_b^2)$). The link functions for each category are as follows:

$$\log\left(\frac{\pi_{0ij}}{1 - \pi_{0ij}}\right) = \eta_{(0)ij}$$

$$\log\left(\frac{\pi_{0ij} + \pi_{1ij}}{1 - (\pi_{0ij} + \pi_{1ij})}\right) = \eta_{(1)ij}$$

The following GLIMMIX commands fit a cumulative logit model with an ordinal multinomial response.

```
proc glimmix data=multinomial_ord;
class block trt;
model categoria (order=data)= trt/dist=Multinomial link=clogit
solution oddsratio (DIFF=LAST LABEL);
random intercept/subject=block;
estimate 'c=0, t=1' intercept 1 0 trt 1 0,
'c=1, t=1' intercept 0 1 trt 1 0,
'c=0, t=2' intercept 1 0 trt 0 1 0,
'c=1, t=2' intercept 0 1 trt 0 1 0,
'c=0, t=3' intercept 1 0 trt 0 0 1 0,
'c=1, t=3' intercept 0 1 trt 0 0 1 0,
'c=0, t=4' intercept 0 1 trt 0 0 0 1,
'c=1, t=4' intercept 1 0 trt 0 0 0 1/ilink;
freq y;
run;
```

The data should have one column for block, treatment, lesion category, and frequency or number of observations (Y), which, in this case, is referenced by the variables block, trt, category, and frequency, respectively.

Most of the options in the above syntax have already been explained previously; the “order = data” option specifies that the order in which the categories appear in the dataset will be treated as ordinal categories from the lowest to the highest for the analysis. If this option is not used with the response variable in the model specification, “proc GLIMMIX” will rearrange its categories in an alphabetical or numerical order, but this will depend on whether the categories are entered as a number or a name. The “freq y” option orders GLIMMIX to use y as the number of observations in the corresponding category. The “estimate” command specifies the estimable functions that form the boundaries between categories of each of the four treatments. For example, the first estimate “ $c = 0, t = 1$ ” defines $\eta_0 + \tau_1$, that is, the boundary between the categories “Without” (no lesion) and “Moderate” (slight lesion) for treatment 1. This first estimate corresponds to logit $\log\left(\frac{\pi_{01}}{1 - \pi_{01}}\right)$, which is the probability that a chicken that received treatment 1 will respond to a degree of lesion classified under category 0 (no lesion). The second estimation “ $c = 1, t = 1$ ” defines $\eta_1 + \tau_1$, that is, the boundary between the categories “Moderate” (slight lesion) and

Table 8.9 Results of the analysis of variance in the multinomial cumulative logit model

(a) Type III tests of fixed effects				
Effect	Num DF	Den DF	F-value	Pr > F
Trt	3	794	22.45	<0.0001
(b) Solutions for fixed effects				
Effect	Categoría	Trt	Estimate	Standard error
Intercept ($\hat{\eta}_1$)	Without		0.6144	0.1799
Intercept ($\hat{\eta}_2$)	Moderate		3.8787	0.2465
Trt ($\hat{\tau}_1$)		1	-1.5034	0.2086
Trt ($\hat{\tau}_2$)		2	-0.2509	0.2055
Trt ($\hat{\tau}_3$)		3	-1.0365	0.2036
Trt ($\hat{\tau}_4$)		4	0	.

“Severe” (severe lesion) for treatment 1 and corresponds to logit $\log\left(\frac{\pi_{11}}{1-\pi_{11}}\right)$, and so on. By taking the inverse of these links values, we can obtain the estimated probabilities of π_{01} and π_{11} . Part of the Statistical Analysis Software (SAS) glimmix output is presented below:

The results of the analysis of variance in part (a) of Table 8.9 indicate that the degree of lesion in the chicken footpad (pododermatitis) in the treatments tested were significantly different ($P < 0.0001$). Therefore, the hypothesis of proportional odds of treatments is rejected ($H_0 : \tau_i = 0$ for all i , that is, *oddsratio* = 1).

In part (b) of Table 8.9, we can see that the estimated intercepts $\hat{\eta}_1 = 0.6144$ and $\hat{\eta}_2 = 3.8787$ define the boundary between the categories “Without” lesion and “Moderate” lesion and the boundary between the categories “Moderate” lesion and “Severe” lesion, respectively. The estimated effect of the treatments ($\hat{\tau}_i$) shows that the boundaries move either upward or downward when a certain treatment is applied. In this sense, all estimated treatment coefficients have a negative effect with respect to treatment 4. This means that chickens under treatments 1–3 have a low probability of developing a moderate lesion and a higher probability of developing a severe lesion than when treatment 4 is applied.

To calculate the probability that a chicken will not develop footpad dermatitis ($c = 0$) when receiving treatment 1, that is, “ $c = 0$, Trt = 1,” we first estimate the linear predictor $\hat{\eta}_{01} = \hat{\eta}_0 + \hat{\tau}_1 = 0.6144 + (-1.5034) = -0.889$, and, taking the inverse, we obtain $\hat{\pi}_{01} = 1/(1+e^{-(-0.889)}) = 0.29$. This value is the estimated probability that a chicken will not develop footpad dermatitis when receiving treatment 1. However, now, for “ $c = 1$, Trt = 1,” $\hat{\eta}_{11} = \hat{\eta}_1 + \hat{\tau}_1 = 3.8787 + (-1.5034) = 2.3753$, whose inverse value is 0.915. This value is an estimate of the probability $\hat{\pi}_{01} + \hat{\pi}_{11}$. From this value, we obtain the probability that a chicken will develop a moderate lesion and a severe lesion. For a moderate lesion, the probability is $\hat{\pi}_{11} = 0.915 - \hat{\pi}_{01} = 0.915 - 0.29 = 0.624$, and, for a severe lesion, the probability is $\hat{\pi}_{21} = 1 - 0.915 = 0.085$. In a similar way the probabilities for the categories ($c = 0, 1, 2$) of the rest of the treatments are computed.

Table 8.10 Estimated odds ratio

Odds ratio estimates				
Comparison	Estimate	DF	95% Confidence limits	
trt 1 vs. 4	0.222	794	0.148	0.335
trt 2 vs. 4	0.778	794	0.520	1.165
trt 3 vs. 4	0.355	794	0.238	0.529

Table 8.11 Estimates on the model scale (Estimate) and on the data scale (Mean) for footpad dermatitis categories in the multinomial cumulative logit model

Estimates							
Label	Estimate	Standard error	DF	<i>t</i> -value	Pr > <i>t</i>	Mean	Standard error mean
<i>c</i> = 0, <i>t</i> = 1	-0.8893	0.2428	794	-4.93	<0.0001	0.2914	0.001174
<i>c</i> = 1, <i>t</i> = 1	2.3753	0.2214	794	10.73	<0.0001	0.9149	0.01724
<i>c</i> = 0, <i>t</i> = 2	0.3634	0.1757	794	2.07	0.0390	0.5899	0.04252
<i>c</i> = 1, <i>t</i> = 2	3.6277	0.2420	794	14.99	<0.0001	0.9741	0.006103
<i>c</i> = 0, <i>t</i> = 3	-0.4222	0.1740	794	-2.43	0.0155	0.3960	0.04162
<i>c</i> = 1, <i>t</i> = 3	2.8422	0.2304	794	12.34	<0.0001	0.9449	0.01199
<i>c</i> = 0, <i>t</i> = 4	3.8787	0.2465	794	15.73	<0.0001	0.9797	0.004893
<i>c</i> = 1, <i>t</i> = 4	0.6144	0.1799	794	3.41	0.0007	0.6489	0.04098

The odds ratios tabulated in Table 8.10 are the odds ratios for treatments 1 through 4, i.e., $e^{\hat{\tau}_i}$ for treatments 1–4. These are the estimated odds ratios of adjacent categories of treatments i ($i = 1, 2, 3$) relative to treatment 4. Values of τ_i are not category-specific; the odds ratios for “Without” lesion versus “Moderate” lesion and those for “Moderate” lesion versus “Severe” lesion are listed below (hence the name “proportional odds”).

From the above odds ratio results, it should be obvious why the F - and P -values in the fixed effects tests are what they are. Adding the “ilink” option to the end of the “estimate” command prompts GLIMMIX to estimate the inverse of the linear predictors ($\hat{\eta}_{ci}$), i.e., the probabilities per category $\hat{\pi}_{ci} = 1 / (1 + e^{-\hat{\eta}_{ci}})$ (Table 8.11).

In the above table, several estimates are shown for $\hat{\eta}_c + \hat{\tau}_i$. For example, the probability that a chicken will not develop a lesion under treatment 1 can be represented by “ $c = 0, t = 1$,” that is, $\hat{\eta}_c + \hat{\tau}_1 = -0.8893$. This result matches the one obtained from the fixed effects table “Solutions for fixed effects” previously shown. Taking the inverse of the link yields the probability

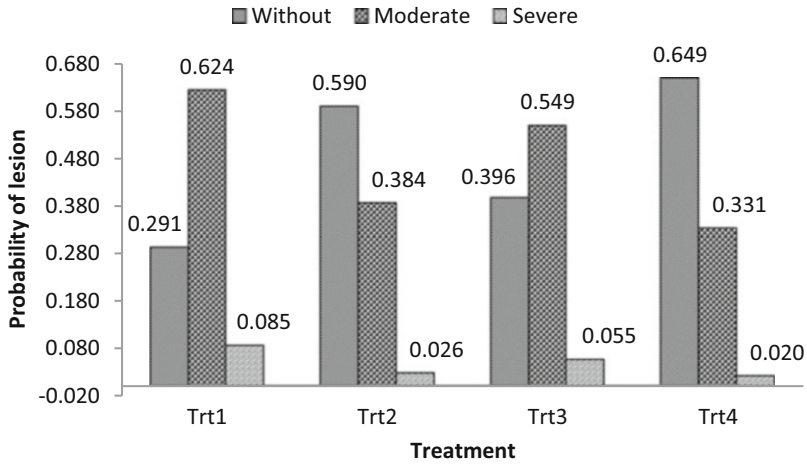


Fig. 8.1 Estimated probabilities for the footpad lesion categories in the treatments tested, using the cumulative logit model

$\hat{\pi}_{01} = 1 / (1 + e^{0.8893}) = 0.2914$. This probability is the maximum likelihood estimate that a chicken will have no footpad lesion with treatment 1. The inverse of the link function is under the “Mean” column of Table 8.11. Now, for the category “ $c = 1, t = 1$,” the inverse of the linear predictor is 0.9149, this is the estimate of $\hat{\pi}_{01} + \hat{\pi}_{11}$. From this value, we can obtain the probability of a chicken showing a “Moderate” lesion when receiving treatment 1, that is, $\hat{\pi}_{01} + \hat{\pi}_{11} = 0.9149$, and, substituting the value of $\hat{\pi}_{01}$, we obtain the value $\hat{\pi}_{11} = 0.9141 - 0.2914 = 0.6227$. Finally, for a “Severe” lesion (category “ $c = 2, t = 1$ ”), the probability that a chicken will present a severe lesion is $\hat{\pi}_{21} = 1 - 0.9141 = 0.0859$. Following the same procedure, we can obtain the probabilities for each of the following categories ($c = 0, 1, 2$) of the rest of the treatments (2–4).

Figure 8.1 shows that under the traditional feeding program with a litter density of 1 kg m^{-2} of rice husks (Trt1), there is a high probability that broilers will develop moderate and severe footpad lesions, as shown by $\hat{\pi}_{11} = 0.624$ and $\hat{\pi}_{21} = 0.085$, respectively. When the litter density was increased from 1 to 2 kg m^{-2} of rice husks under the traditional broiler program (Trt2), the probability of the risk of developing moderate and severe footpad lesions in broilers decreased significantly to $\hat{\pi}_{12} = 0.384$ and $\hat{\pi}_{22} = 0.026$, respectively, compared to Trt1, whereas the probability of not developing a footpad lesion increased to $\hat{\pi}_{02} = 0.590$ (Trt2) compared to $\hat{\pi}_{01} = 0.291$ (Trt1). Regarding the implementation of the two foot care programs plus the litter density of 2 kg husk m^{-2} of rice husks, the probability of chickens of not developing a footpad lesion is $\hat{\pi}_{04} = 0.649$ (Trt4) compared to $\hat{\pi}_{03} = 0.396$ in Trt3, whereas the probability of chickens developing moderate and severe lesions decreased from $\hat{\pi}_{14} = 0.331$ and $\hat{\pi}_{24} = 0.025$ in Trt4 compared to $\hat{\pi}_{13} = 0.549$ and $\hat{\pi}_{23} = 0.055$ in Trt3.

8.4 Cumulative Probit Models

An ordinal cumulative probit model, first considered by Aitchison and Silvey (1957), generalizes a binary probit model to ordinal responses. This model results from the probit modeling of the cumulative probabilities as a linear function of the covariates. The link functions for the cumulative probit model with C categories are listed below:

$$\begin{aligned}\eta_1 &= \Phi^{-1}(\pi_1) = \eta_1 + \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b} \\ \eta_2 &= \Phi^{-1}(\pi_1 + \pi_2) = \eta_2 + \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b} \\ &\vdots \\ \eta_{C-1} &= \Phi^{-1}(\pi_1 + \pi_2 + \cdots + \pi_{C-1}) = \eta_{C-1} + \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b}\end{aligned}$$

where \mathbf{X} and \mathbf{Z} are the design matrices, $\boldsymbol{\beta}$ and \mathbf{b} are the vectors of fixed and random effects parameters, respectively, and $\Phi^{-1}(\cdot)$ is the inverse function of the standard normal cumulative distribution. The inverse link of each of the link functions is as follows:

$$\begin{aligned}\pi_1 &= \Phi(\eta_1) = h(\eta_1) \\ \pi_1 + \pi_2 &= \Phi(\eta_2) = h(\eta_2) \\ &\vdots \\ \pi_1 + \pi_2 + \cdots + \pi_{C-1} &= \Phi(\eta_{C-1}) = h(\eta_{C-1}).\end{aligned}$$

Once $h(\eta_1)$, $h(\eta_2)$, ..., $h(\eta_{C-1})$ are estimated, we can estimate $\hat{\pi}_1$, ..., $\hat{\pi}_C$. The quality of the estimates of the ordinal cumulative probit model are usually very similar to those of an ordinal cumulative logit model for some datasets but not all. Both involve stochastic ordering at different levels of the response variable and are designed to detect the location of changes in the response variable.

Returning to Example 8.3.1, for the cumulative probit model, we change the “LINK = CPROBIT” option in the model’s definition of the above program syntax. The output will contain all the same elements, except the odds ratios. The analysis for the cumulative probit is exactly the same as that one we performed in the cumulative logit model. Part of the output is shown in parts (a)–(c) of Table 8.12.

The estimated variance component due to blocks is $\hat{\sigma}_{\text{block}}^2 = 0.0092$. The results of the analysis of variance showed that the degrees of lesion in the chickens’ footpad (pododermatitis) in the tested treatments differ significantly ($P < 0.0001$).

In part (b) of Table 8.12, it is possible to observe that the estimated intercepts $\hat{\eta}_1 = 0.3880$ and $\hat{\eta}_2 = 2.2407$ define the boundary between the “Without” lesion and “Moderate” lesion categories and the boundary between the “Moderate” lesion and “Severe” lesion categories, respectively. The estimated effect of the treatments ($\hat{\tau}_i$) moves the boundaries either upward or downward, when a certain treatment is applied. In this sense, all estimated treatment coefficients have a negative effect

Table 8.12 Results of the analysis of variance in the multinomial cumulative probit model

(a) Covariance parameter estimates				
Cov Parm	Subject	Estimate	Standard error	
Intercept	Blk	0.009262	0.01817	
(b) Type III tests of fixed effects				
Effect	Num DF	Den DF	F-value	Pr > F
Trt	3	794	24.57	<0.0001
(c) Solutions for fixed effects				
Effect	Categoría	Trt	Estimate	Standard error
Intercept ($\hat{\eta}_1$)	Without		0.3880	0.1124
Intercept ($\hat{\eta}_2$)	Moderate		2.2407	0.1375
Trt ($\hat{\tau}_1$)		1	-0.9278	0.1227
Trt ($\hat{\tau}_2$)		2	-0.1595	0.1242
Trt ($\hat{\tau}_3$)		3	-0.6459	0.1219
Trt ($\hat{\tau}_4$)		4	0	.

with respect to treatment 4. This means that chickens under treatments 1–3 have a low probability of developing a footpad lesion and a higher probability of developing a severe lesion with respect to treatment 4.

From “Type III tests of fixed effects” (Table 8.12, part (b)), the probabilities for each of the categories can be obtained. For the probability that a chicken will not develop a footpad lesion ($c = 0$) under treatment 1, i.e., “ $c = 0, Trt = 1,$ ” the estimated linear predictor is obtained as $\hat{\eta}_{01} = \hat{\eta}_0 + \hat{\tau}_1 = 0.3880 + (-0.9278) = -0.5398$ and, taking the inverse, gives $\hat{\pi}_{01} = \Phi(-0.5398) = 0.2946$, that is, the estimated probability that a chicken will not develop a footpad lesion when receiving treatment 1. For “ $c = 1, Trt = 1,$ ” $\hat{\eta}_{11} = \hat{\eta}_1 + \hat{\tau}_1 = 2.2407 + (-0.9278) = 1.3129$, whose inverse value is 0.9054. This value is an estimator of $\hat{\pi}_{01} + \hat{\pi}_{11}$. From this value, we can obtain the probability that a chicken will develop a moderate lesion and a severe lesion. For a moderate lesion, $\hat{\pi}_{11} = 0.9054 - \hat{\pi}_{01} = 0.9054 - 0.2946 = 0.6108$, and, for a severe lesion, $\hat{\pi}_{21} = 1 - 0.9054 = 0.0946$. Similarly, we can obtain the probabilities of the categories for the other treatments ($c = 0, 1, 2$) for the rest of the treatments.

Similar to the previous example, adding the “ILINK” option to the end of the “ESTIMATE” command prompts GLIMMIX to estimate the values of the linear predictors ($\hat{\eta}_{ci}$) and the inverse of the linear predictors, which are the probabilities per category ($\hat{\pi}_{ci} = \Phi(\hat{\eta}_{ci})$). Table 8.13 shows the estimates of the linear predictors as well as their inverse values (probabilities in this case).

From the above table, we show the estimates of $\hat{\eta}_c + \hat{\tau}_i$. For example, the estimated linear predictor that a chicken will not develop a footpad lesion under treatment 1, i.e., “ $c = 0, t = 1,$ ” is calculated as $\hat{\eta}_c + \hat{\tau}_1 = -0.5398$. This result matches the values obtained from the fixed effects table (“Solutions for fixed effects”) previously shown. Taking the inverse of the link function, $\hat{\pi}_{01} = \Phi(0.5398) = 0.2947$. This is the

Table 8.13 Estimates on the model scale (Estimate) and on the data scale (Mean) for footpad lesion categories in the multinomial cumulative probit model

Estimates							
Label	Estimate	Standard error	DF	<i>t</i> -value	Pr > <i>t</i>	Mean	Standard error mean
$c = 0,$ $t = 1$	-0.5398	0.1100	794	-4.91	<0.0001	0.2947	0.03793
$c = 1,$ $t = 1$	1.3129	0.1208	794	10.87	<0.0001	0.9054	0.02035
$c = 0,$ $t = 2$	0.2285	0.1105	794	2.07	0.0389	0.5904	0.04293
$c = 1,$ $t = 2$	2.0812	0.1345	794	15.47	<0.0001	0.9813	0.006153
$c = 0,$ $t = 3$	-0.2578	0.1085	794	-2.38	0.0178	0.3983	0.04189
$c = 1,$ $t = 3$	1.5949	0.1258	794	12.68	<0.0001	0.9446	0.01407
$c = 0,$ $t = 4$	2.2407	0.1375	794	16.29	<0.0001	0.9875	0.004457
$c = 1,$ $t = 4$	0.3880	0.1124	794	3.45	0.0006	0.6510	0.04158

probability that a chicken will not develop a footpad lesion when receiving treatment 1. This probability is under the “Mean” column.

Now, for the category “ $c = 1, t = 1,$ ” the inverse of the link function is a probability of 0.9054, which results from the inverse value of the linear predictor $\hat{\eta}_1 + \hat{\tau}_1 = 1.3129$. This value is the estimate in terms of probability $\hat{\pi}_{01} + \hat{\pi}_{11}$. From this value, we can obtain the probability that a chicken presents a “Moderate” lesion when receiving treatment 1, that is, $\hat{\pi}_{01} + \hat{\pi}_{11} = 0.9054$, and, using the value of $\hat{\pi}_{01}$, we obtain the values $\hat{\pi}_{11} = 0.9054 - 0.2947 = 0.6107$ and $\hat{\pi}_{21} = 1 - 0.9054 = 0.0946$. Following the same procedure, we can obtain the rest of the probabilities for each one of the categories ($c = 0, 1, 2$) and for the rest of the treatments (2–4).

8.5 Effect of Judges’ Experience on Canned Bean Quality Ratings

Canning quality is one of the most essential traits required in all new dry bean (*Phaseolus vulgaris* L.) varieties, and the selection for this trait is a critical part of bean breeding programs. Advanced lines that are candidates for release as varieties must be evaluated for canning quality for at least 3 years from samples grown at different locations. Quality is evaluated by a panel of judges with varying levels of experience in evaluating breeding lines for visual quality traits. A total of 264 bean breeding lines from 4 commercial classes were retained according to the procedures described by Walters et al. (1997). These included 62 white (navy), 65 black,

Table 8.14 Frequency of ratings of different types of beans as a function of the bean-rating experience

	Black		Kidney		Navy		Pinto	
	< 5 Years	> 5 Years	< 5 Years	> 5 Years	< 5 Years	> 5 Years	< 5 Years	> 5 Years
1	13	32	7	10	10	22	13	2
2	91	78	32	31	56	51	29	17
3	123	124	136	96	84	107	91	68
4	72	122	101	104	84	98	109	124
5	24	31	47	71	51	52	60	109
6	2	3	6	18	24	37	25	78
7	0	0	1	0	1	5	1	12

55 kidney, and 82 pinto bean lines plus control or “check” lines. The visual appearance of the processed beans was determined subjectively by a panel of 13 judges on a 7-point hedonic scale (1 = very undesirable, ..., 4 = neither desirable nor undesirable, ..., 7 = very desirable). Beans were presented to the panel of judges in random order at the same time. Before evaluating the samples, all judges were shown examples of samples rated as satisfactory.

There is concern that certain judges, due to lack of experience, may not be able to correctly score the canned samples. From attribute-based product evaluations, inferences about the effects of experience can be drawn from the psychology literature (Wallsten and Budescu 1981). Prior to the bean canning quality rating experiment, it was postulated that not only do less experienced judges have a more severe rating than do more experienced judges but also that experience should have little or no effect on white beans, for which the canning procedure was developed. Judges are stratified for the purpose of analysis by experience (less than 5 years, greater than 5 years). Counts by canning quality, judge experience, and bean breeding lines are listed in the following table (Table 8.14).

The link functions for the cumulative logit model for describing a variable with *C* categories are as follows:

$$\eta_{(1)ij} = \log\left(\frac{\pi_{1ij}}{1 - \pi_{1ij}}\right) = \eta_1 + \alpha_i + \beta_j + (\alpha\beta)_{ij}$$

$$\eta_{(2)ij} = \log\left(\frac{\pi_{1ij} + \pi_{2ij}}{1 - (\pi_{1ij} + \pi_{2ij})}\right) = \eta_2 + \alpha_i + \beta_j + (\alpha\beta)_{ij}$$

$$\vdots$$

$$\eta_{C-1} = \log\left(\frac{\pi_{1ij} + \pi_{2ij} + \dots + \pi_{(C-1)ij}}{1 - (\pi_{1ij} + \pi_{2ij} + \dots + \pi_{(C-1)ij})}\right) = \eta_{C-1} + \alpha_i + \beta_j + (\alpha\beta)_{ij}$$

The components of the GLMM with an ordinal multinomial response are as follows:

Distributions: $y_{1ij}, y_{2ij}, y_{3ij}, y_{4ij}, y_{5ij}, y_{6ij}, y_{7ij} \sim \text{Multinomial}(N_{ij}, \pi_{1ij}, \pi_{2ij}, \pi_{3ij}, \pi_{4ij}, \pi_{5ij}, \pi_{6ij}, \pi_{7ij})$, where $y_{1ij}, y_{2ij}, y_{3ij}, y_{4ij}, y_{5ij}, y_{6ij}$, and y_{7ij} are the observed frequencies of the responses in each category c of the hedonic scale (1 = very undesirable, ..., 4 = neither desirable nor undesirable, ..., 7 = very desirable).

Linear predictor: $\eta_{(c)ij} = \eta_c + \alpha_i + \beta_j + (\alpha\beta)_{ij}$, where $\eta_{(c)ij}$ is the c th link ($c = 1, 2, \dots, 6$) for bean type i and judge's experience j ; η_c is the intercept for the c th link; α_i is the fixed effect due to the bean type for i th bean class; β_j is the fixed effect due to the j th experience of the judge; and $(\alpha\beta)_{ij}$ is the fixed effect due to the interaction between bean class and judge experience. The link functions for each category are as follows:

$$\log\left(\frac{\pi_{1ij}}{1 - \pi_{1ij}}\right) = \eta_{1ij}$$

$$\log\left(\frac{\pi_{1ij} + \pi_{2ij}}{1 - (\pi_{1ij} + \pi_{2ij})}\right) = \eta_{2ij}$$

$$\log\left(\frac{\pi_{1ij} + \pi_{2ij} + \pi_{3ij}}{1 - (\pi_{1ij} + \pi_{2ij} + \pi_{3ij})}\right) = \eta_{3ij}$$

$$\log\left(\frac{\pi_{1ij} + \pi_{2ij} + \pi_{3ij} + \pi_{4ij}}{1 - (\pi_{1ij} + \pi_{2ij} + \pi_{3ij} + \pi_{4ij})}\right) = \eta_{4ij}$$

$$\log\left(\frac{\pi_{1ij} + \pi_{2ij} + \pi_{3ij} + \pi_{4ij} + \pi_{5ij}}{1 - (\pi_{1ij} + \pi_{2ij} + \pi_{3ij} + \pi_{4ij} + \pi_{5ij})}\right) = \eta_{5ij}$$

$$\log\left(\frac{\pi_{1ij} + \pi_{2ij} + \pi_{3ij} + \pi_{4ij} + \pi_{5ij} + \pi_{6ij}}{1 - (\pi_{1ij} + \pi_{2ij} + \pi_{3ij} + \pi_{4ij} + \pi_{5ij} + \pi_{6ij})}\right) = \eta_{6ij}$$

The following GLIMMIX commands fit a cumulative logit model with an ordinal multinomial response.

```
proc glimmix data=beans ;
class Exper;
model cal (order=data) = Exper|Class/dist=Multinomial link=clogit
```

```

solution oddsratio;
Contrast 'Effect of Experience on Black bean' exper 1 -1 class*exper 1 -1
0 0 0 0 0 0 0 0 0 0;
Contrast 'Effect of Experience on Kidney Bean' exper 1 -1 class*exper 0 0
1 -1 0 0 0 0 0 0 0 0;
Contrast 'Effect of Experience on Navies bean' exper 1 -1 class*exper 0 0
0 0 0 0 0 1 -1 0 0 0;
Contrast 'Effect of Experience on Pinto beans' exper 1 -1 class*exper 0 0
0 0 0 0 0 0 0 0 1 -1;
estimate 'Black, < 5 year, Rating = 1' Intercept 1 0 0 0 0 0 0 0 0 class 1 0
0 0 0 0 0 exper 1 0 class*exper 1 0 0 0 0 0 0 0 0 0/ilink;
estimate 'Black, < 5 year, Rating <= 2' Intercept 0 1 0 0 0 0 0 0 0 class 1 0
0 0 0 0 0 exper 1 0 0 class*exper 1 0 0 0 0 0 0 0 0 0/ilink;
estimate 'Black, < 5 year, Rating <= 3' Intercept 0 0 0 1 0 0 0 0 0 class 1 0
0 0 0 0 0 exper 1 0 0 class*exper 1 0 0 0 0 0 0 0 0 0/ilink;
estimate 'Black, < 5 year, Rating <= 4' Intercept 0 0 0 0 1 0 0 0 class 1 0 0 0
0 0 0 0 0 exper 1 0 0 class*exper 1 0 0 0 0 0 0 0 0 0/ilink;
estimate 'Black, < 5 year, Rating <= 5' Intercept 0 0 0 0 0 0 1 0 0 class 1 0
0 0 0 0 0 exper 1 0 0 class*exper 1 0 0 0 0 0 0 0 0 0/ilink;
estimate 'Black, > 5 year, Rating <= 6' Intercept 0 0 0 0 0 0 0 0 1 class 1 0
0 0 0 0 0 exper 1 0 class*exper 1 0 0 0 0 0 0 0 0 0/ilink;
estimate 'Black, > 5 year, Rating = 1' Intercept 1 0 0 0 0 0 0 0 0 class 1 0
0 0 0 0 0 exper 0 1 class*exper 0 1 0 0 0 0 0 0 0 0/ilink;
estimate 'Black, > 5 year, Rating <= 2' Intercept 0 1 0 0 0 0 0 0 0 class 1 0
0 0 0 0 0 exper 0 1 class*exper 0 1 0 0 0 0 0 0 0 0/ilink;
estimate 'Black, > 5 year, Rating <= 3' Intercept 0 0 0 1 0 0 0 0 0 class 1 0
0 0 0 0 0 exper 0 1 class*exper 0 1 0 0 0 0 0 0 0 0/ilink;
estimate 'Black, > 5 year, Rating <= 4' Intercept 0 0 0 0 1 0 0 0 class 1 0 0 0
0 0 0 0 0 exper 0 1 class*exper 0 1 0 0 0 0 0 0 0 0/ilink;
estimate 'Black, > 5 year, Rating <= 5' Intercept 0 0 0 0 0 0 1 0 0 class 1 0
0 0 0 0 0 exper 0 1 class*exper 0 1 0 0 0 0 0 0 0 0/ilink;
estimate 'Black, > 5 year, Rating <= 6' Intercept 0 0 0 0 0 0 0 0 1 class 1 0
0 0 0 0 0 exper 0 1 class*exper 0 1 0 0 0 0 0 0 0 0/ilink;
estimate 'Kidney, < 5 year, Rating = 1' Intercept 1 0 0 0 0 0 0 0 0 class 0 1
0 0 0 0 0 exper 1 0 0 class*exper 0 0 0 1 0 0 0 0 0 0/ilink;
estimate 'Kidney, < 5 year, Rating <= 2' Intercept 0 1 0 0 0 0 0 0 0 class 0 1
0 0 0 0 0 exper 1 0 0 class*exper 0 0 0 1 0 0 0 0 0 0/ilink;
estimate 'Kidney, < 5 yr, Rating <= 3' Intercept 0 0 0 1 0 0 0 0 0 class 0 1
0 0 0 0 0 exper 1 0 0 class*exper 0 0 0 1 0 0 0 0 0 0/ilink;
estimate 'Kidney, < 5 year, Rating <= 4' Intercept 0 0 0 0 0 1 0 0 0 class 0 1
0 0 0 0 0 exper 1 0 0 class*exper 0 0 0 1 0 0 0 0 0 0/ilink;
estimate 'Kidney, < 5 year, Rating <= 5' Intercept 0 0 0 0 0 0 1 0 0 class 0 1
0 0 0 0 0 exper 1 0 0 class*exper 0 0 0 1 0 0 0 0 0 0/ilink;
estimate 'Kidney, < 5 year, Rating <= 6' Intercept 0 0 0 0 0 0 0 0 1 class 0 1
0 0 0 0 0 exper 1 0 0 class*exper 0 0 0 1 0 0 0 0 0 0/ilink;
estimate 'Kidney, > 5 year, Rating = 1' Intercept 1 0 0 0 0 0 0 0 0 class 0 1
0 0 0 0 0 exper 0 1 class*exper 0 0 0 0 1 0 0 0 0 0/ilink;
estimate 'Kidney, > 5 year, Rating <= 2' Intercept 0 1 0 0 0 0 0 0 0 class 0 1
0 0 0 0 0 exper 0 1 class*exper 0 0 0 0 1 0 0 0 0 0/ilink;
estimate 'Kidney, > 5 year, Rating <= 3' Intercept 0 0 0 1 0 0 0 0 0 class 0 1
0 0 0 0 0 exper 0 1 class*exper 0 0 0 0 1 0 0 0 0 0/ilink;
estimate 'Kidney, > 5 year, Rating <= 4' Intercept 0 0 0 0 0 1 0 0 0 class 0 1
0 0 0 0 0 exper 0 1 class*exper 0 0 0 0 1 0 0 0 0 0/ilink;

```

```

estimate 'Kidney, > 5 year, Rating <= 5' Intercept 0 0 0 0 0 0 1 0 0 class 0 1
0 0 0 0 exper 0 1 class*exper 0 0 0 0 1 0 0 0 0 0 0/ilink;
estimate 'Kidney, > 5 year, Rating <= 6' Intercept 0 0 0 0 0 0 0 1 class 0 1
0 0 0 0 exper 0 1 class*exper 0 0 0 0 1 0 0 0 0 0 0/ilink;
estimate 'Navies, < 5 year, Rating = 1' Intercept 1 0 0 0 0 0 0 0 class 0 0
0 1 0 0 exper 1 0 0 class*exper 0 0 0 0 0 1 0 0 0 0 0/ilink;
estimate 'Navies, < 5 year, Qualification <= 2' Intercept 0 1 0 0 0 0 0 0
0 class 0 0 0 1 0 0 exper 1 0 0 class*exper 0 0 0 0 0 1 0 0 0 0 0/ilink;
estimate 'Navies, < 5 year, Qualification <= 3' Intercept 0 0 0 0 1 0 0 0 0
class 0 0 0 1 0 0 exper 1 0 0 class*exper 0 0 0 0 0 1 0 0 0 0 0/ilink;
estimate 'Navies, < 5 year, Rating <= 4' Intercept 0 0 0 0 0 1 0 0 0 class 0 0
0 1 0 0 exper 1 0 0 class*exper 0 0 0 0 0 1 0 0 0 0 0/ilink;
estimate 'Navies, < 5 year, Rating <= 5' Intercept 0 0 0 0 0 0 0 1 0 0 class
0 0 0 1 0 0 exper 1 0 0 class*exper 0 0 0 0 0 1 0 0 0 0 0/ilink;
estimate 'Navies, < 5 year, Qualification <= 6' Intercept 0 0 0 0 0 0 0 0 1
class 0 0 0 1 0 0 exper 1 0 0 class*exper 0 0 0 0 0 1 0 0 0 0 0/ilink;
estimate 'Navies, > 5 year, Qualification = 1' Intercept 1 0 0 0 0 0 0 0
0 class 0 0 0 1 0 0 exper 0 1 class*exper 0 0 0 0 0 0 0 1 0 0 0/ilink;
estimate 'Navies, > 5 year, Qualification <= 2' Intercept 0 1 0 0 0 0 0 0
0 class 0 0 0 1 0 0 exper 0 1 class*exper 0 0 0 0 0 0 0 1 0 0 0/ilink;
estimate 'Navies, > 5 year, Rating <= 3' Intercept 0 0 0 1 0 0 0 0 0 class 0 0
0 1 0 0 exper 0 1 class*exper 0 0 0 0 0 0 0 1 0 0 0/ilink;
estimate 'Navies, > 5 year, Rating <= 4' Intercept 0 0 0 0 0 1 0 0 0 class 0 0
0 1 0 0 exper 0 1 class*exper 0 0 0 0 0 0 0 1 0 0 0/ilink;
estimate 'Navies, > 5 year, Rating <= 5' Intercept 0 0 0 0 0 0 0 1 0 0 class
0 0 0 1 0 0 exper 0 1 class*exper 0 0 0 0 0 0 0 1 0 0 0/ilink;
estimate 'Navies, > 5 year, Rating <= 6' Intercept 0 0 0 0 0 0 0 0 1 class
0 0 0 1 0 0 exper 0 1 class*exper 0 0 0 0 0 0 0 1 0 0 0/ilink;
estimate 'Pinto, < 5 year, Qualification = 1' Intercept 1 0 0 0 0 0 0 0
0 class 0 0 0 0 0 0 1 exper 1 0 class*exper 0 0 0 0 0 0 0 0 1 0 0/ilink;
estimate 'Pinto, < 5 year, Qualification <= 2' Intercept 0 1 0 0 0 0 0 0
0 class 0 0 0 0 0 0 1 exper 1 0 0 class*exper 0 0 0 0 0 0 0 1 0 0/ilink;
estimate 'Pinto, < 5 year, Qualification <= 3' Intercept 0 0 0 0 1 0 0 0 0
class 0 0 0 0 0 1 exper 1 0 0 class*exper 0 0 0 0 0 0 0 0 1 0 0/ilink;
estimate 'Pinto, < 5 year, Rating <= 4' Intercept 0 0 0 0 0 1 0 0 0 class 0 0
0 0 0 1 exper 1 0 0 class*exper 0 0 0 0 0 0 0 0 1 0 0/ilink;
estimate 'Pinto, < 5 year, Rating <= 5' Intercept 0 0 0 0 0 0 0 1 0 0 class 0 0
0 0 0 1 exper 1 0 class*exper 0 0 0 0 0 0 0 0 1 0 0/ilink;
estimate 'Pinto, < 5 year, Rating <= 6' Intercept 0 0 0 0 0 0 0 0 1 class 0 0
0 0 0 1 exper 1 0 class*exper 0 0 0 0 0 0 0 0 1 0 0/ilink;
estimate 'Pinto, > 5 years, Qualification = 1' Intercept 1 0 0 0 0 0 0 0
0 class 0 0 0 0 0 1 exper 0 1 class*exper 0 0 0 0 0 0 0 0 0 1/ilink;
estimate 'Pinto, > 5 year, Qualification <= 2' Intercept 0 1 0 0 0 0 0 0
0 class 0 0 0 0 0 1 exper 0 1 class*exper 0 0 0 0 0 0 0 0 0 1/ilink;
estimate 'Pinto, > 5 year, Qualification <= 3' Intercept 0 0 0 1 0 0 0 0
0 class 0 0 0 0 0 1 exper 0 1 class*exper 0 0 0 0 0 0 0 0 0 1/ilink;
estimate 'Pinto, > 5 year, Rating <= 4' Intercept 0 0 0 0 0 1 0 0 0 class 0 0
0 0 0 1 exper 0 1 class*exper 0 0 0 0 0 0 0 0 0 1/ilink;
estimate 'Pinto, > 5 year, Rating <= 5' Intercept 0 0 0 0 0 0 0 1 0 0 class 0 0
0 0 0 1 exper 0 1 class*exper 0 0 0 0 0 0 0 0 0 1/ilink;
estimate 'Pinto, > 5 year, Qualification <= 6' Intercept 0 0 0 0 0 0 0 0 1
class 0 0 0 0 0 1 exper 0 1 class*exper 0 0 0 0 0 0 0 0 0 1/ilink;
freq y;
run;

```

Table 8.15 Fixed effects hypothesis testing in the multinomial cumulative logit model

Type III tests of fixed effects				
Effect	Num DF	Den DF	F-value	Pr > F
Class	1	2779	36.19	<0.0001
Exper	3	2779	85.20	<0.0001
Class*Exper	3	2779	10.13	<0.0001

Table 8.16 Hypothesis testing in quality assessment

Contrasts				
Label	Num DF	Den DF	F-value	Pr > F
Effect of experience on black beans	1	2779	2.77	0.0961
Effect of experience on kidney beans	1	2779	7.86	0.0051
Effect of experience on navy beans	1	2779	0.02	0.8822
Effect of experience on pinto beans	1	2779	58.06	<0.0001

Part of the results is shown below. The results of the analysis of variance show that the class of bean (Class), experience of the evaluator (Exper), and the interaction between class and experience (Class×Exper) on bean canning scores differ significantly ($P = 0.0001$). That is, the results of comparing judges with more and less years of experience will depend on the line (variety) of beans (Table 8.15).

The contrasts address this interaction (Table 8.16). Hypothesis testing is as follows: $\pi_{\text{class of bean, } < 5 \text{ years of experience}} = \pi_{\text{class of bean, } > 5 \text{ years of experience}}$.

The results show that judges with more than 5 years of experience differ from those with less than 5 years of experience in evaluating the quality of canned kidney and pinto beans (Table 8.16). With the “solution” option in the model specification, the fixed parameter estimates table shows the solution of the fixed effects parameters under maximum likelihood. In this table, we can observe the values of the estimated intercepts: $\hat{\eta}_1 = -4.6421$ defines the boundary between the categories, “1 = highly undesirable” and “2 = moderately undesirable”, whereas $\hat{\eta}_2 = -2.9316$ defines the boundary between the categories “2 = moderately undesirable” and “3 = slightly undesirable.” The third intercept defines the boundary between the categories “3 = moderately undesirable” and “3 = slightly undesirable,” $\hat{\eta}_3 = -1.3995$ defines the boundary between the categories “3 = slightly undesirable” and “4 = neither undesirable nor desirable,” and so on.

The estimated effects of bean type ($\hat{\alpha}_i$), evaluator ($\hat{\beta}_i$), and their interaction ($\hat{\alpha}\hat{\beta}_{ij}$) are shown below. From these values, we can estimate the linear predictors for each of the categories. For example, the linear predictor for canned black beans evaluated by an inexperienced judge who assigns the category “1 = very undesirable” is $\hat{\eta}_{111} = \hat{\eta}_1 + \hat{\alpha}_1 + \hat{\beta}_1 + \hat{\alpha}\hat{\beta}_{11} = -4.6421 + 1.9670 + 1.0284 - 0.8066 = -2.4533$, for category “2 = moderately undesirable,” it is $\hat{\eta}_{211} = \hat{\eta}_2 + \hat{\alpha}_1 + \hat{\beta}_1 + \hat{\alpha}\hat{\beta}_{11} = -2.9316 + 1.9670 + 1.0284 - 0.8066 = -0.7428$, for category “3 = slightly undesirable,” it is $\hat{\eta}_{311} = \hat{\eta}_3 + \hat{\alpha}_1 + \hat{\beta}_1 + \hat{\alpha}\hat{\beta}_{11} = -1.3995 + 1.9670 + 1.0284 - 0.8066 = 0.7893$, and,

Table 8.17 Maximum likelihood estimation of the estimated parameters in the fixed effects solution of canned bean quality ratings in the multinomial cumulative logit model

Fixed parameter estimates								
Effect	Cal $\hat{\eta}_i$	Class $\hat{\alpha}_i$	Expert $\hat{\beta}_1$	Estimate	Standard error	DF	t-value	Pr > t
Intercept $\hat{\eta}_1$	1			-4.6421	0.1363	2779	-34.05	<0.0001
Intercept $\hat{\eta}_2$	2			-2.9316	0.1057	2779	-27.74	<0.0001
Intercept $\hat{\eta}_3$	3			-1.3995	0.09643	2779	-14.51	<0.0001
Intercept $\hat{\eta}_4$	4			0.004287	0.09230	2779	0.05	0.9630
Intercept $\hat{\eta}_5$	5			1.4191	0.1026	2779	13.84	<0.0001
Intercept $\hat{\eta}_6$	6			3.8925	0.2346	2779	16.59	<0.0001
Class		Black $\hat{\alpha}_1$		1.9670	0.1318	2779	14.93	<0.0001
Class		Kidney $\hat{\alpha}_2$		1.0472	0.1342	2779	7.80	<0.0001
Class		Navy $\hat{\alpha}_3$		1.3076	0.1345	2779	9.72	<0.0001
Class		Pinto $\hat{\alpha}_4$		0
Exper			1 $\hat{\beta}_1$	1.0284	0.1350	2779	7.62	<0.0001
Exper			2	0
Class*Exper		Black	1 $\hat{\alpha}\hat{\beta}_{11}$	-0.8066	0.1894	2779	-4.26	<0.0001
Class*Exper		Black	2	0
Class*Exper		Kidney	1 $\hat{\alpha}\hat{\beta}_{21}$	-0.6457	0.1912	2779	-3.38	0.0007
Class*Exper		Kidney	2	0
Class*Exper		Navy	1 $\hat{\alpha}\hat{\beta}_{31}$	-1.0072	0.1969	2779	-5.12	<0.0001
Class*Exper		Navy	2	0
Class*Exper		Pinto	1 $\hat{\alpha}\hat{\beta}_{41}$	0
Class*Exper		Pinto	2	0

for category “4 = neither undesirable nor desirable,” it is $\hat{\eta}_{411} = \hat{\eta}_4 + \hat{\alpha}_1 + \hat{\beta}_1 + \hat{\alpha}\hat{\beta}_{11} = 0.004287 + 1.9670 + 1.0284 - 0.8066 = 2.1931$. This is how the other categories are calculated for each type of bean and assessor (Table 8.17).

The results of Table 8.18 were obtained with the “estimate” command in conjunction with the “ilink” option that prompts GLIMMIX to compute the values of the linear predictors, $\hat{\eta}_{cij}$, tabulated under the “Estimate” column, and the estimated probabilities $\hat{\pi}_{cij}$ for all categories of each treatment are tabulated under the “Mean” column ($\hat{\pi}_{cij}$), except the reference category.

From Table 8.18 (“Estimates”), we can obtain the probabilities reported under the “Mean” column in which an inexperienced (<5 years) panelist (judge) would rate canned black beans as category 1 (1 = highly undesirable) with a probability of $\hat{\pi}_{111} = 0.08$ compared to an experienced panelist (>5 years) who would give a

Table 8.18 Estimates on the model scale (Estimate) and on the data scale (Mean) based on judges' experience in canned bean quality ratings in the multinomial cumulative logit model

Estimates							
Label	Estimate	Standard error	DF	t-value	Pr > t	Mean	Standard error mean
Black <5 years, score = 1	-2.4533	0.1292	2779	-18.99	<0.0001	0.07920	0.009419
Black <5 years, score ≤ 2	-0.7428	0.1004	2779	-7.40	<0.0001	0.3224	0.02194
Black <5 years, score ≤ 3	0.7893	0.1008	2779	7.83	<0.0001	0.6877	0.02164
Black <5 years, score ≤ 4	2.1931	0.1076	2779	20.38	<0.0001	0.8996	0.009716
Black <5 years, score ≤ 5	3.6079	0.1238	2779	29.15	<0.0001	0.9736	0.003180
Black >5 years, score ≤ 6	6.0814	0.2467	2779	24.65	<0.0001	0.9977	0.000561
Black >5 years, score = 1	-2.6751	0.1264	2779	-21.17	<0.0001	0.06446	0.007621
Black >5 years, score ≤ 2	-0.9646	0.09577	2779	-10.07	<0.0001	0.2760	0.01913
Black >5 years, score ≤ 3	0.5675	0.09314	2779	6.09	<0.0001	0.6382	0.02151
Black >5 years, score ≤ 4	1.9713	0.09967	2779	19.78	<0.0001	0.8778	0.01069
Black >5 years, score ≤ 5	3.3861	0.1170	2779	28.95	<0.0001	0.9673	0.003704
Black >5 years, score ≤ 6	5.8595	0.2434	2779	24.07	<0.0001	0.9972	0.000690
Kidney <5 years, score = 1	-3.2122	0.1333	2779	-24.11	<0.0001	0.03871	0.004958
Kidney <5 years, score ≤ 2	-1.5017	0.1018	2779	-14.74	<0.0001	0.1822	0.01517

(continued)

Table 8.18 (continued)

Estimates							
Label	Estimate	Standard error	DF	<i>t</i> -value	Pr > <i>t</i>	Mean	Standard error mean
Kidney <5 years, score ≤ 3	0.03040	0.09608	2779	0.32	0.7517	0.5076	0.02401
Kidney <5 years, score ≤ 4	1.4342	0.1011	2779	14.19	<0.0001	0.8076	0.01571
Kidney <5 years, score ≤ 5	2.8490	0.1178	2779	24.18	<0.0001	0.9453	0.006096
Kidney >5 years, score ≤ 6	5.3225	0.2438	2779	21.83	<0.0001	0.9951	0.001179
Kidney >5 years, score = 1	-3.5949	0.1372	2779	-26.20	<0.0001	0.02673	0.003569
Kidney >5 years, score ≤ 2	-1.8844	0.1071	2779	-17.60	<0.0001	0.1319	0.01226
Kidney >5 years, score ≤ 3	-0.3523	0.09988	2779	-3.53	0.0004	0.4128	0.02421
Kidney >5 years, score ≤ 4	1.0515	0.1020	2779	10.31	<0.0001	0.7411	0.01957
Kidney >5 years, score ≤ 5	2.4663	0.1176	2779	20.98	<0.0001	0.9217	0.008480
Kidney >5 years, score ≤ 6	4.9397	0.2436	2779	20.27	<0.0001	0.9929	0.001719
Navies <5 years, score = 1	-3.3133	0.1404	2779	-23.60	<0.0001	0.03512	0.004757
Navies <5 years, score ≤ 2	-1.6027	0.1119	2779	-14.33	<0.0001	0.1676	0.01561
Navies <5 years, score ≤ 3	-0.07066	0.1068	2779	-0.66	0.5084	0.4823	0.02667
Navies <5 years, score ≤ 4	1.3332	0.1102	2779	12.10	<0.0001	0.7914	0.01820
Navies <5 years, score ≤ 5	2.7479	0.1251	2779	21.97	<0.0001	0.9398	0.007077

(continued)

Table 8.18 (continued)

Estimates							
Label	Estimate	Standard error	DF	t-value	Pr > t	Mean	Standard error mean
Navies >5 years, score ≤ 6	5.2214	0.2473	2779	21.12	<0.0001	0.9946	0.001321
Navies >5 years, score = 1	-3.3345	0.1348	2779	-24.74	<0.0001	0.03441	0.004479
Navies >5 years, score ≤ 2	-1.6240	0.1047	2779	-15.51	<0.0001	0.1647	0.01440
Navies >5 years, score ≤ 3	-0.09190	0.09897	2779	-0.93	0.3532	0.4770	0.02469
Navies >5 years, score ≤ 4	1.3119	0.1028	2779	12.76	<0.0001	0.7878	0.01719
Navies >5 years, score ≤ 5	2.7267	0.1186	2779	22.99	<0.0001	0.9386	0.006836
Navies >5 years, score ≤ 6	5.2002	0.2439	2779	21.32	<0.0001	0.9945	0.001331
Pinto <5 years, score = 1	-3.6137	0.1380	2779	-26.19	<0.0001	0.02624	0.003527
Pinto <5 years, score ≤ 2	-1.9032	0.1081	2779	-17.61	<0.0001	0.1297	0.01221
Pinto <5 years, score ≤ 3	-0.3711	0.1008	2779	-3.68	0.0002	0.4083	0.02436
Pinto <5 years, score ≤ 4	1.0327	0.1030	2779	10.03	<0.0001	0.7374	0.01994
Pinto <5 years, score ≤ 5	2.4475	0.1184	2779	20.67	<0.0001	0.9204	0.008678
Pinto >5 years, score ≤ 6	4.9210	0.2439	2779	20.17	<0.0001	0.9928	0.001753
Pinto >5 years, score = 1	-4.6421	0.1363	2779	-34.05	<0.0001	0.009545	0.001289
Pinto >5 years, score ≤ 2	-2.9316	0.1057	2779	-27.74	<0.0001	0.05061	0.005078

(continued)

Table 8.18 (continued)

Estimates							
Label	Estimate	Standard error	DF	t-value	Pr > t	Mean	Standard error mean
Pinto >5 years, score ≤ 3	-1.3995	0.09643	2779	-14.51	<0.0001	0.1979	0.01531
Pinto >5 years, score ≤ 4	0.004287	0.09230	2779	0.05	0.9630	0.5011	0.02307
Pinto >5 years, score ≤ 5	1.4191	0.1026	2779	13.84	<0.0001	0.8052	0.01609
Pinto >5 years, score ≤ 6	3.8925	0.2346	2779	16.59	<0.0001	0.9800	0.004595

probability of $\hat{\pi}_{112} = 0.0646$. To calculate the probability that a judge with less than 5 years experience would assign a rating of 2 (2 = moderately undesirable) to canned black beans, we derive this probability from the cumulative probability of 0.3224, which corresponds to $\hat{\pi}_{211} + \hat{\pi}_{111}$, from which we get $\hat{\pi}_{211} = 0.3224 - \hat{\pi}_{111} = 0.3224 - 0.08 = 0.24$. On the other hand, for a judge with experience (>5 years), the probability of assigning a score of 2 to canned black beans is $\hat{\pi}_{212} = 0.2760 - \hat{\pi}_{112} = 0.2760 - 0.06446 = 0.2115$.

Following the same procedure, the other probabilities for the rest of the categories are obtained. The probabilities calculated for each of the categories are shown in Table 8.19 and can be seen in Fig. 8.2.

8.6 Generalized Logit Models: Nominal Response Variables

In a model with unordered data, the polytomous response variable does not have an ordered structure. Two classes of models, generalized logit models and conditional logit models, can be used with nominal response data. A generalized logit model consists of a combination of several binary logits estimated simultaneously. A logit model is the simplest and best-known probabilistic choice model. However, there are problems in making use of a multinomial logit model because of its inflexibility. A generalized logit model is essentially more flexible than the traditional multinomial cumulative logit model.

A generalized logit model shows the same flexibility as a probit model but is much more tractable. Like cumulative logit and probit models, a generalized logit model has $C - 1$ link functions, where C denotes the number of response categories.

Table 8.19 Probabilities calculated for each of the canned bean grades

		Cal1	Cal2	Cal3	Cal4	Cal5	Cal6	Cal7
Black	J1	0.08	0.24	0.37	0.21	0.07	0.02	0.00
	J2	0.06	0.21	0.36	0.24	0.09	0.03	0.00
Kidney	J1	0.04	0.14	0.33	0.30	0.14	0.05	0.00
	J2	0.03	0.11	0.28	0.33	0.18	0.07	0.01
Navy	J1	0.04	0.13	0.31	0.31	0.15	0.05	0.01
	J2	0.03	0.13	0.31	0.31	0.15	0.06	0.01
Pinto	J1	0.03	0.10	0.28	0.33	0.18	0.07	0.01
	J2	0.01	0.04	0.15	0.30	0.30	0.17	0.02

Cal1 = qualification 1, Cal2 = qualification 2,....., Cal7 = qualification 7; J1 = panelist with less than 5 years' experience, and J2 = panelist with more than 5 years' experience

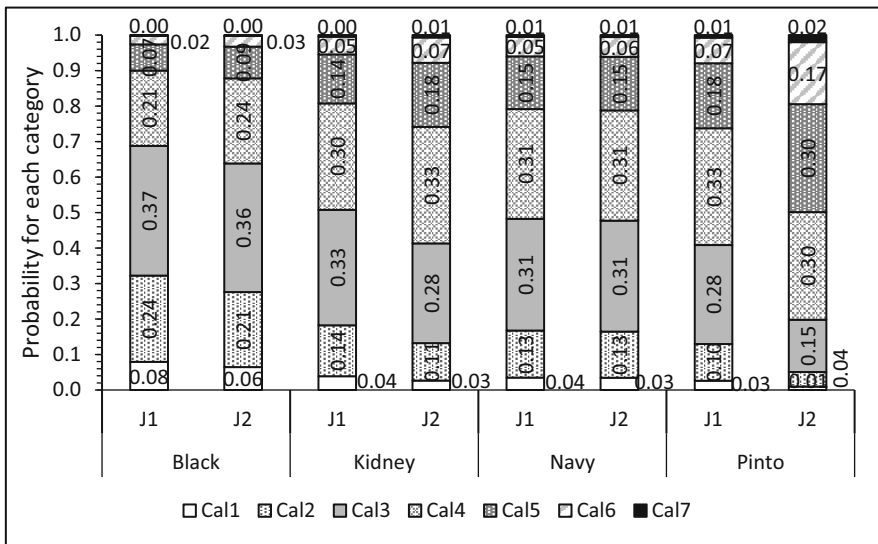


Fig. 8.2 Estimated probabilities for each category of the acceptability of canned beans, according to the experience of the panelist (judge)

Moreover, in this class of models, a category is first defined as the reference category. This may be arbitrary or it may make compelling logical sense in the study to designate a particular response category as the reference. In practice and throughout the analysis, the category used as the reference is irrelevant, as long as we are consistent about it. For example, if *C* is used as the reference category, then the generalized logits are defined as shown below:

$$\begin{aligned}\eta_1 &= \log\left(\frac{\pi_{1ij}}{\pi_{Cij}}\right) = \alpha_1 + \mathbf{X}\beta_1 + \mathbf{Z}\mathbf{b}_1 \\ \eta_2 &= \log\left(\frac{\pi_{2ij}}{\pi_{Cij}}\right) = \alpha_2 + \mathbf{X}\beta_2 + \mathbf{Z}\mathbf{b}_2 \\ &\vdots \\ \eta_{C-1} &= \log\left(\frac{\pi_{(C-1)ij}}{\pi_{Cij}}\right) = \alpha_{C-1} + \mathbf{X}\beta_{C-1} + \mathbf{Z}\mathbf{b}_{C-1}\end{aligned}$$

Given the different effects in the models, the intercepts (α 's), β 's, and \mathbf{b} 's vary across the pairs of response variable categories for each link function. Using algebra, it can be shown that the general form of the inverse of the link functions is given by

$$\pi_c = \frac{e^{\eta_c}}{C-1 + \sum_{c=1}^{C-1} e^{\eta_c}}, \quad c = 1, 2, \dots, C-1$$

Once $\pi_1, \pi_2, \dots, \pi_{C-1}$ are estimated, the reference category is estimated as

$$\pi_C = 1 - \sum_{c=1}^{C-1} \pi_c.$$

8.6.1 CRDs with a Nominal Multinomial Response

In practice, cumulative models are used for analyzing ordinal data and generalized logit models for nominal data. Returning to Example 8.3.1, we will now implement the analysis of a generalized logit model. This model relaxes the assumptions of proportionality; but it is less parsimonious than the “odds ratio” model since they fit $C - 1$ binary logit models, where C is the number of categories of the response variable. The linear predictor and distribution are the same as in the previous example.

The following GLIMMIX syntax implements the analysis of the generalized logit model:

```
proc glimmix data=chickens ;
class trt block category;
model category(reference='severe')= trt/dist=Multinomial
link=glogit oddsratio;
random intercept/subject=block solution group=category;
estimate 't=1' intercept 1 trt 1 0,
't=2' intercept 1 trt 0 1 0,
't=3' intercept 1 trt 0 0 0 1 0,
't=4' intercept 1 trt 0 0 0 0 1/ilink bycat;
freq y;
run;
```

Table 8.20 Analysis of variance in the generalized multinomial logit model

Type III tests of fixed effects				
Effect	Num DF	Den DF	F-value	Pr > F
Trt	6	790	10.78	<0.0001

Table 8.21 Maximum likelihood estimates on the model scale (Estimate) for footpad lesion level in the multinomial generalized logit model

Solutions for fixed effects							
Effect	Category	Trt	Estimate	Standard error	DF	t-value	Pr > t
Intercept	Without lesion		4.8525	1.0059	2	4.82	0.0404
Intercept	Moderate lesion		4.2485	1.0071	2	4.22	0.0519
trt	Without lesion	1	-3.8447	1.0330	790	-3.72	0.0002
trt	Moderate lesion	1	-2.6478	1.0327	790	-2.56	0.0105
trt	Without lesion	2	-1.1888	1.1618	790	-1.02	0.3065
trt	Moderate lesion	2	-0.9651	1.1662	790	-0.83	0.4082
trt	Without lesion	3	-2.7860	1.0585	790	-2.63	0.0087
trt	Moderate lesion	3	-1.8326	1.0598	790	-1.73	0.0842
trt	Without lesion	4	0
trt	Moderate lesion	4	0

Most of the syntax of the program has already been explained. The “reference=” option is new to this program in the command, where the model is defined and is used to designate the reference category. By not specifying the “reference=” option, GLIMMIX by default uses the last category in the dataset. Moreover, the “link = glogit” option prompts GLIMMIX to fit a generalized logit model. The “bycat” option in the “estimate” command is unique to the generalized logit model. Finally, the “ilink” option asks GLIMMIX to estimate all category probabilities for each treatment, except those for the reference category. Part of the output is shown in Table 8.20. The fixed effects test shows that there are highly significant differences ($P = 0.0001$) on the average percentage of footpad lesion level between treatments.

Unlike the cumulative logit model, in the generalized logit model, the estimates of the fixed effects (treatments), as well as the intercepts, are separated for each link function. For the estimation of linear predictors, we use the estimated values of Table 8.21 (“Solutions for fixed effects”). The estimated intercepts $\hat{\alpha}_1 = 4.8525$ and $\hat{\alpha}_2 = 4.2485$ define the boundary between the categories “Without” lesion and “Moderate” lesion and the boundary between the categories “Moderate” lesion and “Severe” lesion, respectively. For treatment 1, the treatment effects ($\hat{\tau}_i$) estimated for the “Without” lesion category is $\hat{\tau}_1 = -3.8447$ and for the “Moderate” lesion category, it is $\hat{\tau}_1 = -2.6478$. With these values, the linear predictors for the “Without” lesion and “Moderate” lesion categories under treatment 1 are $\hat{\eta}_{01} = 4.8525 - 3.8447 = 1.0077$ and $\hat{\eta}_{11} = 4.2485 - 2.6478 = 1.6007$, respectively.

The estimated probabilities for each of the categories (“Without” lesion and “Moderate” lesion) in each treatment, except for the reference category, are found

Table 8.22 Estimates on the model scale (“Estimate”) and on the data scale (“Mean”) for footpad lesion level observed in treatments in the multinomial generalized logit model

Estimates								
Label	Category	Estimate	Standard error	DF	t-value	Pr > t	Mean	Standard error mean
t = 1	Without lesion	1.0077	0.2515	790	4.01	<0.0001	0.3150	0.03552
t = 1	Moderate lesion	1.6007	0.2286	790	7.00	<0.0001	0.5700	0.03677
t = 2	Without lesion	3.6637	0.5881	790	6.23	<0.0001	0.5850	0.03801
t = 2	Moderate lesion	3.2834	0.5881	790	5.58	<0.0001	0.4000	0.03761
t = 3	Without lesion	2.0665	0.3414	790	6.05	<0.0001	0.3929	0.03755
t = 3	Moderate lesion	2.4159	0.3300	790	7.32	<0.0001	0.5573	0.03762
t = 4	Without lesion	4.8525	1.0059	790	4.82	<0.0001	0.6433	0.03687
t = 4	Moderate lesion	4.2485	1.0071	790	4.22	<0.0001	0.3517	0.03669

under the “Mean” column of Table 8.22. The probability that a chick has no footpad lesion when receiving treatment 1 is $\hat{\pi}_{01} = 0.315$, whereas the value 0.57 corresponds to the cumulative probability $\hat{\pi}_{01} + \hat{\pi}_{11}$. From this value, we can calculate the probability of observing a moderate lesion, which is $\hat{\pi}_{11} = 0.57 - \hat{\pi}_{01} = 0.57 - 0.315 = 0.255$. From these probabilities, we can estimate the probability of observing a severe footpad lesion under treatment 1 as $\hat{\pi}_{21} = 1 - (0.57) = 0.43$. Following the same logic, we can estimate the reference probabilities for the rest of the other treatments.

Another important result is the odds ratio estimates. These estimates are shown in Table 8.23.

These odds ratios compare the odds for the labeled category to those for the reference category for treatments 1–3 relative to treatment 4. These odds ratio values are derived from the estimated probabilities in each of the categories. For example, the probabilities that a chicken does not present a lesion and a moderate lesion are $\hat{\pi}_{04} = 0.6433$ and $\hat{\pi}_{14} = 0.3517$, respectively. From these probabilities, we can estimate the probability of observing a severe lesion as follows: $\hat{\pi}_{24} = 1 - (0.6433 + 0.3517) = 0.005$. The estimated odds ratio of not observing a lesion (“Without” lesion) between treatments 1 and 4 is

$$\text{Odds ratio}_{T_{t1}, T_{t4}} = \frac{\hat{\pi}_{01}}{\hat{\pi}_{21}} / \frac{\hat{\pi}_{04}}{\hat{\pi}_{24}} = \frac{0.315}{0.115} / \frac{0.6433}{0.005} = 0.0213$$

the value provided in the odds ratio estimates table. If we compare the analysis using the cumulative logit link and the generalized logit link, we observe insignificant

Table 8.23 Estimated odds ratio

Odds ratio estimates						
Category	Trt	_trt	Estimate	DF	95% Confidence limits	
Without lesion	1	4	0.021	790	0.003	0.163
Moderate lesion	1	4	0.071	790	0.009	0.538
Without lesion	2	4	0.305	790	0.031	2.979
Moderate lesion	2	4	0.381	790	0.039	3.759
Without lesion	3	4	0.062	790	0.008	0.493
Moderate lesion	3	4	0.160	790	0.020	1.281

changes in the estimated category probabilities by treatment as well as in the significance level in the test of treatment effects.

8.6.2 CRD: Cheese Tasting

Consider a study in which you want to know the effects of various additives on the flavor of cheese. Researchers tested 4 cheese additives and obtained 52 response ratings for each additive. Each response was measured on a scale of 9 categories ranging from: I dislike it very much (1) to I like it very much or excellent flavor (9). Data are obtained from the study by McCullagh and Nelder (1989) (Table 8.24).

The components of the GLMM with an ordinal multinomial response are as follows:

Distributions: $y_{1i}, y_{2i}, y_{3i}, y_{4i}, y_{5i}, y_{6i}, y_{7i}, y_{8i}, y_{9i}$ ~ Multinomial ($N_i, \pi_{1i}, \pi_{2i}, \pi_{3i}, \pi_{4i}, \pi_{5i}, \pi_{6i}, \pi_{7i}, \pi_{8i}, \pi_{9i}$), where $y_{1i}, y_{2i}, y_{3i}, y_{4i}, y_{5i}, y_{6i}, y_{7i}, y_{8i}$, and y_{9i} are the observed frequencies of the responses in each category c of the hedonic scale (1 = very undesirable, ..., 5 = neither desirable nor undesirable, ..., 9 = very desirable).

Linear predictor: $\eta_{(c)i} = \eta_c + \alpha_i$, where $\eta_{(c)ij}$ is c th link ($c = 1, 2, \dots, 8$) for the additive type i , η_c is the intercept for the c th link, and α_i is the fixed effect due to the i th additive. The link functions for each category are as follows:

$$\log\left(\frac{\pi_{1i}}{1 - \pi_{1i}}\right) = \eta_{1i}$$

$$\log\left(\frac{\pi_{1i} + \pi_{2i}}{1 - (\pi_{1i} + \pi_{2i})}\right) = \eta_{2i}$$

$$\log\left(\frac{\pi_{1i} + \pi_{2i} + \pi_{3i}}{1 - (\pi_{1i} + \pi_{2i} + \pi_{3i})}\right) = \eta_{3i}$$

Table 8.24 Effect of additives on cheese flavor

id	Additive	Y	Freq	id	Additive	Y	Freq
1	1	1	0	19	3	1	1
2	1	2	0	20	3	2	1
3	1	3	1	21	3	3	6
4	1	4	7	22	3	4	8
5	1	5	8	23	3	5	23
6	1	6	8	24	3	6	7
7	1	7	19	25	3	7	5
8	1	8	8	26	3	8	1
9	1	9	1	27	3	9	0
10	2	1	6	28	4	1	0
11	2	2	9	29	4	2	0
12	2	3	12	30	4	3	0
13	2	4	11	31	4	4	1
14	2	5	7	32	4	5	3
15	2	6	6	33	4	6	7
16	2	7	1	34	4	7	14
17	2	8	0	35	4	8	16
18	2	9	0	36	4	9	11

$$\log\left(\frac{\pi_{1i} + \pi_{2i} + \pi_{3i} + \pi_{4i}}{1 - (\pi_{1i} + \pi_{2i} + \pi_{3i} + \pi_{4i})}\right) = \eta_{4i}$$

$$\log\left(\frac{\pi_{1i} + \pi_{2i} + \pi_{3i} + \pi_{4i} + \pi_{5i}}{1 - (\pi_{1i} + \pi_{2i} + \pi_{3i} + \pi_{4i} + \pi_{5i})}\right) = \eta_{5i}$$

$$\log\left(\frac{\pi_{1i} + \pi_{2i} + \pi_{3i} + \pi_{4i} + \pi_{5i} + \pi_{6i}}{1 - (\pi_{1i} + \pi_{2i} + \pi_{3i} + \pi_{4i} + \pi_{5i} + \pi_{6i})}\right) = \eta_{6i}$$

$$\log\left(\frac{\pi_{1i} + \pi_{2i} + \pi_{3i} + \pi_{4i} + \pi_{5i} + \pi_{6i} + \pi_{7i}}{1 - (\pi_{1i} + \pi_{2i} + \pi_{3i} + \pi_{4i} + \pi_{5i} + \pi_{6i} + \pi_{7i})}\right) = \eta_{7i}$$

$$\log\left(\frac{\pi_{1i} + \pi_{2i} + \pi_{3i} + \pi_{4i} + \pi_{5i} + \pi_{6i} + \pi_{7i} + \pi_{8i}}{1 - (\pi_{1i} + \pi_{2i} + \pi_{3i} + \pi_{4i} + \pi_{5i} + \pi_{6i} + \pi_{7i} + \pi_{8i})}\right) = \eta_{8i}$$

The following GLIMMIX commands fit a cumulative logit model with an ordinal multinomial response.

```
proc glimmix ;
class id additive scale;
model scale(order=data)= additive/dist=Multinomial link=clogit
solution oddsratio;
estimate 'c=1, a=1' intercept 1 0 0 0 0 0 0 0 additive 1 0 0 0,
'c=2, a=1' intercept 0 1 0 0 0 0 0 0 additive 1 0 0 0,
'c=3, a=1' intercept 0 0 1 0 0 0 0 0 additive 1 0 0 0,
'c=4, a=1' intercept 0 0 0 1 0 0 0 0 additive 1 0 0 0,
```


Table 8.25 Fixed effects tests in the multinomial cumulative logit model

Type III tests of fixed effects				
Effect	Num DF	Den DF	F-value	Pr > F
Additive	3	197	38.11	<0.0001

```
'c=5, a=1' intercept 0 0 0 0 1 0 0 0 additive 1 0 0 0,
'c=6, a=1' intercept 0 0 0 0 0 1 0 0 additive 1 0 0 0,
'c=7, a=1' intercept 0 0 0 0 0 0 1 0 additive 1 0 0 0,
'c=8, a=1' intercept 0 0 0 0 0 0 0 1 additive 1 0 0 0,
'c=1, a=2' intercept 1 0 0 0 0 0 0 0 additive 0 1 0 0,
'c=2, a=2' intercept 0 1 0 0 0 0 0 0 additive 0 1 0 0,
'c=3, a=2' intercept 0 0 1 0 0 0 0 0 additive 0 1 0 0,
'c=4, a=2' intercept 0 0 0 1 0 0 0 0 additive 0 1 0 0,
'c=5, a=2' intercept 0 0 0 0 1 0 0 0 additive 0 1 0 0,
'c=6, a=2' intercept 0 0 0 0 0 1 0 0 additive 0 1 0 0,
'c=7, a=2' intercept 0 0 0 0 0 0 1 0 additive 0 1 0 0,
'c=8, a=2' intercept 0 0 0 0 0 0 0 1 additive 0 1 0 0,
'c=1, a=3' intercept 1 0 0 0 0 0 0 0 additive 0 0 1 0,
'c=2, a=3' intercept 0 1 0 0 0 0 0 0 additive 0 0 1 0,
'c=3, a=3' intercept 0 0 1 0 0 0 0 0 additive 0 0 1 0,
'c=4, a=3' intercept 0 0 0 1 0 0 0 0 additive 0 0 1 0,
'c=5, a=3' intercept 0 0 0 0 1 0 0 0 additive 0 0 1 0,
'c=6, a=3' intercept 0 0 0 0 0 1 0 0 additive 0 0 1 0,
'c=7, a=3' intercept 0 0 0 0 0 0 1 0 additive 0 0 1 0,
'c=8, a=3' intercept 0 0 0 0 0 0 0 1 additive 0 0 1 0,
'c=1, a=4' intercept 1 0 0 0 0 0 0 0 additive 0 0 0 1,
'c=2, a=4' intercept 0 1 0 0 0 0 0 0 additive 0 0 0 1,
'c=3, a=4' intercept 0 0 1 0 0 0 0 0 additive 0 0 0 1,
'c=4, a=4' intercept 0 0 0 1 0 0 0 0 additive 0 0 0 1,
'c=5, a=4' intercept 0 0 0 0 1 0 0 0 additive 0 0 0 1,
'c=6, a=4' intercept 0 0 0 0 0 1 0 0 additive 0 0 0 1,
'c=7, a=4' intercept 0 0 0 0 0 0 1 0 additive 0 0 0 1,
'c=8, a=4' intercept 0 0 0 0 0 0 0 1 additive 0 0 0 1/ilink;
freq freq;
run;
```

Part of the results is shown in Table 8.25. The results of the analysis of variance show that the type of additive used in the manufacture of cheese significantly affects the degree of consumer acceptance ($P = 0.0001$). That is, the type of additive affects the sensory characteristics of the cheese.

The contrast of hypothesis are presented in Table 8.26. The hypothesis tests are as follows:

$$\pi_{\text{additive}_i} = \pi_{\text{additive}_j}; \forall i \neq j$$

The results show that the additives provide different sensory characteristics that are reflected in the evaluation of preference.

With the “solution” option in the model specification, Table 8.27 (fixed parameter estimates) shows the solution of the maximum likelihood estimates for the fixed

Table 8.26 Contrast of hypothesis in the acceptance of cheese made with four additives

Contrasts				
Label	Num DF	Den DF	F-value	Pr > F
Additive effect 1 vs. 2	1	197	61.13	<0.0001
Additive effect 1 vs. 3	1	197	21.19	<0.0001
Additive effect 2 vs. 3	1	197	19.14	<0.0001
Additive effect 2 vs. 4	1	197	108.45	<0.0001
Additive effect 3 vs. 4	1	197	62.04	<0.0001

Table 8.27 Maximum likelihood estimates of the fixed effects in the preference ratings of cheese made with different types of additives in the multinomial cumulative logit model

Fixed parameter estimates							
Effect	escala	Additive	Estimate	Standard error	DF	t-value	Pr > t
Intercept $\hat{\eta}_1$	1		-7.0802	0.5640	197	-12.55	<0.0001
Intercept $\hat{\eta}_2$	2		-6.0250	0.4764	197	-12.65	<0.0001
Intercept $\hat{\eta}_3$	3		-4.9254	0.4257	197	-11.57	<0.0001
Intercept $\hat{\eta}_4$	4		-3.8568	0.3880	197	-9.94	<0.0001
Intercept $\hat{\eta}_5$	5		-2.5206	0.3453	197	-7.30	<0.0001
Intercept $\hat{\eta}_6$	6		-1.5685	0.3122	197	-5.02	<0.0001
Intercept $\hat{\eta}_7$	7		-0.06688	0.2738	197	-0.24	0.8073
Intercept $\hat{\eta}_8$	8		1.4930	0.3357	197	4.45	<0.0001
Aditivo $\hat{\alpha}_1$	1		1.6128	0.3805	197	4.24	<0.0001
Aditivo $\hat{\alpha}_2$	2		4.9646	0.4767	197	10.41	<0.0001
Aditivo $\hat{\alpha}_3$	3		3.3227	0.4218	197	7.88	<0.0001
Aditivo $\hat{\alpha}_4$	4		0

effects parameters. In this table, we observe the values of the estimated intercepts: $\hat{\eta}_1 = -7.0802$ defines the boundary between categories “1” and “2,” whereas $\hat{\eta}_2 = -6.0250$ defines the boundary between categories “2” and “3.” The third intercept, $\hat{\eta}_3 = -4.9254$, defines the boundary between categories “3” and “4” and so forth. The estimated effects of the additive type ($\hat{\alpha}_i, i = 1, 2, 3,$ and 4) are 1.628, 4.9646, 3.3227, and 0, respectively. From these values, linear predictors are estimated for each of the categories.

For example, the estimated linear predictor for a cheese made with additive 1, where the evaluator (consumer) assigns it category “1 = highly undesirable,” is represented as $\hat{\eta}_{11} = \hat{\eta}_1 + \hat{\alpha}_1 = -7.0802 + 1.6128 = -5.4674$; for the category “2 = moderately undesirable,” it is $\hat{\eta}_{21} = \hat{\eta}_2 + \hat{\alpha}_1 = -6.0250 + 1.6128 = -4.4122$; for the category “3 = slightly undesirable,” it is $\hat{\eta}_{31} = \hat{\eta}_3 + \hat{\alpha}_1 = -4.9254 + 1.6128 = -3.3126$; and for the category “4 = neither undesirable nor desirable,” it is $\hat{\eta}_{41} = \hat{\eta}_4 + \hat{\alpha}_1 = -3.8568 + 1.6128 = -2.2440$. These values are shown in the “Estimate” column of Table 8.28; other categories are similarly calculated for each type of additive.

The estimated values in Table 8.27 obtained with the “estimate” command in conjunction with the “ilink” option prompts GLIMMIX to calculate the values of the

linear predictors $\hat{\eta}_{Ci}$ tabulated in the “Estimate” column and estimated probabilities $\hat{\pi}_{Ci}$ of all categories of each treatment, tabulated in the “Mean” column ($\hat{\pi}_{cij}$), except for the reference category.

From Table 8.28 (Estimates), we obtain the probabilities for each category that is reported under the “Mean” column. In this case, the probability for $\hat{\pi}_{11} = 0.004205$. This value is obtained by taking the inverse value of the linear predictor $\hat{\eta}_{11} = -5.4674$ ($\hat{\pi}_{11} = 1 / (1 + \exp^{(5.4674)}) = 0.004205$). To calculate the probability that a panelist would assign a rating of 2 (2 = moderately undesirable) to cheese made with additive 1, we use the cumulative probability of 0.01198, which corresponds to $\hat{\pi}_{21} + \hat{\pi}_{11}$. From this value, we obtain $\hat{\pi}_{21} = 0.01198 - \hat{\pi}_{11} = 0.01198 - 0.004205 = 0.007775$ and for the probability of assigning a rating of 3 to cheese made with additive 1, $\hat{\pi}_{31} = 0.03514 - (\hat{\pi}_{21} + \hat{\pi}_{11}) = 0.03514 - 0.01198 = 0.033942$. Following the same procedure, we obtain the other probabilities for the rest of the categories of each of the additives used in the manufacturing of cheese, which are tabulated in Table 8.29 and can be seen in Fig. 8.3.

Figure 8.3 shows the probability results of each flavor rating for each of the additives (it should be noted that some probability values were suppressed to avoid overwriting). It can be seen that additive 1 primarily receives ratings of 5–7; additive 2 primarily receives ratings of 2–5; additive 3 primarily receives ratings of 4–6; and additive 4 primarily receives ratings of 7–9.

The odds ratio results (Table 8.30) show the preferences more clearly. For example, the odds ratio additive 1 vs. 4 states that the first additive is 5.017 times more likely to receive a lower score than the fourth additive.

8.7 Exercises

Exercise 8.7.1 The dataset for this exercise corresponds to the results of 9 judges who rated 2 classes of wine, namely, white wine (WW = 1) and red wine (RW = 2), and, within each wine class, they rated 10 wines on a scale of 1–20 points. The minimum rating for a particular wine was 7, and the maximum rating was 19.5. For didactic purposes, ratings between 7 and 11 were assigned low quality, a rating between 12 and 15 as medium quality, and anything above 15 was considered excellent quality. The data are shown in Table 8.31 of the wine evaluation experiment under columns “Judge” (wine evaluator panelist), “Wine_type” (white wine: 1, red wine: 2), “Quality” (low, medium, and excellent), and the frequency of the observed qualities (“y”).

- (a) Fit the cumulative logit proportional odds model to these data. Perform a complete and appropriate analysis of the data, focusing on:
 - (i) An evaluation of the effects of the combination of treatments
 - (ii) Interpretation of the odds ratios
 - (iii) The expected probability per category for each treatment

Table 8.28 Estimates on the model scale (Estimate) and on the data scale (Mean) based on judges' preference ratings of cheese made with different types of additives in the multinomial cumulative logit model

Estimates							
Label	Estimate	Standard error	DF	<i>t</i> -value	Pr > <i>t</i>	Mean	Standard error mean
<i>c</i> = 1, <i>a</i> = 1	-5.4674	0.5236	197	-10.44	<0.0001	0.004205	0.002192
<i>c</i> = 2, <i>a</i> = 1	-4.4122	0.4278	197	-10.31	<0.0001	0.01198	0.005064
<i>c</i> = 3, <i>a</i> = 1	-3.3126	0.3700	197	-8.95	<0.0001	0.03514	0.01255
<i>c</i> = 4, <i>a</i> = 1	-2.2440	0.3267	197	-6.87	<0.0001	0.09587	0.02832
<i>c</i> = 5, <i>a</i> = 1	-0.9078	0.2833	197	-3.20	0.0016	0.2875	0.05804
<i>c</i> = 6, <i>a</i> = 1	0.04425	0.2646	197	0.17	0.8673	0.5111	0.06611
<i>c</i> = 7, <i>a</i> = 1	1.5459	0.3017	197	5.12	<0.0001	0.8243	0.04369
<i>c</i> = 8, <i>a</i> = 1	3.1058	0.4057	197	7.65	<0.0001	0.9571	0.01665
<i>c</i> = 1, <i>a</i> = 2	-2.1155	0.4106	197	-5.15	<0.0001	0.1076	0.03942
<i>c</i> = 2, <i>a</i> = 2	-1.0603	0.3009	197	-3.52	0.0005	0.2572	0.05749
<i>c</i> = 3, <i>a</i> = 2	0.03922	0.2735	197	0.14	0.8861	0.5098	0.06836
<i>c</i> = 4, <i>a</i> = 2	1.1078	0.2969	197	3.73	0.0002	0.7517	0.05542
<i>c</i> = 5, <i>a</i> = 2	2.4441	0.3397	197	7.19	<0.0001	0.9201	0.02497
<i>c</i> = 6, <i>a</i> = 2	3.3961	0.3724	197	9.12	<0.0001	0.9676	0.01168
<i>c</i> = 7, <i>a</i> = 2	4.8978	0.4249	197	11.53	<0.0001	0.9926	0.003124
<i>c</i> = 8, <i>a</i> = 2	6.4576	0.5045	197	12.80	<0.0001	0.9984	0.000789
<i>c</i> = 1, <i>a</i> = 3	-3.7575	0.4761	197	-7.89	<0.0001	0.02281	0.01061
<i>c</i> = 2, <i>a</i> = 3	-2.7023	0.3677	197	-7.35	<0.0001	0.06284	0.02165
<i>c</i> = 3, <i>a</i> = 3	-1.6027	0.3001	197	-5.34	<0.0001	0.1676	0.04186
<i>c</i> = 4, <i>a</i> = 3	-0.5341	0.2556	197	-2.09	0.0379	0.3696	0.05955
<i>c</i> = 5, <i>a</i> = 3	0.8021	0.2610	197	3.07	0.0024	0.6904	0.05579

(continued)

Table 8.28 (continued)

Estimates							
Label	Estimate	Standard error	DF	<i>t</i> -value	Pr > <i>t</i>	Mean	Standard error mean
<i>c</i> = 6, <i>a</i> = 3	1.7541	0.2984	197	5.88	<0.0001	0.8525	0.03752
<i>c</i> = 7, <i>a</i> = 3	3.2558	0.3618	197	9.00	<0.0001	0.9629	0.01293
<i>c</i> = 8, <i>a</i> = 3	4.8157	0.4528	197	10.63	<0.0001	0.9920	0.003610
<i>c</i> = 1, <i>a</i> = 4	-7.0802	0.5640	197	-12.55	<0.0001	0.000841	0.000474
<i>c</i> = 2, <i>a</i> = 4	-6.0250	0.4764	197	-12.65	<0.0001	0.002412	0.001146
<i>c</i> = 3, <i>a</i> = 4	-4.9254	0.4257	197	-11.57	<0.0001	0.007207	0.003046
<i>c</i> = 4, <i>a</i> = 4	-3.8568	0.3880	197	-9.94	<0.0001	0.02070	0.007865
<i>c</i> = 5, <i>a</i> = 4	-2.5206	0.3453	197	-7.30	<0.0001	0.07443	0.02379
<i>c</i> = 6, <i>a</i> = 4	-1.5685	0.3122	197	-5.02	<0.0001	0.1724	0.04455
<i>c</i> = 7, <i>a</i> = 4	-0.06688	0.2738	197	-0.24	0.8073	0.4833	0.06838
<i>c</i> = 8, <i>a</i> = 4	1.4930	0.3357	197	4.45	<0.0001	0.8165	0.05029

Exercise 8.7.2 Data were obtained from a series of experiments conducted to reduce damage to potato tubers due to a potato lifter. The experiments were conducted at the Institute of Agricultural Engineering (IMAG-DLO) in Wageningen, the Netherlands. One source of damage was the type of rod used in the lifter. In the experiment – under consideration – eight types of rods were compared. It is an empirical fact that the degree of damage varies considerably between potato varieties with the type of rope used in the lifting of full potato sacks. Three blocks of observations were obtained for the combinations of varieties and rope types. Most of the combinations involved about 20 potatoes. For some combinations, there are no data due to an insufficient number of large potatoes. Tubers were dropped from a given height. To determine the damage, all tubers were peeled and the degree of blue coloration was classified into one of four classes (class 1: no damage; class 2: slight damage; class 3: moderate damage; and class 4: severe damage). The observations, in the form of counts per class and combination, are shown in Table 8.32 of the tuber experiment whose columns are “Variety” (1, 2, 3, 4, 4, 5, 6), “String” (1, 2, 3, 4, 5, 6, 7, 8), “Block” (1, 2, 3), Type of damage (sd = no damage, dl = light damage, dm = moderate damage, ds = severe damage), and the observed frequency (*Y*).

Table 8.29 Probabilities calculated for each of the ratings by additives used in the manufacture of cheese

	Rating								
	Cal1	Cal2	Cal3	Cal4	Cal5	Cal6	Cal7	Cal8	Cal9
Additive 1	0.00421	0.00778	0.02316	0.06073	0.19163	0.22336	0.3132	0.1328	0.0429
Additive 2	0.1076	0.1496	0.2526	0.2419	0.1684	0.0475	0.025	0.0058	0.0016
Additive 3	0.02281	0.04003	0.10476	0.202	0.3208	0.1621	0.1104	0.0291	0.008
Additive 4	0.00084	0.00157	0.0048	0.01349	0.05373	0.09797	0.3109	0.3332	0.1835

Note: Grade1 = Grade 1, Grade2 = Grade 2,....., Grade9 = Grade 9

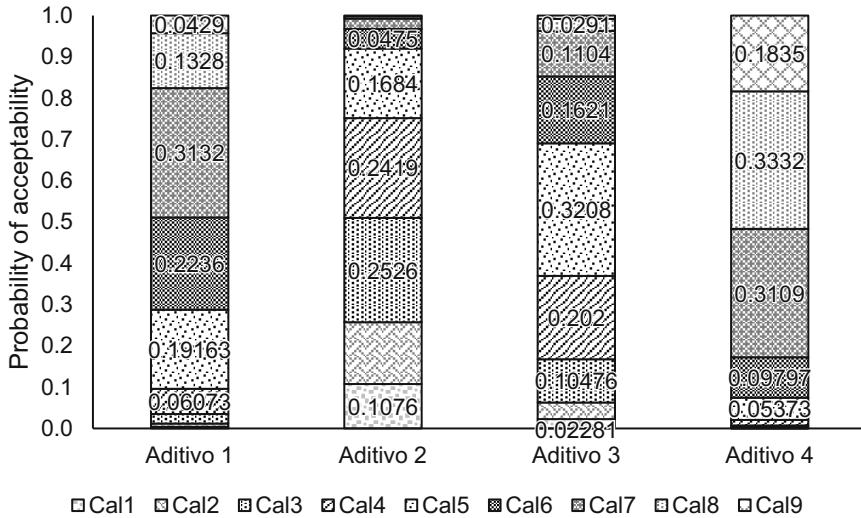


Fig. 8.3 Estimated probabilities for the categories of acceptability for the cheese according to the type of additive

Table 8.30 Odds ratio

Odds ratio estimates					
Additivo	_Additivo	Estimate	DF	95% Confidence limits	
1	4	5.017	197	2.369	10.625
2	4	143.257	197	55.953	366.783
3	4	27.735	197	12.071	63.724

- (a) List the components of the multinomial GLMM.
- (b) Fit the cumulative logit proportional odds model to these data. Perform a complete and appropriate analysis of the data, focusing on:
 - (i) An evaluation of the effects of the combination of treatments
 - (ii) Interpretation of the odds ratios
 - (iii) The expected probability per category for each treatment
- (c) Test whether the proportional odds assumption is viable. Cite relevant evidence to support your conclusion regarding the adequacy of the assumption.
- (d) If as a result of b), you consider that an alternative cumulative logit model is better, then revise your analysis in a) accordingly.

Exercise 8.7.3 Fit a generalized multinomial logit model using the dataset from Exercise 8.7.2 of this section, following the instructions:

Table 8.31 Results of the wine evaluation experiment

Judge	Wine_type	Quality	y
1	1	Low	4
1	1	Medium	6
1	1	Excellent	0
1	2	Low	3
1	2	Medium	6
1	2	Excellent	1
2	1	Low	3
2	1	Medium	5
2	1	Excellent	2
2	2	Low	2
2	2	Medium	7
2	2	Excellent	1
3	1	Low	0
3	1	Medium	4
3	1	Excellent	6
3	2	Low	0
3	2	Medium	1
3	2	Excellent	9
4	1	Low	1
4	1	Medium	3
4	1	Excellent	6
4	2	Low	4
4	2	Medium	3
4	2	Excellent	3
5	1	Low	6
5	1	Medium	3
5	1	Excellent	1
5	2	Low	3
5	2	Medium	5
5	2	Excellent	2
6	1	Low	0
6	1	Medium	4
6	1	Excellent	6
6	2	Low	0
6	2	Medium	6
6	2	Excellent	4
7	1	Low	1
7	1	Medium	9
7	1	Excellent	0
7	2	Low	3
7	2	Medium	5
7	2	Excellent	2

(continued)

Table 8.31 (continued)

Judge	Wine_type	Quality	y
8	1	Low	0
8	1	Medium	7
8	1	Excellent	3
8	2	Low	0
8	2	Medium	6
8	2	Excellent	4
9	1	Low	0
9	1	Medium	5
9	1	Excellent	5
9	2	Low	2
9	2	Medium	4
9	2	Excellent	4

- (a) List the components of this model.
- (b) Perform a thorough and appropriate analysis of the data, focusing on:
- (i) An evaluation of the main effects and treatment interaction
 - (ii) Odds ratio interpretation
 - (iii) The expected probability per category for each treatment
- (c) Comment on and discuss your results. Cite relevant evidence to support your conclusion regarding the adequacy of the assumption.

Exercise 8.7.4 In this exercise, the effects of judges' experience on quality ratings of canned beans are assessed. Canning quality is one of the most essential traits required in all new dry bean (*Phaseolus vulgaris* L.) varieties, and selection for this trait is a critical part of bean breeding programs. Advanced lines, which are candidates for release as varieties, must be evaluated for canning quality for at least 3 years from samples grown at different locations. Quality is evaluated by a panel of judges with varying levels of experience in evaluating breeding lines for visual quality traits. In all, 264 bean breeding lines from 4 commercial classes were conserved according to the procedures described by Walters et al. (1997).

These included 62 white (navy), 65 black, 55 kidney, and 82 pinto bean lines plus control lines and "checks." The visual appearance of the processed beans was determined subjectively by a panel of 13 judges on a 7-point hedonic scale (1 = very undesirable, 4 = neither desirable nor undesirable, 7 = very desirable). The beans were presented to the panel of judges in random order at the same time. Prior to evaluating the samples, all judges were shown examples of samples rated as satisfactory (4). There is concern that certain judges, due to lack of experience, may not be able to score canned samples correctly.

From attribute-based product evaluations, inferences about the effects of experience can be drawn from the psychology literature (Wallsten and Budescu (1981). Prior to the bean canning quality rating experiment, it was postulated that not only do

Table 8.32 Results of the tuber experiment. V = variety, C = string, B = block, D = damage (sd = no damage, dl = slight damage, dm = moderate damage, ds = severe damage), and Y = observed frequency

V	C	B	D	Y	V	C	B	D	Y	V	C	B	D	Y
1	1	1	sd	5	2	1	1	sd	4	3	1	1	sd	3
1	1	1	dl	14	2	1	1	dl	5	3	1	1	dl	2
1	1	1	dm	1	2	1	1	dm	4	3	1	1	dm	8
1	1	1	ds	0	2	1	1	ds	7	3	1	1	ds	7
1	2	1	sd	6	2	2	1	sd	8	3	2	1	sd	18
1	2	1	dl	11	2	2	1	dl	3	3	2	1	dl	1
1	2	1	dm	1	2	2	1	dm	0	3	2	1	dm	0
1	2	1	ds	0	2	2	1	ds	0	3	2	1	ds	0
1	3	1	sd	6	2	3	1	sd	3	3	3	1	sd	5
1	3	1	dl	13	2	3	1	dl	10	3	3	1	dl	7
1	3	1	dm	0	2	3	1	dm	6	3	3	1	dm	4
1	3	1	ds	0	2	3	1	ds	1	3	3	1	ds	4
1	4	1	sd	2	2	4	1	sd	1	3	4	1	sd	1
1	4	1	dl	9	2	4	1	dl	3	3	4	1	dl	4
1	4	1	dm	6	2	4	1	dm	11	3	4	1	dm	6
1	4	1	ds	3	2	4	1	ds	5	3	4	1	ds	9
1	5	1	sd	11	2	5	1	sd	16	3	5	1	sd	12
1	5	1	dl	8	2	5	1	dl	3	3	5	1	dl	7
1	5	1	dm	0	2	5	1	dm	1	3	5	1	dm	1
1	5	1	ds	0	2	5	1	ds	0	3	5	1	ds	0
1	6	1	sd	12	2	6	1	sd	16	3	6	1	sd	16
1	6	1	dl	5	2	6	1	dl	3	3	6	1	dl	3
1	6	1	dm	2	2	6	1	dm	0	3	6	1	dm	0
1	6	1	ds	0	2	6	1	ds	0	3	6	1	ds	1
1	7	1	sd	8	2	7	1	sd	11	3	7	1	sd	20
1	7	1	dl	12	2	7	1	dl	9	3	7	1	dl	0
1	7	1	dm	0	2	7	1	dm	0	3	7	1	dm	0
1	7	1	ds	0	2	7	1	ds	0	3	7	1	ds	0
1	8	1	sd	12	2	8	1	sd	10	3	8	1	sd	18
1	8	1	dl	4	2	8	1	dl	10	3	8	1	dl	2
1	8	1	dm	0	2	8	1	dm	0	3	8	1	dm	0
1	8	1	ds	0	2	8	1	ds	0	3	8	1	ds	0
1	1	2	sd	5	2	1	2	sd	5	3	1	2	sd	6
1	1	2	dl	31	2	1	2	dl	7	3	1	2	dl	8
1	1	2	dm	2	2	1	2	dm	5	3	1	2	dm	5
1	1	2	ds	1	2	1	2	ds	1	3	1	2	ds	1
1	2	2	sd	6	2	2	2	sd	13	3	2	2	sd	12
1	2	2	dl	11	2	2	2	dl	6	3	2	2	dl	6
1	2	2	dm	1	2	2	2	dm	1	3	2	2	dm	1
1	2	2	ds	0	2	2	2	ds	0	3	2	2	ds	1
1	3	2	sd	5	2	3	2	sd	5	3	3	2	sd	10

(continued)

Table 8.32 (continued)

V	C	B	D	Y	V	C	B	D	Y	V	C	B	D	Y
1	3	2	dl	13	2	3	2	dl	12	3	3	2	dl	8
1	3	2	dm	0	2	3	2	dm	3	3	3	2	dm	0
1	3	2	ds	0	2	3	2	ds	0	3	3	2	ds	2
1	4	2	sd	2	2	4	2	sd	0	3	4	2	sd	2
1	4	2	dl	11	2	4	2	dl	8	3	4	2	dl	6
1	4	2	dm	9	2	4	2	dm	11	3	4	2	dm	10
1	4	2	ds	0	2	4	2	ds	1	3	4	2	ds	2
1	5	2	sd	16	2	5	2	sd	10	3	5	2	sd	16
1	5	2	dl	4	2	5	2	dl	9	3	5	2	dl	4
1	5	2	dm	0	2	5	2	dm	1	3	5	2	dm	0
1	5	2	ds	0	2	5	2	ds	0	3	5	2	ds	0
1	6	2	sd	15	2	6	2	sd	10	3	6	2	sd	14
1	6	2	dl	5	2	6	2	dl	7	3	6	2	dl	5
1	6	2	dm	0	2	6	2	dm	1	3	6	2	dm	0
1	6	2	ds	0	2	6	2	ds	1	3	6	2	ds	0
1	7	2	sd	9	2	7	2	sd	11	3	7	2	sd	16
1	7	2	dl	7	2	7	2	dl	6	3	7	2	dl	3
1	7	2	dm	3	2	7	2	dm	1	3	7	2	dm	0
1	7	2	ds	0	2	7	2	ds	0	3	7	2	ds	0
1	8	2	sd	0	2	8	2	sd	13	3	8	2	sd	16
1	8	2	dl	0	2	8	2	dl	6	3	8	2	dl	3
1	8	2	dm	0	2	8	2	dm	0	3	8	2	dm	0
1	8	2	ds	0	2	8	2	ds	0	3	8	2	ds	1
1	1	3	sd	0	2	1	3	sd	9	3	1	3	sd	3
1	1	3	dl	0	2	1	3	dl	7	3	1	3	dl	15
1	1	3	dm	0	2	1	3	dm	2	3	1	3	dm	2
1	1	3	ds	1	2	1	3	ds	0	3	1	3	ds	0
1	2	3	sd	7	2	2	3	sd	18	3	2	3	sd	16
1	2	3	dl	10	2	2	3	dl	2	3	2	3	dl	6
1	2	3	dm	2	2	2	3	dm	0	3	2	3	dm	2
1	2	3	ds	0	2	2	3	ds	0	3	2	3	ds	2
1	3	3	sd	1	2	3	3	sd	13	3	3	3	sd	8
1	3	3	dl	19	2	3	3	dl	6	3	3	3	dl	9
1	3	3	dm	0	2	3	3	dm	0	3	3	3	dm	2
1	3	3	ds	0	2	3	3	ds	0	3	3	3	ds	1
1	4	3	sd	4	2	4	3	sd	0	3	4	3	sd	3
1	4	3	dl	13	2	4	3	dl	9	3	4	3	dl	5
1	4	3	dm	3	2	4	3	dm	9	3	4	3	dm	10
1	4	3	ds	0	2	4	3	ds	2	3	4	3	ds	1
1	5	3	sd	15	2	5	3	sd	16	3	5	3	sd	16
1	5	3	dl	4	2	5	3	dl	2	3	5	3	dl	3
1	5	3	dm	0	2	5	3	dm	1	3	5	3	dm	0

(continued)

Table 8.32 (continued)

V	C	B	D	Y	V	C	B	D	Y	V	C	B	D	Y
1	5	3	ds	1	2	5	3	ds	0	3	5	3	ds	1
1	6	3	sd	11	2	6	3	sd	15	3	6	3	sd	15
1	6	3	dl	4	2	6	3	dl	3	3	6	3	dl	5
1	6	3	dm	0	2	6	3	dm	2	3	6	3	dm	0
1	6	3	ds	5	2	6	3	ds	0	3	6	3	ds	0
1	7	3	sd	11	2	7	3	sd	14	3	7	3	sd	15
1	7	3	dl	9	2	7	3	dl	5	3	7	3	dl	5
1	7	3	dm	0	2	7	3	dm	1	3	7	3	dm	0
1	7	3	ds	0	2	7	3	ds	0	3	7	3	ds	0
1	8	3	sd	17	2	8	3	sd	12	3	8	3	sd	16
1	8	3	dl	2	2	8	3	dl	3	3	8	3	dl	4
1	8	3	dm	1	2	8	3	dm	2	3	8	3	dm	0
1	8	3	ds	0	2	8	3	ds	1	3	8	3	ds	0
4	1	1	sd	4	5	1	1	sd	9	6	1	1	sd	5
4	1	1	dl	9	5	1	1	dl	10	6	1	1	dl	14
4	1	1	dm	7	5	1	1	dm	1	6	1	1	dm	1
4	1	1	ds	0	5	1	1	ds	0	6	1	1	ds	0
4	2	1	sd	12	5	2	1	sd	17	6	2	1	sd	18
4	2	1	dl	8	5	2	1	dl	2	6	2	1	dl	2
4	2	1	dm	0	5	2	1	dm	0	6	2	1	dm	0
4	2	1	ds	0	5	2	1	ds	0	6	2	1	ds	0
4	3	1	sd	10	5	3	1	sd	15	6	3	1	sd	5
4	3	1	dl	10	5	3	1	dl	5	6	3	1	dl	14
4	3	1	dm	0	5	3	1	dm	0	6	3	1	dm	0
4	3	1	ds	0	5	3	1	ds	0	6	3	1	ds	0
4	4	1	sd	2	5	4	1	sd	12	6	4	1	sd	5
4	4	1	dl	8	5	4	1	dl	7	6	4	1	dl	14
4	4	1	dm	4	5	4	1	dm	0	6	4	1	dm	1
4	4	1	ds	6	5	4	1	ds	0	6	4	1	ds	0
4	5	1	sd	17	5	5	1	sd	16	6	5	1	sd	15
4	5	1	dl	2	5	5	1	dl	4	6	5	1	dl	5
4	5	1	dm	0	5	5	1	dm	0	6	5	1	dm	0
4	5	1	ds	0	5	5	1	ds	0	6	5	1	ds	0
4	6	1	sd	18	5	6	1	sd	20	6	6	1	sd	19
4	6	1	dl	2	5	6	1	dl	0	6	6	1	dl	1
4	6	1	dm	0	5	6	1	dm	0	6	6	1	dm	0
4	6	1	ds	0	5	6	1	ds	0	6	6	1	ds	0
4	7	1	sd	19	5	7	1	sd	20	6	7	1	sd	18
4	7	1	dl	1	5	7	1	dl	0	6	7	1	dl	2
4	7	1	dm	0	5	7	1	dm	0	6	7	1	dm	0
4	7	1	ds	0	5	7	1	ds	0	6	7	1	ds	0
4	8	1	sd	15	5	8	1	sd	20	6	8	1	sd	18

(continued)

Table 8.32 (continued)

V	C	B	D	Y	V	C	B	D	Y	V	C	B	D	Y
4	8	1	dl	3	5	8	1	dl	0	6	8	1	dl	2
4	8	1	dm	0	5	8	1	dm	0	6	8	1	dm	0
4	8	1	ds	1	5	8	1	ds	0	6	8	1	ds	0
4	1	2	sd	4	5	1	2	sd	10	6	1	2	sd	6
4	1	2	dl	11	5	1	2	dl	10	6	1	2	dl	13
4	1	2	dm	3	5	1	2	dm	0	6	1	2	dm	0
4	1	2	ds	1	5	1	2	ds	0	6	1	2	ds	0
4	2	2	sd	17	5	2	2	sd	19	6	2	2	sd	18
4	2	2	dl	2	5	2	2	dl	1	6	2	2	dl	2
4	2	2	dm	0	5	2	2	dm	0	6	2	2	dm	0
4	2	2	ds	1	5	2	2	ds	0	6	2	2	ds	0
4	3	2	sd	10	5	3	2	sd	14	6	3	2	sd	13
4	3	2	dl	9	5	3	2	dl	6	6	3	2	dl	7
4	3	2	dm	0	5	3	2	dm	0	6	3	2	dm	0
4	3	2	ds	0	5	3	2	ds	0	6	3	2	ds	0
4	4	2	sd	4	5	4	2	sd	7	6	4	2	sd	0
4	4	2	dl	8	5	4	2	dl	11	6	4	2	dl	15
4	4	2	dm	6	5	4	2	dm	1	6	4	2	dm	5
4	4	2	ds	1	5	4	2	ds	0	6	4	2	ds	0
4	5	2	sd	18	5	5	2	sd	15	6	5	2	sd	16
4	5	2	dl	2	5	5	2	dl	4	6	5	2	dl	3
4	5	2	dm	0	5	5	2	dm	1	6	5	2	dm	0
4	5	2	ds	0	5	5	2	ds	0	6	5	2	ds	0
4	6	2	sd	19	5	6	2	sd	19	6	6	2	sd	19
4	6	2	dl	1	5	6	2	dl	1	6	6	2	dl	1
4	6	2	dm	0	5	6	2	dm	0	6	6	2	dm	0
4	6	2	ds	0	5	6	2	ds	0	6	6	2	ds	0
4	7	2	sd	20	5	7	2	sd	17	6	7	2	sd	17
4	7	2	dl	0	5	7	2	dl	2	6	7	2	dl	3
4	7	2	dm	0	5	7	2	dm	0	6	7	2	dm	0
4	7	2	ds	0	5	7	2	ds	0	6	7	2	ds	0
4	8	2	sd	20	5	8	2	sd	18	6	8	2	sd	15
4	8	2	dl	0	5	8	2	dl	2	6	8	2	dl	4
4	8	2	dm	0	5	8	2	dm	0	6	8	2	dm	0
4	8	2	ds	0	5	8	2	ds	0	6	8	2	ds	0
4	1	3	sd	10	5	1	3	sd	5	6	1	3	sd	3
4	1	3	dl	9	5	1	3	dl	11	6	1	3	dl	15
4	1	3	dm	1	5	1	3	dm	4	6	1	3	dm	2
4	1	3	ds	0	5	1	3	ds	0	6	1	3	ds	0
4	2	3	sd	16	5	2	3	sd	12	6	2	3	sd	13
4	2	3	dl	3	5	2	3	dl	8	6	2	3	dl	7
4	2	3	dm	1	5	2	3	dm	0	6	2	3	dm	0

(continued)

Table 8.32 (continued)

V	C	B	D	Y	V	C	B	D	Y	V	C	B	D	Y
4	2	3	ds	0	5	2	3	ds	0	6	2	3	ds	0
4	3	3	sd	15	5	3	3	sd	7	6	3	3	sd	2
4	3	3	dl	4	5	3	3	dl	12	6	3	3	dl	18
4	3	3	dm	0	5	3	3	dm	1	6	3	3	dm	0
4	3	3	ds	0	5	3	3	ds	0	6	3	3	ds	0
4	4	3	sd	5	5	4	3	sd	6	6	4	3	sd	4
4	4	3	dl	11	5	4	3	dl	6	6	4	3	dl	10
4	4	3	dm	4	5	4	3	dm	6	6	4	3	dm	6
4	4	3	ds	0	5	4	3	ds	2	6	4	3	ds	0
4	5	3	sd	17	5	5	3	sd	16	6	5	3	sd	6
4	5	3	dl	2	5	5	3	dl	4	6	5	3	dl	13
4	5	3	dm	0	5	5	3	dm	0	6	5	3	dm	1
4	5	3	ds	0	5	5	3	ds	0	6	5	3	ds	0
4	6	3	sd	16	5	6	3	sd	17	6	6	3	sd	18
4	6	3	dl	2	5	6	3	dl	3	6	6	3	dl	2
4	6	3	dm	0	5	6	3	dm	0	6	6	3	dm	0
4	6	3	ds	0	5	6	3	ds	0	6	6	3	ds	0
4	7	3	sd	17	5	7	3	sd	18	6	7	3	sd	15
4	7	3	dl	2	5	7	3	dl	2	6	7	3	dl	5
4	7	3	dm	1	5	7	3	dm	0	6	7	3	dm	0
4	7	3	ds	0	5	7	3	ds	0	6	7	3	ds	0
4	8	3	sd	17	5	8	3	sd	17	6	8	3	sd	19
4	8	3	dl	2	5	8	3	dl	2	6	8	3	dl	1
4	8	3	dm	0	5	8	3	dm	0	6	8	3	dm	0
4	8	3	ds	0	5	8	3	ds	0	6	8	3	ds	0

less experienced judges have a more severe rating than do more experienced judges but also that experience should have little or no effect on the white beans for which the canning procedure was developed. Judges are stratified for the purpose of analysis by experience (less than 5 years, greater than 5 years).

Counts by canning quality, judge experience, and bean breeding lines are listed in the following table (Table 8.33).

Table 8.33 Bean experiment results

Cal	Black		Kidney		Navies		Pinto	
	< 5 Years	> 5 Years	< 5 Years	> 5 Years	< 5 Years	> 5 Years	< 5 Years	> 5 Years
1	13	32	7	10	10	22	13	2
2	91	78	32	31	56	51	29	17
3	123	124	136	96	84	107	91	68
4	72	122	101	104	84	98	109	124
5	24	31	47	71	51	52	60	109
6	2	3	6	18	24	37	25	78
7	0	0	1	0	1	5	1	12

- (a) Fit the generalized logit model to these data. Perform a complete and appropriate analysis of the data, focusing on:
 - (i) An evaluation of the effects of the combination of treatments
 - (ii) Interpretation of the odds ratios
 - (iii) The expected probability per category for each treatment
- (b) Test whether the proportional odds assumption is viable. Cite relevant evidence to support your conclusion regarding the adequacy of the assumption.

Exercise 8.7.5 An experiment was conducted to look at the damage levels (ordinal categories 0–4) of *Picea sitchensis* shoots in two time periods (10 November and 8 December), at four temperatures (different on each date), and at four ozone levels (Table 8.34).

- (a) Fit the cumulative logit proportional odds model to these data. Perform a complete and appropriate analysis of the data, focusing on:
 - (i) An evaluation of the effects of the combination of treatments
 - (ii) Interpretation of the odds ratios
 - (iii) The expected probability per category for each treatment
- (b) Test whether the proportional odds assumption is viable. Cite relevant evidence to support your conclusion regarding the adequacy of the assumption.

Table 8.34 Experimental results of *Picea sitchensis* sprouts

Weather	Temperature (°C)	Ozone	Category	Frequency
1	-9	170	0	1
1	-9	170	1	10
1	-9	170	2	2
1	-9	170	3	2
1	-9	170	4	0
1	-12	170	0	0
1	-12	170	1	8
1	-12	170	2	3
1	-12	170	3	1
1	-12	170	4	3
1	-15	170	0	0
1	-15	170	1	3
1	-15	170	2	2
1	-15	170	3	4
1	-15	170	4	6
1	-18	170	0	0
1	-18	170	1	1
1	-18	170	2	1
1	-18	170	3	4
1	-18	170	4	9
1	-9	120	0	1
1	-9	120	1	9
1	-9	120	2	4
1	-9	120	3	1
1	-9	120	4	0
1	-12	120	0	0
1	-12	120	1	7
1	-12	120	2	7
1	-12	120	3	1
1	-12	120	4	0
1	-15	120	0	0
1	-15	120	1	1
1	-15	120	2	5
1	-15	120	3	6
1	-15	120	4	3
1	-18	120	0	0
1	-18	120	1	0
1	-18	120	2	4
1	-18	120	3	5
1	-18	120	4	6
1	-9	70	0	4
1	-9	70	1	6

(continued)

Table 8.34 (continued)

Weather	Temperature (°C)	Ozone	Category	Frequency
1	-9	70	2	3
1	-9	70	3	2
1	-9	70	4	0
1	-12	70	0	1
1	-12	70	1	6
1	-12	70	2	6
1	-12	70	3	2
1	-12	70	4	0
1	-15	70	0	0
1	-15	70	1	3
1	-15	70	2	6
1	-15	70	3	4
1	-15	70	4	2
1	-18	70	0	0
1	-18	70	1	1
1	-18	70	2	0
1	-18	70	3	5
1	-18	70	4	9
1	-9	0	0	2
1	-9	0	1	11
1	-9	0	2	2
1	-9	0	3	0
1	-9	0	4	0
1	-12	0	0	1
1	-12	0	1	6
1	-12	0	2	6
1	-12	0	3	2
1	-12	0	4	0
1	-15	0	0	2
1	-15	0	1	4
1	-15	0	2	4
1	-15	0	3	3
1	-15	0	4	2
1	-18	0	0	0
1	-18	0	1	4
1	-18	0	2	3
1	-18	0	3	5
1	-18	0	4	3
2	-15	170	0	3
2	-15	170	1	8
2	-15	170	2	4
2	-15	170	3	1

(continued)

Table 8.34 (continued)

Weather	Temperature (°C)	Ozone	Category	Frequency
2	-15	170	4	3
2	-19	170	0	0
2	-19	170	1	10
2	-19	170	2	5
2	-19	170	3	0
2	-19	170	4	4
2	-23	170	0	0
2	-23	170	1	1
2	-23	170	2	8
2	-23	170	3	4
2	-23	170	4	6
2	-27	170	0	0
2	-27	170	1	0
2	-27	170	2	2
2	-27	170	3	3
2	-27	170	4	14
2	-15	120	0	6
2	-15	120	1	6
2	-15	120	2	8
2	-15	120	3	0
2	-15	120	4	0
2	-19	120	0	1
2	-19	120	1	12
2	-19	120	2	7
2	-19	120	3	0
2	-19	120	4	0
2	-23	120	0	0
2	-23	120	1	0
2	-23	120	2	7
2	-23	120	3	7
2	-23	120	4	6
2	-27	120	0	0
2	-27	120	1	0
2	-27	120	2	1
2	-27	120	3	2
2	-27	120	4	17
2	-15	70	0	9
2	-15	70	1	4
2	-15	70	2	5
2	-15	70	3	2
2	-15	70	4	0
2	-19	70	0	2

(continued)

Table 8.34 (continued)

Weather	Temperature (°C)	Ozone	Category	Frequency
2	-19	70	1	10
2	-19	70	2	6
2	-19	70	3	0
2	-19	70	4	2
2	-23	70	0	0
2	-23	70	1	3
2	-23	70	2	5
2	-23	70	3	4
2	-23	70	4	8
2	-27	70	0	0
2	-27	70	1	0
2	-27	70	2	0
2	-27	70	3	3
2	-27	70	4	17
2	-15	0	0	5
2	-15	0	1	6
2	-15	0	2	3
2	-15	0	3	1
2	-15	0	4	2
2	-19	0	0	6
2	-19	0	1	5
2	-19	0	2	3
2	-19	0	3	1
2	-19	0	4	2
2	-23	0	0	0
2	-23	0	1	4
2	-23	0	2	2
2	-23	0	3	1
2	-23	0	4	3
2	-27	0	0	0
2	-27	0	1	1
2	-27	0	2	0
2	-27	0	3	5
2	-27	0	4	11

Appendix

Data: CRD with a multinomial response: ordinal

Rep	Trt	Cat	Freq	Rep	Trt	Cat	Freq
rep1	M1H1	Without	0	rep1	M2H3	Moderate	3
rep2	M1H1	Without	2	rep2	M2H3	Moderate	1
rep3	M1H1	Without	2	rep3	M2H3	Moderate	3
rep4	M1H1	Without	2	rep4	M2H3	Moderate	2
rep1	M1H2	Without	2	rep1	M2H4	Moderate	4
rep2	M1H2	Without	0	rep2	M2H4	Moderate	2
rep3	M1H2	Without	4	rep3	M2H4	Moderate	2
rep4	M1H2	Without	2	rep4	M2H4	Moderate	5
rep1	M1H3	Without	3	rep1	M2H1	Severe	4
rep2	M1H3	Without	7	rep2	M2H1	Severe	6
rep3	M1H3	Without	1	rep3	M2H1	Severe	7
rep4	M1H3	Without	2	rep4	M2H1	Severe	4
rep1	M1H4	Without	0	rep1	M2H2	Severe	5
rep2	M1H4	Without	5	rep2	M2H2	Severe	2
rep3	M1H4	Without	2	rep3	M2H2	Severe	3
rep4	M1H4	Without	1	rep4	M2H2	Severe	4
rep1	M1H1	Moderate	3	rep1	M2H3	Severe	3
rep2	M1H1	Moderate	2	rep2	M2H3	Severe	4
rep3	M1H1	Moderate	3	rep3	M2H3	Severe	4
rep4	M1H1	Moderate	5	rep4	M2H3	Severe	4
rep1	M1H2	Moderate	3	rep1	M2H4	Severe	5
rep2	M1H2	Moderate	3	rep2	M2H4	Severe	6
rep3	M1H2	Moderate	6	rep3	M2H4	Severe	0
rep4	M1H2	Moderate	3	rep4	M2H4	Severe	3
rep1	M1H3	Moderate	4	rep1	M3H1	Without	0
rep2	M1H3	Moderate	2	rep2	M3H1	Without	3
rep3	M1H3	Moderate	1	rep3	M3H1	Without	2
rep4	M1H3	Moderate	3	rep4	M3H1	Without	0
rep1	M1H4	Moderate	5	rep1	M3H2	Without	5
rep2	M1H4	Moderate	4	rep2	M3H2	Without	3
rep3	M1H4	Moderate	8	rep3	M3H2	Without	3
rep4	M1H4	Moderate	4	rep4	M3H2	Without	2
rep1	M1H1	Severe	6	rep1	M3H3	Without	0
rep2	M1H1	Severe	6	rep2	M3H3	Without	2
rep3	M1H1	Severe	5	rep3	M3H3	Without	1
rep4	M1H1	Severe	3	rep4	M3H3	Without	0
rep1	M1H2	Severe	5	rep1	M3H4	Without	3
rep2	M1H2	Severe	7	rep2	M3H4	Without	5
rep3	M1H2	Severe	0	rep3	M3H4	Without	7
rep4	M1H2	Severe	5	rep4	M3H4	Without	3

(continued)

Data: CRD with a multinomial response: ordinal

Rep	Trt	Cat	Freq	Rep	Trt	Cat	Freq
rep1	M1H3	Severe	3	rep1	M3H1	Moderate	0
rep2	M1H3	Severe	1	rep2	M3H1	Moderate	5
rep3	M1H3	Severe	7	rep3	M3H1	Moderate	5
rep4	M1H3	Severe	5	rep4	M3H1	Moderate	0
rep1	M1H4	Severe	5	rep1	M3H2	Moderate	3
rep2	M1H4	Severe	1	rep2	M3H2	Moderate	2
rep3	M1H4	Severe	0	rep3	M3H2	Moderate	6
rep4	M1H4	Severe	5	rep4	M3H2	Moderate	1
rep1	M2H1	Without	1	rep1	M3H3	Moderate	3
rep2	M2H1	Without	2	rep2	M3H3	Moderate	5
rep3	M2H1	Without	1	rep3	M3H3	Moderate	3
rep4	M2H1	Without	1	rep4	M3H3	Moderate	3
rep1	M2H2	Without	1	rep1	M3H4	Moderate	0
rep2	M2H2	Without	3	rep2	M3H4	Moderate	2
rep3	M2H2	Without	1	rep3	M3H4	Moderate	3
rep4	M2H2	Without	4	rep4	M3H4	Moderate	4
rep1	M2H3	Without	4	rep1	M3H1	Severe	9
rep2	M2H3	Without	5	rep2	M3H1	Severe	2
rep3	M2H3	Without	3	rep3	M3H1	Severe	3
rep4	M2H3	Without	4	rep4	M3H1	Severe	10
rep1	M2H4	Without	1	rep1	M3H2	Severe	2
rep2	M2H4	Without	1	rep2	M3H2	Severe	5
rep3	M2H4	Without	8	rep3	M3H2	Severe	1
rep4	M2H4	Without	2	rep4	M3H2	Severe	7
rep1	M2H1	Moderate	4	rep1	M3H3	Severe	6
rep2	M2H1	Moderate	2	rep2	M3H3	Severe	3
rep3	M2H1	Moderate	2	rep3	M3H3	Severe	6
rep4	M2H1	Moderate	5	rep4	M3H3	Severe	7
rep1	M2H2	Moderate	4	rep1	M3H4	Severe	7
rep2	M2H2	Moderate	4	rep2	M3H4	Severe	3
rep3	M2H2	Moderate	6	rep3	M3H4	Severe	0
rep4	M2H2	Moderate	2	rep4	M3H4	Severe	3

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



Chapter 9

Generalized Linear Mixed Models for Repeated Measurements



9.1 Introduction

Repeated measures data, also known as longitudinal data, are those derived from experiments in which observations are made on the same experimental units at various planned times. These experiments can be of the regression or analysis of variance (ANOVA) type, can contain two or more treatments, and are set up using familiar designs, such as completely randomized design (CRD), randomized complete block design (RCBD), or randomized incomplete blocks, if blocking is appropriate, or using row and column designs such as Latin squares when appropriate. Repeated measures designs are widely used in the biological sciences and are fairly well understood for normally distributed data but less so with binary, ordinal, count data, and so on. Nevertheless, recent developments in statistical computing methodology and software have greatly increased the number of tools available for analyzing categorical data.

A generalized linear mixed model (GLMM) is one of the most useful and sophisticated structures in modern statistics, as it allows complex structures to be incorporated into the framework of a general linear model. Fitting such models has been the subject of much research over the last three decades. GLMMs, for repeated measures, combine both generalized linear model (GLM) theory (e.g., a binomial, multinomial, or Poisson response variable) and linear mixed effects models.

Experimentation is sometimes not well understood since researchers believe that it involves only the manipulation of the levels of independent variables and the observation of subsequent responses in dependent variables. Independent variables, whose levels are determined or set by the experimenter, are said to have fixed effects, although random effects are also very common, where the levels of the effects are assumed to be randomly selected from an infinite population of possible levels. Many variables of interest in research are not fully amenable to experimental manipulation but can nevertheless be studied by considering them to have random effects. For example, the genetic composition of individuals of a species cannot be

manipulated experimentally, but it is of great interest to geneticists aiming to assess the genetic contribution to individual variation of some specific behaviors.

A GLMM with repeated measures is a generalization of the standard linear model, and this generalization is due to (1) the presence of more than one response variable that can be binary, ordinal, count, and so on and (2) the nonconstant correlation and/or variability exhibited by the data. The linear mixed model, therefore, gives you the flexibility to model not only the means of your data (as in the standard linear model) but also their variances and covariances. Usually, a normal distribution is assumed for random effects. Since normally distributed data can be modeled entirely in terms of their means and variances/covariances, the two sets of parameters in a linear mixed model actually specify the full probability distribution of the data. The parameters of the mean structure in the model are called (known as) fixed effects parameters, which can be qualitative (as in traditional analysis of variance) or quantitative (as in standard regression), and the parameters of the variance–covariance of the model are known as covariance parameters, which help distinguish a linear mixed model from the standard linear model. Covariance parameters come up quite frequently in the following applications, with two more typical scenarios:

- (a) Experimental units on which data are measured can be grouped into clusters, and data from a common cluster are correlated.
- (b) Repeated measurements of the same experimental unit are taken, and these repeated measurements correlate or show some variability.

The first scenario can be generalized to include a set of clusters nested within one another. For example, if students are the experimental unit, they can be grouped into classes (clusters), which, in turn, can be grouped into schools. Each level of this hierarchy may present an additional source of variability and correlation. The second scenario occurs in longitudinal studies, in which repeated measurements of the same experimental unit over time are taken. Alternatively, these repeated measures could be spatial or multivariate.

9.2 Example of Turf Quality

The proportional odds model, introduced by McCullagh (1980), was proposed as an extension of the generalized linear model used for ordinal responses. One can recall that the proportional odds model is a special case of a GLM with a cumulative link function in which the probability of an observation falling into a category or below is modeled. In the case of a logit link, with only two categories (a binary response), the proportional odds model reduces to a standard logistic regression or a classification model. As with any other type of response variable, repeated measurements are common in agronomic research. They result in clustered data structures with correlations between repeated observations in the same experimental unit that must be taken into account in the analysis.

Table 9.1 Turf quality of five grass varieties (low, Med = medium, Excel = Excellent, Sept = September)

Variety	No. of plots	May			July			Sept		
		Low	Med	Excel	Low	Med	Excel	Low	Med	Excel
1	18	4	10	4	1	9	8	0	12	6
2	17	2	11	4	0	7	10	0	9	8
3	17	2	11	4	2	8	7	2	11	4
4	18	8	7	3	4	8	6	4	13	1
5	18	1	11	6	3	4	11	3	6	9

The data were obtained from an experiment studying the turf quality of five grass varieties. The varieties were sown independently in 17 or 18 plots. The evaluations of the plots (experimental units) were carried out in the months of May, July, and September of the growing season, and turf quality was classified on an ordinal scale into three categories: low quality, medium quality, and excellent quality, as demonstrated in Table 9.1.

The components of the GLMM, with repeated measures with an ordinal multinomial response, are as follows:

Distributions: $y_{1ij}, y_{2ij}, y_{3ij} | \rho_{ij} \sim \text{Multinomial}(N_{ij}, \pi_{1ij}, \pi_{2ij}, \pi_{3ij})$, where y_{1ij}, y_{2ij} , and y_{3ij} are the observed frequencies of the responses (turf quality) in each c category (low, medium, and excellent), and ρ_{ij} is the random effect due to the combination variety \times month (measurement time), assuming $\rho_{ij} \sim N(0, \sigma_\rho^2)$.

Linear predictor: $\eta_{(c)ij} = \eta_c + \tau_i + \rho_{ij}$, where $\eta_{(c)ij}$ is the c th link ($c = 1, 2$) in the ij th combination variety \times month, η_c is the intercept for the c th link, τ_i is the fixed effect due to the i th treatment, and ρ_{ij} is the random effect due to the ij th measurement of variety \times month ($\rho_{ij} \sim N(0, \sigma_{\text{variety} \times \text{month}}^2)$). The link functions for each category are as follows:

$$\log\left(\frac{\pi_{0ij}}{1 - \pi_{0ij}}\right) = \eta_{0ij}$$

$$\log\left(\frac{\pi_{0ij} + \pi_{1ij}}{1 - (\pi_{0ij} + \pi_{1ij})}\right) = \eta_{1ij}$$

The following Statistical Analysis Software (SAS) program fits a repeated measures GLMM with an ordinal response.

```
proc glimmix data=turfgrass method=laplace;
class Variety time;
model cat (order=data)=variety|time/dist=Multinomial link=clogit
solution oddsratio;
random intercept/subject=variety type=cs solution ;
```

Table 9.2 Fit statistics under different correlation structures

Fit statistics	Covariance structure			
	CS	AR(1)	Toep (1)	UN
-2 Log likelihood	497.38	497.46	497.37	n o c o n v e r g e
AIC (smaller is better)	513.38	513.46	511.37	
AICC (smaller is better)	513.94	514.02	511.81	
BIC (smaller is better)	510.26	510.34	508.64	
CAIC (Consistent Akaike's information criterion) (smaller is better)	518.26	518.34	515.64	
HQIC (Hannan Quinn information criterion) (smaller is better)	504.99	505.07	504.03	

```
estimate 'c=1, var=1' intercept 1 0 variedad 1 0 0 0 0,
'c=2, var=1' intercept 0 1 variedad 1 0 0 0 0,
'c=1, var=2' intercept 1 0 variedad 0 1 0 0 0,
'c=2, var=2' intercept 0 1 variedad 0 1 0 0 0,
'c=1, var=3' intercept 1 0 variedad 0 0 1 0 0,
'c=2, var=3' intercept 0 1 variedad 0 0 1 0 0,
'c=1, var=4' intercept 1 0 variedad 0 0 0 1 0,
'c=2, var=4' intercept 0 1 variedad 0 0 0 1 0,
'c=1, var=5' intercept 1 0 variedad 0 0 0 0 1,
'c=2, var=5' intercept 0 1 variedad 0 0 0 0 1/ilink;
freq y;
run;
```

Mixed models have advantages over fixed linear models (Littell et al. 1996) because they have the ability to incorporate fixed ($X\beta$) and random effects (Zb) that allow us to select different variance–covariance structures for repeated measures experiments (with or without missing data) to see which covariance structure best fits the model (Henderson 1984; Smith et al. 2005). Selecting or building a good enough model involves selecting a covariance structure that best fits the dataset. The information criteria minus two Restricted Log Likelihood ($-2RLL$), Akaike information criterion (AIC), Corrected Akaike’s information criterion (AICC), Bayesian information criterion (BIC), etc.) provided by proc GLIMMIX are used as statistical fit measures to select the variance structure (compound symmetry (“CS”), first-order autoregressive (“AR(1)”), Toeplitz (“Toep(1)”), unstructured (“UN”)) that best models the dataset.

Most of the commands have already been explained. To provide the correlation structure that you want to model, with the above program, you vary the “TYPE” option = (CS, AR(1), Toep(1), and UN) separately to specify each of the covariance structures in the parentheses. Part of the results is shown below.

According to the fit statistics (Table 9.2), the covariance structure that best fits the dataset is Toeplitz of order 1 (Toep(1)). The type III tests of fixed effects, shown in Table 9.3 part (a), indicate that grass variety provides different turfgrass qualities

Table 9.3 Results of the analysis of variance

(a) Type III tests of fixed effects							
Effect	Num degree of freedom (DF)		Den DF	F-value	Pr > F		
Variety	4		10	4.80	0.0202		
(b) Solutions for fixed effects							
Effect	Cat	Variety	Estimate	Standard error	DF	t-value	Pr > t
Intercept	Low		-2.4509	0.3219	10	-7.61	<0.0001
Intercept	Medium		0.1961	0.2721	10	0.72	0.4875
Variety		Var1	0.4261	0.3753	10	1.14	0.2827
Variety		Var2	-0.01502	0.3785	10	-0.04	0.9691
Variety		Var3	0.6125	0.3825	10	1.60	0.1404
Variety		Var4	1.4904	0.3943	10	3.78	0.0036
Variety		Var5	0

Table 9.4 Estimated linear predictors and means on the model scale (Estimate) and on the data scale (Mean) for observed turfgrass quality in grass varieties in the multinomial generalized logit model

Estimates							
Label	Estimate	Standard error	DF	t-value	Pr > t	Mean	Standard error mean
c = 1, var = 1	-2.0248	0.3018	10	-6.71	<0.0001	0.1166	0.03110
c = 2, var = 1	0.6222	0.2646	10	2.35	0.0405	0.6507	0.06013
c = 1, var = 2	-2.4659	0.3177	10	-7.76	<0.0001	0.07828	0.02292
c = 2, var = 2	0.1811	0.2667	10	0.68	0.5125	0.5452	0.06613
c = 1, var = 3	-1.8384	0.3040	10	-6.05	0.0001	0.1372	0.03599
c = 2, var = 3	0.8086	0.2760	10	2.93	0.0150	0.6918	0.05884
c = 1, var = 4	-0.9605	0.2791	10	-3.44	0.0063	0.2768	0.05588
c = 2, var = 4	1.6865	0.2992	10	5.64	0.0002	0.8438	0.03944
c = 1, var = 5	-2.4509	0.3219	10	-7.61	<0.0001	0.07937	0.02352
c = 2, var = 5	0.1961	0.2721	10	0.72	0.4875	0.5489	0.06737

($P = 0.0202$). The “solution” option in the model specification “Model” provides the solution of fixed effects of the model (intercepts and treatments), which we use to estimate the linear predictors $\hat{\eta}_{ci} = \hat{\eta}_c + \text{Variety}_i$ (part (b)).

The probabilities $\hat{\pi}_{ci}$ obtained using the “Estimate” information are tabulated under the “Mean” column of Table 9.4.

From these values, we can observe that for the category " $c = 1, \text{var} = 1$," the value of the linear predictor is $\hat{\eta}_{11} = \hat{\eta}_1 + \widehat{\text{variety}}_1 = -2.0248$. Taking the inverse of $\hat{\eta}_{11}$ corresponds to the probability of $\hat{\pi}_{11} = 0.1166$ of observing "Low"-quality grass of variety 1. Now, for the category " $c = 2, \text{var} = 1$," the inverse of the linear predictor is 0.6507, which is the estimate of the probability $\hat{\pi}_{11} + \hat{\pi}_{21}$. From this value, we can obtain the probability that variety 1 provides grass of "Medium" quality, that is, $\hat{\pi}_{11} + \hat{\pi}_{21} = 0.6504$, and, substituting the value of $\hat{\pi}_{11}$, we obtain the probability value $\hat{\pi}_{21} = 0.6507 - 0.1166 = 0.5341$. With these two probability estimates $\hat{\pi}_{11}$ and $\hat{\pi}_{21}$, it is possible to estimate the probability that variety 1 will yield an "Excellent" quality turf, which is equal to $\hat{\pi}_{31} = 1 - 0.6504 = 0.3496$. Likewise, we obtain the values of the remaining probabilities $\hat{\pi}_{ci}$ for the rest of the grass varieties.

9.3 Effect of Insecticides on Aphid Growth

A cage experiment was used to investigate the effect of three insecticides on aphid colonies with partial resistance to a common active compound. There were eight treatments: all combinations of the three insecticides and a control (no insecticide) with two types of colonies (susceptible or partially resistant). The experiment was organized as an RCBD with six blocks of eight cages, and each cage was assigned a treatment combination in each block. A colony of aphids was reared in each cage, and the number of live aphids was recorded before insecticide treatment was applied and then 2 and 6 days after application. Both hatches and deaths could occur within each cage between evaluations. The dataset from this experiment is shown below (Table 9.5).

Following the same reasoning as in previous examples, the components of the GLMM with a Poisson response and repeated measures, which models the number of aphids (y_{ijkl}), is described in the following lines.

$$\text{Distributions: } y_{ijkl} \mid b_l, \text{insecticide} \times \text{clone}(\text{block})_{ij(l)} \sim \text{Poisson}(\lambda_{ijkl})$$

$$b_l \sim N(0, \sigma_{\text{block}}^2), \text{insecticide} \times \text{clone}(\text{block})_{ij(l)} \sim N(0, \sigma_{\text{insecticide} \times \text{clone} \times \text{block}}^2).$$

Linear predictor: $\eta_{ijkl} = \theta + I_i + C_j + (IC)_{ij} + b_l + IC(b)_{ij(l)} + \tau_l + (I\tau)_{il} + (C\tau)_{il} + (IC\tau)_{ijkl}$, $C_j + (IC)_{ij} + b_l + IC(b)_{ij(l)} + \tau_l + (I\tau)_{il} + (C\tau)_{il} + (IC\tau)_{ijkl}$, where η_{ijkl} is the linear predictor, θ is the intercept, I_i ($i = 1, 2, 3$) is the fixed effect due to the insecticide, C_j ($j = 1, 2$) is the fixed effect due to the aphid clone, $(IC)_{ij}$ is the fixed effect due to the interaction between the type of insecticide and clone, b_l ($k = 1, 2, 3$) is the random block effect, assuming $b_l \sim N(0, \sigma_{\text{block}}^2)$, $IC(b)_{ij(l)}$ is the random effect of the interaction between the insecticide and clone within blocks, assuming $\text{insecticide} \times \text{clone}(\text{block})_{ij(k)} \sim N(0, \sigma_{\text{insecticide} \times \text{clone} \times \text{block}}^2)$, τ_l ($l = 1, 2, 3$) is the fixed effect due to measurement time, and $(C\tau)_{il}$ and $(IC\tau)_{ijkl}$ are the fixed effects due to interaction.

Table 9.5 Effect of insecticides (C = control, R = resistant, S = susceptible) on aphid growth

Block	Cage	Insecticide	Clone	Dia1	Dia2	Dia6
1	1	Control	R	60	111	220
1	2	Control	S	127	131	220
1	3	D	R	64	30	27
1	4	D	S	110	27	35
1	5	H	R	118	75	121
1	6	H	S	71	10	111
1	7	P	R	66	69	62
1	8	P	S	40	25	19
2	1	Control	R	54	152	156
2	2	Control	S	58	130	362
2	3	D	R	76	60	110
2	4	D	S	48	22	110
2	5	H	R	130	113	101
2	6	H	S	76	76	85
2	7	P	R	93	77	185
2	8	P	S	49	0	8
3	1	Control	R	94	175	292
3	2	Control	S	26	33	52
3	3	D	R	121	73	60
3	4	D	S	78	23	1
3	5	H	R	73	74	56
3	6	H	S	54	27	49
3	7	P	R	25	10	32
3	8	P	S	36	22	1
4	1	Control	R	75	134	238
4	2	Control	S	86	57	194
4	3	D	R	69	32	12
4	4	D	S	122	66	20
4	5	H	R	185	88	251
4	6	H	S	47	23	116

Link function: $\log(\lambda_{ijkl}) = \eta_{ijk}$ is the link function that relates the linear predictor to the mean (λ_{ijkl}).

The following SAS program adjusts the GLMM with a Poisson distribution on repeated measures.

```
proc glimmix nobound method=laplace;
class ID Block Insecticide Cage Clone time;
model y = Insecticide|clone|time/dist=poi;
random intercept Insecticide*Clon/subject=block;
lsmeans Insecticide|Clon|time/lines ilink;
run;quit;
```

Table 9.6 Results of the analysis of variance in the Poisson GLMM

(a) Fit statistics				
Fit statistics	CS	AR(1)	Toep(1)	UN
-2 Log likelihood	1125.54	1113.19	1127.25	No converge
AIC (smaller is better)	1177.54	1165.19	1177.25	
AICC (smaller is better)	1202.17	1189.82	1199.67	
BIC (smaller is better)	1161.58	1149.24	1161.91	
CAIC (smaller is better)	1187.58	1175.24	1186.91	
HQIC (smaller is better)	1142.52	1130.18	1143.59	
(b) Fit statistics for conditional distribution				
-2 Log L (y r. effects)				1006.38
Pearson's chi-square				484.48
Pearson's chi-square/DF				5.77

Before fitting the GLMM, we compare the estimates of covariance structures with a Poisson distribution assumed in the response variable. According to the fit statistics, the covariance structure that best models the data is the autoregressive type of order 1 (AR(1)). The value of the fit statistic of the conditional distribution Pearson's chi - square/DF = 5.77 indicates that there is an extra variation (aka overdispersion) and that the Poisson distribution does not adequately fit the data (Table 9.6).

Since there is overdispersion in the data, a highly recommended alternative is to find another suitable (or more appropriate) distribution for this dataset. In this case, the linear predictor will be the same, although now, a negative binomial distribution will be assumed in the response variable. That is,

$$y_{ijkl}|b_l, \text{insecticide} \times \text{clone}(\text{block})_{ij(l)} \sim \text{Negative Binomial}(\lambda_{ijkl}, \phi)$$

This negative binomial model arises by assuming that the conditional distribution of observations given random blocks and Insecticide*clone(block)_{ij(l)} is as follows: $y_{ijkl}|b_l, \text{insecticide} \times \text{clone}(\text{block})_{ij(l)} \sim \text{Poisson}(\lambda_{ijkl})$, where $\lambda_{ijkl} \sim \text{Gamma}(\frac{1}{\phi}, \phi)$. The result of the new distribution of $y_{ijkl}|b_l, \text{insecticide} \times \text{clone}(\text{Block})_{ij(l)}$ is a negative binomial (Negative binomial (λ_{ijkl}, ϕ)). The link function is $\log(\lambda_{ijkl}) = \eta_{ijkl}$.

The following SAS code fits the GLMM with a negative binomial distribution.

```
proc glimmix nobound method=laplace;
class ID Block Insecticide Cage Clone time;
model y = Insecticide | clone | time / dist=negbi;
random intercept Insecticide*Clone / subject=block;
lsmeans Insecticide | Clone | time / lines ilink;
run;
```

Table 9.7 Fit statistics

(a) Fit statistics	
-2 Log likelihood	878.02
AIC (smaller is better)	932.02
AICC (smaller is better)	956.41
BIC (smaller is better)	915.45
CAIC (smaller is better)	942.45
HQIC (smaller is better)	895.66
(b) Fit statistics for conditional distribution	
-2 Log L (y r. effects)	841.84
Pearson's chi-square	72.47
Pearson's chi-square/DF	0.81

Table 9.8 Estimated variance components and tests of fixed effects

(a) Covariance parameter estimates				
Cov Parm	Subject	Estimate	Standard error	
Variance	Block	0.06138	0.03429	
AR(1)	Block	-0.7143	0.1710	
Scale		0.1654	0.03575	
(b) Type III tests of fixed effects				
Effect	Num DF	Den DF	F-value	Pr > F
Insecticide	3	19	23.04	<0.0001
Clone	1	19	8.60	0.0086
Insecticide*clone	3	19	2.25	0.1161
Time	2	44	6.08	0.0047
Insecticide*time	6	44	7.93	<0.0001
Clone*time	2	44	3.90	0.0275
Insecticide*clone*time	6	44	2.15	0.0663

Part of the results is shown in Table 9.7. The values of the fit statistics, assuming a negative binomial distribution of the data, are shown in part (a), and the value of the conditional statistic is observed in part (b) (Pearson's chi - square/DF = 0.81). This indicates that overdispersion has been eliminated from the data, and, so, the negative binomial distribution adequately models the response variable.

The estimated variance components are shown in part (a) of Table 9.8, under an AR(1) covariance structure. The estimates of the variance components of blocks, the interaction between the insecticide and clone within blocks, and the scale parameter are $\hat{\sigma}_{\text{block}}^2 = 0.06613$, $\hat{\sigma}_{\text{insecticide} \times \text{clone}(\text{block})}^2 = -0.7575$, and $\hat{\phi} = 0.1584$, respectively. The fixed III type effects tests (part (b)) indicate that there is a significant effect of insecticide type ($P < 0.0001$), clone ($P = 0.0387$), measurement time ($P = 0.0137$), and interactions insecticide x measurement time ($P < 0.0001$) and clone x measurement time ($P = 0.0259$) on the average number of aphids. The interaction insecticide x clone x measurement time is close to significance ($P < 0.0663$).

Table 9.9 Estimates of insecticide least squares (LS) means on the model scale (Estimate) and the data scale (Mean)

Insecticide	Estimate	Standard error	DF	<i>t</i> -value	Pr > <i>t</i>	Mean	Standard error mean
C	4.7344	0.1478	19	32.03	<0.0001	113.79	16.8211
D	3.9647	0.1547	19	25.62	<0.0001	52.7043	8.1553
H	4.3733	0.1561	19	28.02	<0.0001	79.3010	12.3753
P	3.4892	0.1753	19	19.90	<0.0001	32.7588	5.7432

Table 9.10 Clone least squares means on the model scale (Estimate) and the data scale (Mean)

Clone	Estimate	Standard error	DF	<i>t</i> -value	Pr > <i>t</i>	Mean	Standard error mean
R	4.4332	0.1320	19	33.58	<0.0001	84.1990	11.1158
S	3.8476	0.1890	19	20.36	<0.0001	46.8785	8.8586

Table 9.11 Insecticide*clone least squares means on the model scale (Estimate) and the data scale (Mean)

Insecticide	Clone	Estimate	Standard error	DF	<i>t</i> -value	Pr > <i>t</i>	Mean	Standard error mean
C	R	4.8836	0.1529	19	31.93	<0.0001	132.10	20.2032
C	S	4.5852	0.2479	19	18.49	<0.0001	98.0186	24.3008
D	R	4.0521	0.1886	19	21.49	<0.0001	57.5153	10.8459
D	S	3.8773	0.2337	19	16.59	<0.0001	48.2958	11.2858
H	R	4.7106	0.1997	19	23.59	<0.0001	111.11	22.1870
H	S	4.0359	0.2244	19	17.98	<0.0001	56.5964	12.7026
P	R	4.0866	0.2322	19	17.60	<0.0001	59.5346	13.8263
P	S	2.8918	0.2534	19	11.41	<0.0001	18.0255	4.5675

Table 9.12 Time least squares means on the model scale (Estimate) and the data scale (Mean)

Time	Estimate	Standard error	DF	<i>t</i> -value	Pr > <i>t</i>	Mean	Standard error mean
1	4.2730	0.1434	44	29.79	<0.0001	71.7375	10.2905
2	3.9108	0.1454	44	26.90	<0.0001	49.9372	7.2603
3	4.2373	0.1457	44	29.09	<0.0001	69.2231	10.0830

The linear predictors and estimated means of the factors and interaction are under the “Estimate” and “Mean” columns, respectively. Average number of aphids for insecticide, clone and time are given below:

For insecticide type (Table 9.9):

For clone (Table 9.10):

For the interaction insecticide*clone (Table 9.11):

For measurement time (Table 9.12):

For the interaction insecticide*time (Table 9.13):

For the interaction clone*time (Table 9.14):

For the interaction insecticide*clone*time (Table 9.15):

Table 9.13 Insecticide*time least squares means on the model scale (Estimate) and the data scale (Mean)

	Time	Estimate	Standard error	DF	t-value	Pr > t	Mean	Standard error mean
C	1	4.2381	0.1930	44	21.95	<0.0001	69.2781	13.3733
C	2	4.6631	0.1913	44	24.38	<0.0001	105.96	20.2696
C	3	5.3019	0.1898	44	27.94	<0.0001	200.71	38.0898
D	1	4.4854	0.1965	44	22.83	<0.0001	88.7111	17.4275
D	2	3.7061	0.2014	44	18.40	<0.0001	40.6940	8.1959
D	3	3.7026	0.2035	44	18.20	<0.0001	40.5537	8.2517
H	1	4.4718	0.1978	44	22.60	<0.0001	87.5164	17.3133
H	2	3.9790	0.2016	44	19.73	<0.0001	53.4625	10.7804
H	3	4.6689	0.1977	44	23.62	<0.0001	106.59	21.0694
P	1	3.8967	0.2241	44	17.39	<0.0001	49.2403	11.0358
P	2	3.2949	0.2357	44	13.98	<0.0001	26.9755	6.3583
P	3	3.2759	0.2399	44	13.65	<0.0001	26.4664	6.3502

Table 9.14 Clone*time least squares means on the model scale (Estimate) and the data scale (Mean)

Clone	Time	Estimate	Standard error	DF	t-value	Pr > t	Mean	Standard error mean
R	1	4.3839	0.1595	44	27.49	<0.0001	80.1482	12.7828
R	2	4.2826	0.1605	44	26.68	<0.0001	72.4270	11.6256
R	3	4.6331	0.1601	44	28.94	<0.0001	102.83	16.4644
S	1	4.1621	0.2092	44	19.90	<0.0001	64.2093	13.4323
S	2	3.5390	0.2131	44	16.60	<0.0001	34.4308	7.3387
S	3	3.8416	0.2144	44	17.91	<0.0001	46.5989	9.9931

9.4 Manufacture of Livestock Feed

In this experiment, two types of pelleted feed were manufactured using different amounts of whole sorghum. Using the whole grain resulted in one feed with a high pellet durability index (PDI) and one with a low PDI. The researcher was interested in how much impact this difference in PDI would have on the amount of intact and pelleted feed distributed to the different positions along the feeding line. The line was fed four times with the high PDI feed and four times with the low PDI feed. After each run, the total weight of the feed in each of the 12 identified trays was measured. The feed was then sieved into each tray, and the crushed fine granules were weighed in the feed line. The response of interest was the ratio (proportion) between the weight of fine granules and the total weight of the feed for each tray. The data for this experiment are in the Appendix (Data: Feeding line experiment).

The experimental design used in this study was a split plot in a randomized completely design. There were 2 fixed factors, feed with 2 levels (high PDI feed (H) and low PDI feed (L)), and a tray with 12 levels (1, 2, 3, ..., 12 locations along

Table 9.15 Insecticide*clone*time least squares means on the model scale (Estimate) and the data scale (Mean)

Insecticide	Clone	Time	Estimate	Standard error	DF	t-value	Pr > t	Mean	Standard error mean
C	R	1	4.2592	0.2321	44	18.35	<0.0001	70.7546	16.4256
C	R	2	4.9631	0.2280	44	21.76	<0.0001	143.04	32.6184
C	R	3	5.4284	0.2269	44	23.92	<0.0001	227.78	51.6892
C	S	1	4.2170	0.3044	44	13.86	<0.0001	67.8323	20.6460
C	S	2	4.3630	0.3031	44	14.40	<0.0001	78.4950	23.7891
C	S	3	5.1754	0.2999	44	17.26	<0.0001	176.87	53.0366
D	R	1	4.4161	0.2538	44	17.40	<0.0001	82.7767	21.0081
D	R	2	3.8625	0.2587	44	14.93	<0.0001	47.5825	12.3088
D	R	3	3.8775	0.2623	44	14.78	<0.0001	48.3054	12.6694
D	S	1	4.5546	0.2908	44	15.66	<0.0001	95.0710	27.6437
D	S	2	3.5497	0.2994	44	11.86	<0.0001	34.8028	10.4203
D	S	3	3.5277	0.3026	44	11.66	<0.0001	34.0460	10.3007
H	R	1	4.8146	0.2622	44	18.36	<0.0001	123.30	32.3313
H	R	2	4.4776	0.2639	44	16.97	<0.0001	88.0246	23.2294
H	R	3	4.8395	0.2644	44	18.30	<0.0001	126.40	33.4231
H	S	1	4.1291	0.2839	44	14.54	<0.0001	62.1193	17.6362
H	S	2	3.4803	0.2935	44	11.86	<0.0001	32.4709	9.5292
H	S	3	4.4984	0.2823	44	15.93	<0.0001	89.8763	25.3740
P	R	1	4.0456	0.3077	44	13.15	<0.0001	57.1426	17.5807
P	R	2	3.8271	0.3117	44	12.28	<0.0001	45.9304	14.3176
P	R	3	4.3870	0.3060	44	14.34	<0.0001	80.3983	24.5990
P	S	1	3.7479	0.3189	44	11.75	<0.0001	42.4308	13.5319
P	S	2	2.7627	0.3453	44	8.00	<0.0001	15.8430	5.4700
P	S	3	2.1648	0.3634	44	5.96	<0.0001	8.7125	3.1659

Table 9.16 Results of the analysis of variance of the experiment

Sources of variation	Degrees of freedom
Feeding	$(2 - 1) = 1$
Feeding (running)	$2(4 - 1) = 6$
Tray	$(12 - 1) = 11$
Feeding*tray	$(2 - 1)(12 - 1) = 11$
Feeding (tray*tray)	$2(12 - 1)(4 - 1) = 66$
Total	$a \times b \times r - 1 = 95$

the feed line). Different run levels (1, 2, 3, 4 runs in the feed line) may influence the inference of this experiment, so it is advisable to analyze which variance structure is suitable for this analysis.

The ANOVA table (Table 9.16) with degrees of freedom for this experiment is shown below.

The researcher aims to draw conclusions about the destructiveness in the feed line with two types of feed, high PDI and low PDI. The following GLMM is used to describe the experiment:

$$y_{ijk} = \mu + \alpha_i + \alpha(r)_{ik} + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}$$

where y_{ijk} is the proportion observed in the run k ($k = 1, 2, 3, 4$), tray j ($j = 1, 2, \dots, 12$), and in feed i ($i = 1, 2$); μ is the overall mean; α_i is the fixed effect of feed i ; $\alpha(r)_{ik}$ is the random effect of the i th feed within the k th run, assuming $\alpha(r)_{ik} \sim N(0, \sigma_{ar}^2)$; β_j is the fixed effect due to the j th tray; $(\alpha\beta)_{ij}$ is the effect of the interaction between the i th feed and the j th tray; and ε_{ijk} is the experimental error. The components of the conditional GLMM assuming that the response variable follows a beta distribution are listed below:

The distribution of the response variable is given by $y_{ijk} | \alpha(r)_{ik} \sim \text{Beta}(\mu + \alpha_i + \alpha(r)_{ik} + \beta_j + (\alpha\beta)_{ij}, \phi)$ whose linear predictor is $\eta_{ijk} = \mu + \alpha_i + \alpha(r)_{ik} + \beta_j + (\alpha\beta)_{ij}$ with link function $\text{logit}\left(\frac{\pi_{ijk}}{1 - \pi_{ijk}}\right) = \eta_{ijk}$. The following GLIMMIX syntax fits a GLMM with a beta distribution.

```
proc glimmix method=laplace;
class tray feed run;
model ratio = feed|tray/dist=beta;
random intercept/subject=feed(run) type=toep(1);
lsmeans feed|tray/lines ilink;
run;
```

Part of the output is shown below. Four covariance structures (“CS,” “AR(1),” “Toep(1),” and “UN”) were tested to see which one best fits the response variable. Of these covariance structures, “Toep(1)” produced the best fit statistics (part (a), Table 9.17).

Another important result that gives the guideline to continue with the analysis is the conditional distribution statistic (Pearson’s chi – square/DF = 0.96), whose

Table 9.17 Results of the analysis of variance

(a) Fit statistics				
-2 Log likelihood				-429.84
AIC (smaller is better)				-377.84
AICC (smaller is better)				-357.19
BIC (smaller is better)				-375.78
CAIC (smaller is better)				-349.78
HQIC (smaller is better)				-391.77
(b) Fit statistics for conditional distribution				
-2 Log L (ratio <i>l r</i> . effects)				-453.24
Pearson's chi-square				91.40
Pearson's chi-square/DF				0.96
(c) Type III tests of fixed effects				
Effect	Num DF	Den DF	<i>F</i> -value	Pr > <i>F</i>
Feed	1	6	1071.19	<0.0001
Tray	11	65	18.03	<0.0001
Tray*feed	11	65	1.83	0.0660

Table 9.18 Feed least squares means on the model scale (Estimate) and the data scale (Mean)

Feed	Estimate	Standard error	DF	<i>t</i> -value	Pr > <i>t</i>	Mean	Standard error mean
H	-2.0009	0.07409	6	-27.01	<0.0001	0.1191	0.007773
L	1.3832	0.07208	6	19.19	<0.0001	0.7995	0.01155

value indicates that the beta model adequately fits the data, whereas the fixed effects tests (part (c)) indicate that there is a statistically significant effect of feeding type ($P = 0.0001$) and tray ($P = 0.0001$).

The linear predictors and estimated probabilities of the factors and interaction are listed under the "Estimate" and "Mean" columns of the following tables, respectively.

For the feeding line (Table 9.18):

For the tray (Table 9.19):

For the interaction feeding*tray (Table 9.20):

9.5 Characterization of Spatial and Temporal Variations in Fecal Coliform Density

During a 1-month period (June 1981), 30 river water samples were collected from the channel at 3 stations, A, B, and C (downstream to upstream) on 5 randomly selected days at 9:00 a.m. and 3:00 p.m. (1 sample per station per hour per day). Each sample was analyzed for fecal coliform by method FC-96. The data from this experiment are shown in Table 9.21.

Table 9.19 Tray least squares means on the model scale (Estimate) and the data scale (Mean)

Tray	Estimate	Standard error	DF	t-value	Pr > t	Mean	Standard error mean
01	-0.5652	0.08182	65	-6.91	<0.0001	0.3623	0.01891
02	-0.6607	0.08531	65	-7.74	<0.0001	0.3406	0.01916
03	-0.6950	0.08822	65	-7.88	<0.0001	0.3329	0.01959
04	-0.2958	0.08100	65	-3.65	0.0005	0.4266	0.01981
05	-0.3773	0.08212	65	-4.59	<0.0001	0.4068	0.01982
06	-0.2947	0.08057	65	-3.66	0.0005	0.4268	0.01971
07	-0.3520	0.08165	65	-4.31	<0.0001	0.4129	0.01979
08	-0.2992	0.07939	65	-3.77	0.0004	0.4258	0.01941
09	-0.1314	0.07670	65	-1.71	0.0916	0.4672	0.01909
10	-0.3935	0.08096	65	-4.86	<0.0001	0.4029	0.01948
11	0.1571	0.07860	65	2.00	0.0499	0.5392	0.01953
12	0.2014	0.07949	65	2.53	0.0137	0.5502	0.01967

Table 9.20 Tray*feed least squares means on the model scale (Estimate) and the data scale (Mean)

Tray	Feed	Estimate	Standard error	DF	t-value	Pr > t	Mean	Standard error mean
01	H	-2.2408	0.1284	65	-17.46	<0.0001	0.09614	0.01116
01	L	1.1104	0.1015	65	10.94	<0.0001	0.7522	0.01892
02	H	-2.4581	0.1369	65	-17.95	<0.0001	0.07885	0.009946
02	L	1.1367	0.1018	65	11.17	<0.0001	0.7571	0.01872
03	H	-2.4724	0.1375	65	-17.98	<0.0001	0.07782	0.009869
03	L	1.0823	0.1105	65	9.79	<0.0001	0.7469	0.02089
04	H	-2.0307	0.1217	65	-16.69	<0.0001	0.1160	0.01248
04	L	1.4391	0.1070	65	13.45	<0.0001	0.8083	0.01658
05	H	-2.1481	0.1254	65	-17.13	<0.0001	0.1045	0.01174
05	L	1.3935	0.1061	65	13.13	<0.0001	0.8011	0.01690
06	H	-2.0087	0.1208	65	-16.62	<0.0001	0.1183	0.01260
06	L	1.4192	0.1066	65	13.31	<0.0001	0.8052	0.01673
07	H	-2.1026	0.1242	65	-16.93	<0.0001	0.1088	0.01204
07	L	1.3987	0.1061	65	13.18	<0.0001	0.8020	0.01685
08	H	-1.9310	0.1192	65	-16.20	<0.0001	0.1266	0.01318
08	L	1.3325	0.1050	65	12.69	<0.0001	0.7913	0.01734
09	H	-1.6240	0.1113	65	-14.59	<0.0001	0.1647	0.01531
09	L	1.3613	0.1056	65	12.89	<0.0001	0.7960	0.01715
10	H	-2.0863	0.1238	65	-16.85	<0.0001	0.1104	0.01216
10	L	1.2994	0.1044	65	12.45	<0.0001	0.7857	0.01757
11	H	-1.4559	0.1075	65	-13.55	<0.0001	0.1891	0.01648
11	L	1.7701	0.1148	65	15.42	<0.0001	0.8545	0.01427
12	H	-1.4519	0.1076	65	-13.50	<0.0001	0.1897	0.01653
12	L	1.8548	0.1171	65	15.83	<0.0001	0.8647	0.01371

Table 9.21 Variation in fecal coliform densities of the river water samples from three sampling stations on five sampling days at 9:00 a.m. (TM = 1) and 3:00 p.m. (TM = 2)

Sampling date	TM	Site	No. of coliforms per milliliter
18 May	9:00 a.m.	A	648
18 May	3:00 p.m.	A	798
18 May	9:00 a.m.	B	517
18 May	3:00 p.m.	B	702
18 May	9:00 a.m.	C	532
18 May	3:00 p.m.	C	55
26 May	9:00 a.m.	A	1421
26 May	3:00 p.m.	A	1388
26 May	9:00 a.m.	B	1883
26 May	3:00 p.m.	B	1855
26 May	9:00 a.m.	C	1724
26 May	3:00 p.m.	C	1769
29 May	9:00 a.m.	A	1523
29 May	3:00 p.m.	A	759
29 May	9:00 a.m.	B	1361
29 May	3:00 p.m.	B	603
29 May	9:00 a.m.	C	2004
29 May	3:00 p.m.	C	541
1 June	9:00 a.m.	A	1987
1 June	3:00 p.m.	A	1056
1 June	9:00 a.m.	B	1796
1 June	3:00 p.m.	B	1579
1 June	9:00 a.m.	C	1221
1 June	3:00 p.m.	C	1223
5 June	9:00 a.m.	A	870
5 June	3:00 p.m.	A	1099
5 June	9:00 a.m.	B	920
5 June	3:00 p.m.	B	951
5 June	9:00 a.m.	C	926
5 June	3:00 p.m.	C	887

To assess the relative magnitudes of sources of variation due to time, site, and subsampling on the number of coliforms per milliliter (y_{ijk}), an analysis of variance using a GLMM with a Poisson response was performed, as described below:

We denote y_{ijk} as the number of colonies per milliliter, whose conditional distribution is given by $y_{ijk} | \text{sampling}(\text{site})_{ik} \sim \text{Poisson}(\lambda_{ijk})$ with the linear predictor η_{ijk} defined by

$$\eta_{ijk} = \theta + \text{site}_i + \text{sampling}(\text{site})_{ik} + \text{time}_j + (\text{site} \times \text{time})_{ij}$$

Table 9.22 Results of the analysis of variance

(a) Estructuras de covarianza				
Fit statistics	Toep(1)	CS	AR(1)	UN
-2 Log likelihood	2022.64	2022.64	2022.64	2022.64
AIC (smaller is better)	2054.64	2056.64	2056.64	2054.64
AICC (smaller is better)	2096.49	2107.64	2107.64	2096.49
BIC (smaller is better)	2051.31	2053.10	2053.10	2051.31
CAIC (smaller is better)	2067.31	2070.10	2070.10	2067.31
HQIC (smaller is better)	2041.31	2042.47	2042.47	2041.31
(b) Fit statistics for conditional distribution				
-2 Log L (ufc r. effects)	1989.36			
Pearson's chi-square	1632.28			
Pearson's chi-square/DF	54.41			
(c) Type III tests of fixed effects				
Effect	Num DF	Den DF	F-value	Pr > F
Site	2	3	1.41	0.3700
T	4	12	956.04	<0.0001
T*site	8	12	82.44	<0.0001

$$(i = 1, 2, 3; j = 1, 2, 3, 4, 5; k = 1, 2)$$

where η_{ijk} is the linear predictor that relates the linear function to the mean, θ is the intercept, $site_i$ is the fixed effect due to the sampling site i , $sampling(site)_{ik}$ is the random effect due to the sampling time nested within the site, assuming $sampling(site)_{ik} \sim N(0, \sigma^2_{sampling(site)})$, $time_j$ is the fixed effect due to sampling date, and $(site \times time)_{ij}$ is the effect of the interaction between the site and sampling date. The link function for this model is $\log(\lambda_{ijk}) = \eta_{ijk}$.

The following GLIMMIX syntax fits a GLMM with a Poisson response.

```
proc glimmix data=ufc nobound method=laplace;
class T TM Site ;
model ufc = Site|T/dist=Poisson link=log;
random intercept/subject=TM(Site) type=toep(1);
lsmeans Site|T/lines ilink;
run;
```

Part of the results is summarized in Table 9.22. To determine which covariance structure best models the response variable, four types were tested (part (a)), all of which produced very similar results. Because of these results, the “Toep(1)” covariance structure was chosen. From this, the fit statistics were obtained, and the value of the conditional distribution statistic is Pearson’s chi – square/DF = 54.41. This value indicates that there is a strong overdispersion in the dataset. Therefore, it is important to look for an alternative distribution that solves this problem.

The hypothesis tests in part (c) indicate that there is a significant difference in the date of sampling ($P = 0.0001$) as well as in the interaction between the site and date

Table 9.23 Fit statistics under the negative binomial distribution

(a) Fit statistics	
-2 Log likelihood	436.98
AIC (smaller is better)	470.98
AICC (smaller is better)	521.98
BIC (smaller is better)	467.44
CAIC (smaller is better)	484.44
HQIC (smaller is better)	456.81
(b) Fit statistics for conditional distribution	
-2 Log L(ufc r. effects)	432.83
Pearson's chi-square	22.66
Pearson's chi-square/DF	0.76

of sampling ($P = 0.0001$). That is, the concentration of fecal coliform units per milliliter is affected by the date of data collection. However, we observed that there is an excessive dispersion in the data. One way to check for and deal with overdispersion is to run a quasi-Poisson model, which, during the fitting process, adds an additional dispersion parameter to account for that additional variance. Another option is to look for a distribution that adequately fits the data; in this case, the negative binomial distribution is a good alternative.

Next, we will implement the analysis assuming that the response variable is distributed under a negative binomial distribution. This means that the distribution of y_{ijk} (number of colonies per mililitro) is given by $y_{ijk} | \text{smampling}(\text{site})_{ik} \sim \text{Negative Binomial}(\lambda_{ijk}, \phi)$, where ϕ is the scale parameter. However, the linear predictor η_{ijk} and the link function remain unchanged.

The following GLIMMIX commands fit a GLMM with a negative binomial distribution.

```
proc glimmix data=ufc nobound method=laplace;
class T TM Site;
model ufc = Site|T/dist=negbin;
random intercept/subject=TM(Site)/type=Toep(1);
lsmeans Site|T/lines ilink;
run;
```

Part of the output of the above program is shown below. The values of the fit statistics under the negative binomial distribution (part (a) of Table 9.23) are much smaller compared to those obtained assuming the Poisson model, indicating that the negative binomial distribution adequately fits the response variable. Furthermore, the value of the conditional distribution statistic indicates that the negative binomial distribution is a good distribution for these data (Pearson's chi - square/DF = 0.76).

This parameter (Pearson's chi - $\frac{\text{square}}{\text{DF}} = 0.76$) refers to how many times the variance is larger than the mean. Since this value is less than 1 (part (b)), the conditional variance is actually smaller than the conditional mean, indicating that overdispersion has been removed in the fitting of the data. Another direct effect

Table 9.24 Type III fixed effects tests

Effect	Num DF	Den DF	F-value	Pr > F
Site	2	3	0.78	0.5346
T	4	12	11.57	0.0004
T*site	8	12	1.13	0.4096

Table 9.25 Means and standard errors on the model scale (Estimate) and on the data scale (Mean) of the sampling site data

Site	Estimate	Standard error	DF	t-value	Pr > t	Mean	Standard error mean
A	7.0195	0.1270	3	55.25	<0.0001	1118.25	142.06
B	7.0243	0.1271	3	55.28	<0.0001	1123.57	142.78
C	6.8237	0.1305	3	52.30	<0.0001	919.40	119.95

Table 9.26 Means and standard errors of measurement time on the model scale (Estimate) and the data scale (Mean)

T	Estimate	Standard error	DF	t-value	Pr > t	Mean	Standard error mean
1	6.2084	0.1467	12	42.32	<0.0001	496.91	72.8990
2	7.4212	0.1420	12	52.27	<0.0001	1670.97	237.23
3	7.0074	0.1455	12	48.16	<0.0001	1104.74	160.75
4	7.2910	0.1418	12	51.42	<0.0001	1466.97	208.00
5	6.8513	0.1422	12	48.19	<0.0001	945.09	134.35

observed when there is no overdispersion is the *F*-values of the fixed effects tests (Table 9.24). In this case, the date on which the samples were collected was significant but not the interaction between the two factors, as the case when the data were fitted using the Poisson GLMM.

The linear predictors and estimated probabilities of the main effects and the interaction between both factors are under the columns “Estimate” and “Mean,” respectively. The sampling site averages are presented below (Table 9.25).

The averages by sampling date are listed below (Table 9.26).

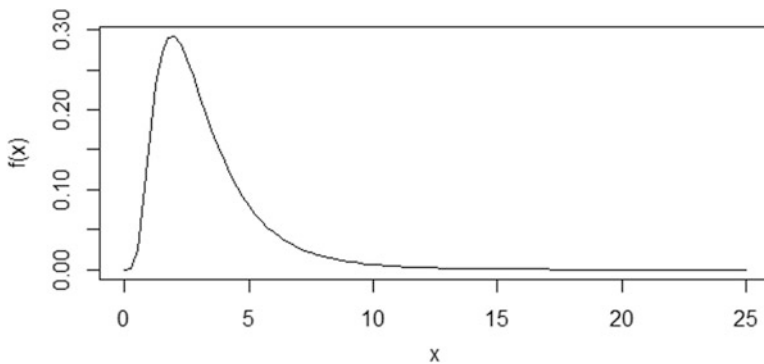
The means of the interaction site × sampling date are shown below (Table 9.27).

9.6 Log-Normal Distribution

Positively skewed distributions are highly common, especially when modeling biological data. Data often have a lower bound, usually 0 or the detection limit, but have no restriction on the upper bound. Therefore, when the data are below the median, no observation can be further away than the lower bound; however, when the data are above the median, there may be values that are many times further away, giving a positively skewed distribution. These skewed distributions can often be approximated by a log-normal distribution (Limpert et al. 2001).

Table 9.27 Means and standard errors for the interaction T*site on the model scale (Estimate) and the data scale (Mean)

Site	T	Estimate	Standard error	DF	t-value	Pr > t	Mean	Standard error mean
A	1	6.5905	0.2463		26.76	<0.0001	728.17	179.35
B	1	6.4197	0.2466		26.03	<0.0001	613.79	151.38
C	1	5.6151	0.2772		20.26	<0.0001	274.53	76.1038
A	2	7.2508	0.2452		29.57	<0.0001	1409.17	345.59
B	2	7.5367	0.2451		30.75	<0.0001	1875.54	459.64
C	2	7.4761	0.2463		30.36	<0.0001	1765.30	434.71
A	3	7.0336	0.2461		28.58	<0.0001	1134.09	279.14
B	3	6.8855	0.2465		27.94	<0.0001	978.01	241.05
C	3	7.1030	0.2586		27.47	<0.0001	1215.59	314.37
A	4	7.3224	0.2458		29.79	<0.0001	1513.87	372.07
B	4	7.4329	0.2450		30.33	<0.0001	1690.62	414.28
C	4	7.1176	0.2463		28.90	<0.0001	1233.47	303.81
A	5	6.9003	0.2460		28.04	<0.0001	992.56	244.21
B	5	6.8467	0.2458		27.85	<0.0001	940.73	231.25
C	5	6.8069	0.2453		27.75	<0.0001	904.07	221.80

**Fig. 9.1** Density function of the log-normal distribution with parameters 1 and 0.6

A log-normal distribution is characterized by having only positive nonzero values, positive skewness, a nonconstant variance that is proportional to the square of the mean value, and a normally distributed natural logarithm. The probability density function for a log-normal distribution has an asymmetric appearance, with a larger amount of data below the expected value and a thinner right tail with higher values. Figure 9.1 shows the positive skewness of a log-normal distribution with mean 1 and standard deviation 0.6.

9.6.1 Emission of Nitrous Oxide (N_2O) in Beef Cattle Manure with Different Percentages of Crude Protein in the Diet

The experiment was conducted between January and February 2017 at the Colegio de Postgraduados Campus Córdoba located in Amatlán de los Reyes, Veracruz, México. The genetic material used were four 5–6-month-old males of the Criollo lechero tropical (CLT) breed, randomly distributed in individual pens of $4.8 \times 2.1 \text{ m}^2$, each one with 75% shade, a cup drinker, and a drawer-type feeder. To ensure the required crude protein percentages for each treatment, the following diets (treatments 1–4) were developed: Trt1 (12% crude protein), Trt2 (14% crude protein), Trt3 (16% crude protein), and Trt4 (commercial feed with 16% crude protein). Each animal randomly received the four treatments in different periods. Each treatment was applied for 11 days, of which the first 7 were considered adaptation days and the following 4 days were used for the measurement of gases in the daily accumulated excreta. The experiment had a total duration of 44 days. The data from this experiment are tabulated in the Appendix (Data: Nitrous oxide emission). The N_2O gas fluxes in ppm were calculated from a linear or nonlinear increase of the concentrations inside the static chambers over time, and these fluxes were converted to micrograms of N_2O-N per m^2 per hour (y); for more details, see the study by Nadia Hernández-Tapia et al., (2019). The statistical model used in this study was an analysis of covariance model in a randomized complete block design with repeated measures, as described below.

$$y_{ijk} = \mu + \tau_i + \text{animal}_j + \text{time}_k + (\tau \times \text{time})_{ik} + \beta_i(x_{ij} - \bar{x}) + \varepsilon_{ijk}$$

where y_{ijk} is the flux of N_2O-N ($\mu\text{g m}^{-2} \text{ h}^{-1}$); μ is the overall mean; τ_i is the fixed effect due to treatment i ($i = 1, 2, 3, 4$); animal_j is the random effect due to animal j ($j = 1, 2, 3, 4$), assuming $\text{animal}_j \sim \mathcal{N}(0, \sigma^2_{\text{animal}})$; time_k is the fixed effect of time k ($k = 1, 2, 3, 4, 5$) at the time of measurement; $(\tau \times \text{time})_{ik}$ is the effect of the interaction between τ_i and time_k , β_i is the coefficient of linear regression of the covariate x_{ij} in treatment i and time j , where x_{ij} can be the pH, humidity (HE), temperature (TE) in the manure, maximum temperature (TMaxA), minimum temperature (TMinA), maximum humidity (HMaxA), minimum humidity (HMinA), or initial weight (kilograms) at the start of a treatment; \bar{x} is the mean of the covariate in question; and ε_{ijk} is the non-normal experimental error.

The linear predictor η_{ijk} for N_2O-N is $\eta_{ijk} = \mu + \tau_i + \text{animal}_j + \text{time}_k + (\tau \times \text{time})_{ik} + \beta_i(x_{ij} - \bar{x})$. The response variable y_{ijk} has a conditional log-normal distribution with a mean μ_{ijk} and variance $(e^{\sigma^2} - 1) \cdot e^{2\mu + \sigma^2}$, that is, $y_{ijk} | \text{animal}_j \sim \text{Log normal}(\mu_{ijk}, (e^{\sigma^2} - 1) \cdot e^{2\mu + \sigma^2})$; the rest of the parameters have already been described above.

The following GLIMMIX syntax adjusts a GLMM with a log-normal response:

```
proc glimmix data=co2 nobound method=laplace;
class animal trt time;
model flox =trt|time xbar/dist=lognormal;
random intercept /subject=animal type=cs;
lsmeans trt|time/lines ilink;
run;
```

Although most of the commands have already been described in previous chapters, in this chapter, we average the TMinA covariate “xbar.” Part of the output is shown below.

The gas emissions from cattle manure, regardless of the treatment applied, are influenced by several factors (covariates) that the researcher cannot control, which have a significant effect on the estimation of means and experimental error. Both are linearly related to the response variable. Covariates such as pH, humidity, and temperature of the excreta, as well as the temperature and humidity (maximum and minimum) of the environment, influence the dynamics of gas emission. These covariates were considered and analyzed in the covariance model to adjust the estimated means of the N₂O–N flux. Based on the fit statistics obtained from the proposed models (Table 9.28), the model that best explains the variability of the N₂O–N flux is model 5 because this model provides the lowest values in AIC, AICC, BIC, and MSE (Mean Square Error). Therefore, the model that provides the best fit or explains the most variability in the N₂O–N flux is the one that includes the minimum environment temperature.

The conditional fit statistics (part (a)) and the estimated variance components (part (b)) are shown in Table 9.29. The type III fixed effects tests (part (c)) indicate that there is a significant effect of Trt ($P = 0.0008$), time ($P = 0.0288$), the interaction Trt \times time ($P = 0.0140$), the covariate Tmin ($P = 0.0079$), and the interaction Tmin \times Trt ($P = 0.038$).

The average N₂O–N emissions between Trt1 (12% CP: Crude Protein) and Trt2 (14% CP) were statistically different from each other. Treatment 1 emitted the highest N₂O–N flux despite being the treatment with the lowest percentage of CP (Table 9.30).

9.7 Effect of a Chemical Salt on the Percentage Inhibition of the *Fusarium sp.*

In order to observe the tolerance of the fungus *Fusarium sp.* to different concentrations of a chemical salt, a bioassay was implemented to evaluate the percentage of inhibition of the fungus. This bioassay consisted of placing a nutritive culture medium in Petri dishes for the fungal development in which different concentrations of the salt in ppm were added (0, 500, 1000, and 2000,). Mycelium growth was measured during 6 days, and the percentage of inhibition of *Fusarium sp.* growth was calculated. Part of the data is shown below, and the complete base is in the Appendix (Data: Percentage inhibition).

Table 9.28 Fit statistics in the different models proposed

Model with the specific covariate	Fit statistics (N ₂ O-N)				
	AIC	AICC	BIC	CME	
1. $Y_{ijk} = \mu + \tau_i + \text{animal}_j + \text{time}_k + (\tau \times \text{time})_{ik} + \beta(pH_{X_{ij}} - \overline{pH}) + \varepsilon_{ijk}$	422.40	433.94	407.67	1.09	
2. $Y_{ijk} = \mu + \tau_i + \text{animal}_j + \text{time}_k + (\tau \times \text{time})_{ik} + \beta(HE_{ij} - \overline{HE}) + \varepsilon_{ijk}$	418.44	429.98	403.71	0.97	
3. $Y_{ijk} = \mu + \tau_i + \text{animal}_j + \text{time}_k + (\tau \times \text{tieme})_{ik} + \beta(TE_{ij} - \overline{TE}) + \varepsilon_{ijk}$	442.74	454.28	428.01	1.16	
4. $Y_{ijk} = \mu + \tau_i + \text{animal}_j + \text{time}_k + (\tau \times \text{time})_{ik} + \beta(TM_{MaxA_{ij}} - \overline{TM_{MaxA}}) + \varepsilon_{ijk}$	450.66	462.20	435.93	1.25	
5. $Y_{ijk} = \mu + \tau_i + \text{animal}_j + \text{time}_k + (\tau \times \text{time})_{ik} + \beta(TMinA_{ij} - \overline{TMinA}) + \varepsilon_{ijk}$	417.62	459.14	391.84	0.76	
6. $Y_{ijk} = \mu + \tau_i + \text{animal}_j + \text{time}_k + (\tau \times \text{tiempo})_{ik} + \beta(HM_{MaxA_{ij}} - \overline{HM_{MaxA}}) + \varepsilon_{ijk}$	443.88	455.42	429.15	1.18	
7. $Y_{ijk} = \mu + \tau_i + \text{animal}_j + \text{time}_k + (\tau \times \text{time})_{ik} + \beta(HM_{MinA_{ij}} - \overline{HM_{MinA}}) + \varepsilon_{ijk}$	416.78	428.32	402.05	0.99	

Table 9.29 Conditional fit statistics, variance components, and type III fixed effect tests

(a) Fit statistics for conditional distribution				
-2 Log L ($F \mid r$. effects)				333.62
Pearson's chi-square				99.06
Pearson's chi-square/DF				0.76
(b) Covariance parameter estimates				
Cov Parm	Subject	Estimate	Standard error	
Variance	A	-0.01561	.	
CS	A	0.000767	.	
Residual		0.7776	0.09845	
(c) Type III tests of fixed effects				
Effect	Num DF	Den DF	F-value	Pr > F
Trt	3	88	6.13	0.0008
Time	4	88	2.84	0.0288
Trt*time	11	88	2.34	0.0140
Tmin	1	88	7.40	0.0079
Tmin*Trt	3	88	4.81	0.0038
Tmin*time	4	88	1.80	0.1351
Tmin*Trt*time	11	88	1.23	0.2814

Table 9.30 Mean and standard error of N flux₂ O (μg of N₂O-N m⁻² h⁻¹) of the different treatments under study

Treatment	N ₂ O(μ) \pm standard error
Trt1 (12% PC)	3.6442 \pm 0.2213a
Trt2 (14% PC)	3.0714 \pm 0.3119b
Trt3 (16% PC)	3.5706 \pm 0.2974ab
Trt4 (16% CP, commercial feed)	3.1205 \pm 0.2130ab

Bio	Day	Conc	Rep	Y	Bio	Day	Conc	Rep	Y
1	1	0	3	5.263	2	1	0	2	0.0016
1	1	0	4	5.263	2	1	0	3	14.285
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
1	2	0	2	1.935	2	2	500	2	31.506
1	2	0	3	4.516	2	2	500	3	42.465
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
1	3	0	2	1.234	2	3	500	3	35.042
1	3	0	3	3.703	2	3	500	4	24.786
1	4	0	3	4.672	2	4	500	2	23.123
1	4	500	1	19.626	2	4	500	3	27.927
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
1	5	0	3	4.065	2	5	500	1	13.253
1	5	0	4	4.065	2	5	500	2	21.285
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
1	6	1000	3	15.862	2	6	2000	1	31.197
1	6	1000	4	18.62	2	6	2000	2	29.173
1	6	2000	1	32.413	2	6	2000	3	29.848

(continued)

Bio	Day	Conc	Rep	Y	Bio	Day	Conc	Rep	Y
1	6	2000	2	29.655	2	6	2000	4	30.522
1	6	2000	3	31.724					
1	6	2000	4	35.172					

Following the same reasoning as in previous examples, the components of the GLMM with beta response distribution repeated-measures for the percentage inhibition of *Fusarium* sp. (y_{ijkl}) are listed below:

Distributions: $y_{ijkl} | \omega_{kl}, \text{conc}(\omega)_{i(kl)} \sim \text{Beta}(\pi_{ijkl}, \phi)$; $i = 1, \dots, 4; j = 1, \dots, 6; k = 1, 2; l = 1, \dots, r_i$. $\omega_{kl} \sim N(0, \sigma_\omega^2)$, $\text{conc}(\omega)_{i(kl)} \sim N(0, \sigma_{\text{conc}(\omega)}^2)$.

Linear predictor: $\eta_{ijk} = \theta + \text{conc}_i + \omega_{kl} + \text{conc}(\omega)_{i(kl)} + \text{time}_j + (\text{conc} \times \text{time})_{ij}$.

where η_{ijk} is the linear predictor, θ is the intercept, conc_i is the fixed effect of salt concentration, ω_{kl} is the random effect of the Petri dish within the bioassay, assuming $\omega_{kl} \sim N(0, \sigma_\omega^2)$, $\text{conc}(\omega)_{i(kl)}$ is the random effect of salt concentration–Petri dish–bioassay, assuming $\text{conc}(\omega)_{i(kl)} \sim N(0, \sigma_{\text{conc}(\omega)}^2)$, time_j is the fixed effect due to the day of measurement, and $(\text{conc} \times \text{time})_{ij}$ is the interaction effect of chemical salt concentration with the day of measurement.

Link function: $\text{logit}(\pi_{ijkl}) = \eta_{ijkl}$ is the link function that relates the linear predictor to the mean (π_{ijkl}).

The following SAS program adjusts the beta GLMM with repeated measures.

```
proc glimmix data=inhibition method=laplace nobound;
class Bio Day Conc Rep;
model pct = Con|Day/dist=beta link=logit;
random intercept/subject=con(bio) type=cs;
lsmeans Con|Day/lines ilink;
run;
```

Before fitting the generalized linear mixed model, we compare the estimates of the covariance structures with the beta distribution in the response variable (Table 9.31 part (a)). According to the fit statistics, the covariance structures that best fit the data are the Toeplitz type (Toep(1)) and unstructured (UN).

Having defined the covariance structure, in this case, Toeplitz of order 1, we present part of the results of the data fit (Table 9.31 part (b)). The fit statistic Pearson's chi – square/DF = 1.07 indicates that there is no overdispersion and that the beta distribution fits the data adequately. The estimated variance component, under Toeplitz (1), of the concentration–repetition bioassay is $\hat{\sigma}_{\text{con}(\omega)}^2 = 0.00285$ and the scale parameter $\hat{\phi} = 52.281$ (c).

Table 9.31 Fit statistics for the conditional distribution and variance components

(a) Fit statistics	CS	AR(1)	Toep(1)	UN
-2 Log likelihood	-523.69	-523.69	-523.69	-523.69
AIC (smaller is better)	-469.69	-469.69	-471.69	-471.69
AICC (smaller is better)	-458.73	-458.73	-461.59	-461.59
BIC (smaller is better)	-467.54	-467.54	-469.62	-469.62
CAIC (smaller is better)	-440.54	-440.54	-443.62	-443.62
HQIC (smaller is better)	-484.16	-484.16	-485.62	-485.62
(b) Fit statistics for conditional distribution				
-2 Log L (pct r. effects)				-529.79
Pearson's chi-square				177.68
Pearson's chi-square/DF				1.07
(c) Covariance parameter estimates				
Cov Parm	Subject	Estimate	Standard error	
Variance	Con(Bio)	0.002849	0.004147	
Scale		52.2809	5.8849	

Table 9.32 Type III fixed effects tests

Type III tests of fixed effects				
Effect	Num DF	Den DF	F-value	Pr > F
Con	3	4	125.40	0.0002
Day	5	138	10.99	<0.0001
Day*Con	15	138	2.25	0.0074

The fixed effects indicate that there is a highly significant effect of salt concentration ($P = 0.0002$), time ($P = 0.0001$), and the interaction concentration x time ($P = 0.0074$) on the growth inhibition of *Fusarium sp.* (Table 9.32).

The linear predictors and estimated probabilities of the factors (Table 9.33 parts (a) and (b)) and interaction (Table 9.34) are found under the columns "Estimate" and "Mean," respectively.

9.8 Carbon Dioxide (CO₂) Emission as a Function of Soil Moisture and Microbial Activity

Productive agricultural soil requires a certain level of ventilation to maintain active plant root growth and soil microbial activity. One scientist found that soil oxygenation levels had been affected in soils fertilized with nutrient-rich sludge from a sewage treatment plant. The level of soil aeration can be reduced by (1) the high water content of the sludge added, through compaction with heavy machinery used to add the sludge and, ironically, (2) the increased microbial activity that occurs when sludge with high organic matter content is added. The objective of the research was to determine the moisture levels at which aeration becomes a limiting factor for

Table 9.33 Concentration and measurement time least square means on the model scale (Estimate) and the data scale (Mean)

(a) Conc least squares means							
Con	Estimate $\hat{\eta}_i$	Standard error	DF	t-value	Pr > t	Mean $\hat{\pi}_i$	Standard error mean
0	-3.5438	0.1499	4	-23.64	<0.0001	0.02809	0.004093
500	-1.0650	0.05941	4	-17.93	<0.0001	0.2563	0.01133
1000	-0.9847	0.05895	4	-16.70	<0.0001	0.2720	0.01167
2000	-0.4487	0.05891	4	-7.62	0.0016	0.3897	0.01401

(b) Day least squares means							
Day	Estimate $\hat{\eta}_j$	Standard error	DF	t-value	Pr > t	Mean $\hat{\pi}_j$	Standard error mean
1	-1.6017	0.1161	138	-13.79	<0.0001	0.1677	0.01621
2	-1.0446	0.08689	138	-12.02	<0.0001	0.2603	0.01673
3	-1.2475	0.08794	138	-14.19	<0.0001	0.2231	0.01524
4	-1.5668	0.1020	138	-15.36	<0.0001	0.1727	0.01457
5	-1.7606	0.1039	138	-16.94	<0.0001	0.1467	0.01301
6	-1.8422	0.1067	138	-17.26	<0.0001	0.1368	0.01260

microbial activity in the soil. The study included a control treatment (no sludge) and three treatments using sludge as a fertilizer with different moisture contents, whose moisture levels for the fertilized soil were 0.24, 0.26, and 0.28 kg water/kg soil.

Soil samples were randomly assigned to the four treatments in a randomized completely design. Soil samples were placed in sealed containers and incubated under favorable conditions for microbial activity. The soil was compacted in the containers simulating a degree of compaction experienced in the field. Microbial activity, measured as an increase in CO₂, was used as a measure of the level of soil oxygenation. The CO₂ evolution/kilogram soil/day in each container was measured on 2, 4, 6, and 8 days after starting of the incubation period. Microbial activity in each soil sample was recorded as the percentage increase in CO₂ produced above the atmospheric level. The data are shown in Table 9.35.

The analysis of variance table for this experiment is shown below (Table 9.36).

Let pct_{ijk} be the percentage of CO₂ emission, assuming that pct_{ijk} has a beta distribution with a mean π_{ijk} and scale parameter ϕ , i.e., $pct_{ijk} \sim \text{Beta}(\pi_{ijk}, \phi)$. The linear predictor η_{ijk} that relates the mean to the link function is given by

$$\eta_{ijk} = \theta + \alpha_i + \alpha(r)_{i(k)} + \tau_j + (\alpha\tau)_{ij}; i = 1, \dots, 4, j = 1, \dots, 4, k = 1, 2, 3$$

where θ is the intercept, α_i is the fixed effect of the treatment i , $\alpha(r)_{i(k)}$ is the random effect of treatment nested in the repetition k , assuming that $\alpha(r)_{i(k)} \sim N(0, \sigma_{\alpha(r)}^2)$, τ_j is the fixed effect of measurement time j , and $(\alpha\tau)_{ij}$ is the interaction effect of treatment with measurement time. The link function is defined by $\text{logit}(\pi_{ijk}) = \eta_{ijk}$.

The following SAS syntax fits a GLMM on repeated measures with a beta distribution.

Table 9.34 Measuring time*salt concentration interaction on the model scale (Estimate) and the data scale (Mean)

Day*con least squares means								
Day	Con	Estimate $\hat{\eta}_{ij}$	Standard error	DF	t-value	Pr > t	Mean $\hat{\pi}_{ij}$	Standard error mean
1	0	-4.0127	0.4083	138	-9.83	<0.0001	0.01776	0.007124
1	500	-0.8709	0.1123	138	-7.76	<0.0001	0.2951	0.02335
1	1000	-0.6848	0.1092	138	-6.27	<0.0001	0.3352	0.02434
1	2000	-0.8382	0.1579	138	-5.31	<0.0001	0.3019	0.03328
2	0	-3.5743	0.2957	138	-12.09	<0.0001	0.02727	0.007844
2	500	-0.4140	0.1061	138	-3.90	0.0001	0.3980	0.02543
2	1000	-0.3519	0.1053	138	-3.34	0.0011	0.4129	0.02554
2	2000	0.1616	0.1043	138	1.55	0.1235	0.5403	0.02590
3	0	-2.9511	0.2944	138	-10.02	<0.0001	0.04969	0.01390
3	500	-0.9923	0.1149	138	-8.64	<0.0001	0.2705	0.02266
3	1000	-0.9044	0.1131	138	-8.00	<0.0001	0.2881	0.02319
3	2000	-0.1423	0.1041	138	-1.37	0.1739	0.4645	0.02590
4	0	-3.5167	0.3558	138	-9.88	<0.0001	0.02884	0.009967
4	500	-1.2429	0.1213	138	-10.25	<0.0001	0.2239	0.02108
4	1000	-1.0361	0.1159	138	-8.94	<0.0001	0.2619	0.02241
4	2000	-0.4716	0.1065	138	-4.43	<0.0001	0.3842	0.02520
5	0	-3.5503	0.3579	138	-9.92	<0.0001	0.02791	0.009710
5	500	-1.4180	0.1269	138	-11.17	<0.0001	0.1950	0.01992
5	1000	-1.4489	0.1277	138	-11.34	<0.0001	0.1902	0.01967
5	2000	-0.6251	0.1083	138	-5.77	<0.0001	0.3486	0.02458
6	0	-3.6579	0.3691	138	-9.91	<0.0001	0.02514	0.009046
6	500	-1.4522	0.1277	138	-11.37	<0.0001	0.1897	0.01963
6	1000	-1.4823	0.1289	138	-11.50	<0.0001	0.1851	0.01944
6	2000	-0.7765	0.1106	138	-7.02	<0.0001	0.3151	0.02388

```
proc glimmix data=co2 method=laplace;
class trt container time;
model pct = trt|time/dist=beta link=logit;
random trt/subject=container;
lsmeans trt|time /lines ilink;
run;
```

Part of the results is shown below. The fit statistics under different covariance structures (Table 9.37 part (a)), such as AIC and AICC indicate that a Toeplitz-type covariance structure of order 1 provides the best fit to the dataset of this experiment.

Table 9.38 part (a) shows the estimated variance component due to treatment x repetition, i.e., $\hat{\sigma}_{a(r)}^2 = 0.03363$, and the estimated scale parameter $\hat{\phi} = 790.82$, and the hypothesis test (part (b)) indicates that the treatments yielded statistically different means ($P = 0.0011$).

Table 9.35 Repeated measurements of emissions of CO₂ by bacterial activity in soil under different moisture conditions

Moisture (kg water/kg soil)	Container	%CO ₂ evolution/kilogram soil/day			
		Day 2	Day 4	Day 6	Day 8
Control	1	0.22	0.56	0.66	0.89
	2	0.68	0.91	1.06	0.8
	3	0.68	0.45	0.72	0.89
0.24	1	2.53	2.7	2.1	1.5
	2	2.59	1.43	1.35	0.74
	3	0.56	1.37	1.87	1.21
0.26	1	0.22	0.22	0.2	0.11
	2	0.45	0.28	1.24	0.86
	3	0.22	0.33	0.34	0.2
0.28	1	0.22	0.8	0.8	0.37
	2	0.22	0.62	0.89	0.95
	3	0.22	0.56	0.69	0.63

Table 9.36 Analysis of variance of an RCD with repeated measures

Sources of variation	Degrees of freedom
Treatment	$(a - 1) = 4 - 1 = 3$
Error ₁	$a(r - 1) = 8$
Measurement time	$(t - 1) = 4 - 1 = 3$
Treatment x time	$(a - 1)(t - 1) = 9$
Error ₂	$a(t - 1)(r - 1) = 4 \times 3 \times 2 = 24$
Total	$a \times t \times r - 1 = 4 \times 4 \times 3 - 1 = 47$

Table 9.37 Fit statistics of the beta GLMM under different covariance structures

(a) Fit statistics	CS	AR(1)	Toep(1)	UN
-2 Log likelihood	-433.28	-433.94	-433.28	No converge
AIC (smaller is better)	-395.28	-395.94	-397.28	
AICC (smaller is better)	-368.14	-368.80	-373.69	
BIC (smaller is better)	-412.41	-413.07	-413.50	
CAIC (smaller is better)	-393.41	-394.07	-395.50	
HQIC (smaller is better)	-429.71	-430.37	-429.89	
(b) Fit statistics for conditional distribution	CS	AR(1)	Toep(1)	UN
-2 Log L (y r. effects)	-446.54	-444.41	-446.58	No converge
Pearson's chi-square	30.46	33.98	30.38	
Pearson's chi-square/DF	0.63	0.71	0.63	

Table 9.39 shows the estimated average emissions of CO₂ in tested treatments, which showed that the treatment with moisture 0.24 kg water/kg soil favored a higher microbial activity, whereas treatments with moisture levels 0.26 and 0.28 kg water/kg soil showed similar microbial activity between them.

Table 9.38 Variance components and fixed effects test

(a) Covariance parameter estimates				
Cov Parm	Subject	Estimate	Standard error	
Variance	Contenedor	0.03363	0.03153	
Scale (ϕ)		790.82	190.89	
(b) Type III tests of fixed effects				
Effect	Num DF	Den DF	F-value	Pr > F
Trt	3	8	15.52	0.0011
Time	3	24	2.94	0.0537
Trt*time	9	24	1.29	0.2914

Table 9.39 Means and standard errors on the model scale (Estimate) and the data scale (Mean)

(a) Trt least squares means							
Trt	Estimate	Standard error	DF	t-value	Pr > t	Mean	Standard error mean
C	-4.9242	0.1595	8	-30.87	<0.0001	0.007216	0.001143
T0.24	-4.1331	0.1343	8	-30.79	<0.0001	0.01578	0.002085
T0.26	-5.5728	0.1898	8	-29.36	<0.0001	0.003786	0.000716
T0.28	-5.1588	0.1728	8	-29.86	<0.0001	0.005716	0.000982

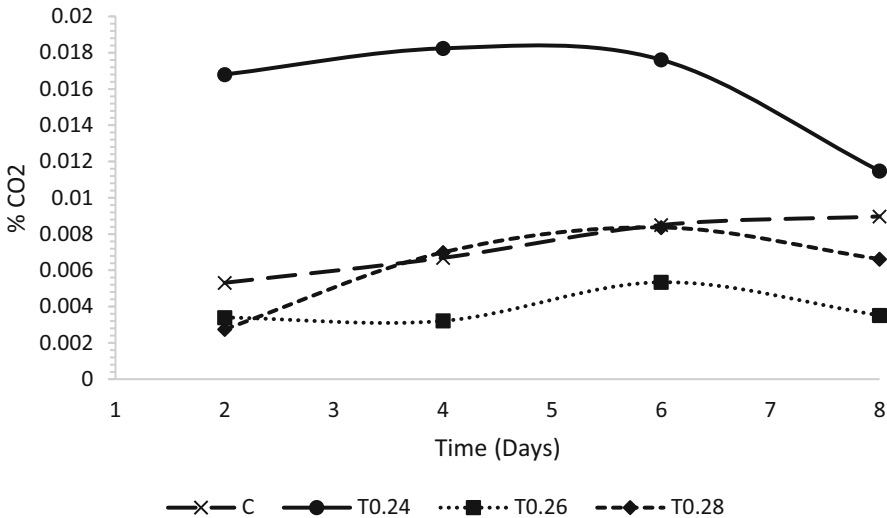


Fig. 9.2 CO₂ emission as a measure of microbial activity

Figure 9.2 clearly shows that the treatment with moisture 0.24 kg water/kg soil provides the best conditions for soil microbial activity, whereas the rest of the treatments significantly affect the activity of microorganisms.

9.9 Effect of Soil Compaction and Soil Moisture on Microbial Activity

A soil scientist conducted an experiment to evaluate the effects of soil compaction and soil moisture on microbial activity. Ventilation levels may be restricted in highly saturated or compacted soils, thus reducing microbial activity. The experiment consisted of three levels of soil compaction (1.1, 1.4, and 1.6 mg soil/m³) and three levels of soil moisture (0.1, 0.2, and 0.24 kg water/kg soil). The treated soil samples were placed in sealed containers and incubated under conditions to microbial activity. The percentage increase in CO₂ produced above atmospheric levels was measured in each soil sample. The experimental design was a completely randomized design (CRD) with a 3 X 3 factorial structure of treatments. Two replicates of the soil container units were prepared for each treatment. The evolution of CO₂/kg soil/day was measured for three successive days. The data from this experiment are shown below in Table 9.40.

The analysis of variance table for this experiment is shown below (Table 9.41).

Let pct_{ijk} be the percentage of CO₂ emission and assume that pct_{ijk} has a beta distribution with a mean π_{ijk} and scale parameter ϕ , i.e., $pct_{ijk} \sim \text{Beta}(\pi_{ijk}, \phi)$. The linear predictor η_{ijkl} that relates the mean to the link function is given by

$$\eta_{ijkl} = \theta + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \alpha\beta(r)_{ij(l)} + \tau_k + (\alpha\tau)_{ik} + (\beta\tau)_{jk} + (\alpha\beta\tau)_{ijk}$$

Table 9.40 Percentage of CO₂ by bacterial activity as a function of soil density (mg soil/m³) and soil humidity (kg water/kg soil)

Density	Humidity	Replication	Day 1	Day 2	Day 3
1.1	0.1	1	2.7	0.34	0.11
		2	2.9	1.57	1.25
	0.2	1	5.2	5.04	3.7
		2	3.6	3.92	2.69
	0.24	1	4	3.47	3.47
		2	4.1	3.47	2.46
1.4	0.1	1	2.6	1.12	0.9
		2	2.2	0.78	0.34
	0.2	1	4.3	3.36	3.02
		2	3.9	2.91	2.35
	0.24	1	1.9	3.02	2.58
		2	3	3.81	2.69
1.6	0.1	1	2	0.67	0.22
		2	3	0.78	0.22
	0.2	1	3.8	2.8	2.02
		2	2.6	3.14	2.46
	0.24	1	1.3	2.69	2.46
		2	0.5	0.34	.

Table 9.41 Analysis of variance of an CRD with factorial structure of treatments in repeated measures

Sources of variation	Degrees of freedom
Treatment	$(a - 1) = 3 - 1 = 2$
Humidity	$(b - 1) = 3 - 1 = 2$
Treatment*humidity	$(a - 1)(b - 1) = 4$
Error ₁	$ab(r - 1) = 3 \times 3 \times 1 = 9$
Time	$(c - 1) = 3 - 1 = 2$
Treatment time	$(a - 1)(c - 1) = 4$
Humidity*time	$(b - 1)(c - 1) = 4$
Treat*hum*time	$(a - 1)(b - 1)(c - 1) = 8$
Error ₂	<i>ldiferencial/17</i>
Total	$a \times b \times c \times r - 1 = 3 \times 3 \times 3 \times 2 - 1 - 1 = 52$

Note: Here, 1 degree of freedom was subtracted from the total observations of the experiment since there is a missing observation

$$i = 1, 2, 3, j = 1, 2, 3, k = 1, 2, 3, l = 1, 2$$

where θ is the intercept, α_i is the fixed effect of the density factor, β_j is the fixed effect of the humidity factor, $(\alpha\beta)_{ij}$ is the effect of the interaction between density and humidity, $\alpha\beta(r)_{ij(l)}$ is the random effect of the interaction density \times humidity \times repetition $\alpha\beta(r)_{ij(l)} \sim N(0, \sigma_{\alpha\beta(r)}^2)$, τ_l is the fixed effect of measurement time, $(\alpha\tau)_{ij}$ is the fixed effect of the interaction between density and measurement time, $(\beta\tau)_{jk}$ is the fixed effect of the interaction between moisture and measurement time, and $(\alpha\beta\tau)_{ijk}$ is the fixed effect of the interaction of density \times humidity \times time. The link function is defined by $\text{logit}(\pi_{ijkl}) = \eta_{ijkl}$.

The following SAS GLIMMIX syntax fits a repeated measures GLMM with a beta distribution.

```
proc glimmix data=co2_fact nobound method=laplace;
class density moisture rep time;
model pct = density|humidity|time/dist=beta link=logit;
random density*humidity/subject=rep type=toep(1);
lsmeans density|humidity|time/lines ilink;
run;
```

Part of the results is listed below. The fit statistics (AIC and AICC) in Table 9.42 part (a) indicate that a Toeplitz covariance structure of order 1 provides the best fit to of the data.

The type III tests of fixed effects in Table 9.43 indicate that soil density ($P = 0.0021$), humidity ($P = 0.0001$), the evolution of emission over time ($P = 0.0001$), and the interaction between moisture and time of measurement ($P = 0.0001$) are statistically significant.

Table 9.42 Fit statistics of a beta GLMM with a factorial structure of treatments under different covariance structures

(a) Fit statistics	CS	AR(1)	Toep(1)	UN
-2 Log likelihood	-413.74	-413.72	-413.72	No converge
AIC (smaller is better)	-353.74	-353.72	-355.72	
AICC (smaller is better)	-269.19	-269.18	-280.07	
BIC (smaller is better)	-392.94	-392.93	-393.62	
CAIC (smaller is better)	-362.94	-362.93	-364.62	
HQIC (smaller is better)	-435.73	-435.71	-434.98	
(b) Fit statistics for conditional distribution	CS	AR(1)	Toep(1)	
-2 Log L (y r. effects)	-413.74	-413.72	-413.72	No converge
Pearson's chi-square	64.60	64.65	64.65	
Pearson's chi-square/DF	1.22	1.22	1.22	

Table 9.43 Hypothesis testing of the factors under study

Type III tests of fixed effects				
Effect	Num DF	Den DF	F-value	Pr > F
Density	2	9	13.14	0.0021
Humidity	2	9	69.66	<0.0001
Density*humidity	4	9	3.57	0.0524
Time	2	17	21.97	<0.0001
Density*time	4	17	0.72	0.5904
Humidity*time	4	17	17.85	<0.0001
Density*humidity*time	8	17	2.12	0.0923

The least mean squares obtained with the “lsmeans” command on the model scale are shown under the “Estimate” column and the data scale under the “Mean” column of Table 9.44.

9.10 Joint Model for Binary and Poisson Data

Another advantage of the GLIMMIX procedure is the ability to fit models to data where the distribution and/or link function varies with response variables. This is accomplished through the specification of DIST = BYOBS or LINK=BYOBS in the model definition. The dataset created below provides an example of a variable with a bivariate outcome. This reflects the condition and length of hospital stay for 32 patients with herniorrhaphy. These data are taken from data provided by Mosteller and Tukey (1977) and reproduced in the study by Hand et al. (1994) (Table 9.45).

For each patient, two responses were recorded. A binary response takes the value one if a patient experienced a routine recovery and the value zero if postoperative intensive care was required. The second response variable is a count variable that

Table 9.44 Means and standard errors and comparison of means (least significance difference (LSD)) on the model scale (Estimate) and data scale (Mean)

(a) Density*humidity least squares means

Density	Humidity	Estimate	Standard error	DF	t-value	Pr > t	Mean	Standard error mean
1.1	0.1	-4.5961	0.1614	9	-28.48	<0.0001	0.009990	0.001596
1.1	0.2	-3.1829	0.07606	9	-41.85	<0.0001	0.03981	0.002908
1.1	0.24	-3.3152	0.08060	9	-41.13	<0.0001	0.03505	0.002726
1.4	0.1	-4.4567	0.1450	9	-30.74	<0.0001	0.01147	0.001643
1.4	0.2	-3.3798	0.08333	9	-40.56	<0.0001	0.03293	0.002654
1.4	0.24	-3.5363	0.08932	9	-39.59	<0.0001	0.02830	0.002456
1.6	0.1	-4.7890	0.1809	9	-26.47	<0.0001	0.008252	0.001481
1.6	0.2	-3.5453	0.08972	9	-39.52	<0.0001	0.02805	0.002446
1.6	0.24	-4.3213	0.1440	9	-30.00	<0.0001	0.01311	0.001863

(b) T grouping of density*humidity least squares means ($\alpha = 0.05$)

LS means with the same letter are not significantly different

Density	Humidity	Estimate	
1.1	0.20	-3.1829	A
1.1	0.24	-3.3152	B A
1.4	0.20	-3.3798	B A
1.4	0.24	-3.5363	B
1.6	0.20	-3.5453	B
1.6	0.24	-4.3213	C
1.4	0.10	-4.4567	C
1.1	0.10	-4.5961	C
1.6	0.10	-4.7890	C

measures the length of hospital stay after the surgery (in days). The binary variable “OKstatus” is a regressor variable that distinguishes patients according to their postoperative physical status (“1” implies better status), and the variable age is the age of the patient.

These data can be modeled with a separate logistic model for the binary outcome and with a Poisson model for the count outcome. Such separate analyses would not take into account the correlation between the two response variables. It is reasonable to assume that the duration of post-surgery hospitalization is correlated and will depend on whether the patient requires intensive care.

In the following analysis, the correlation between the two types of response variables for a patient is modeled with shared random effects (G-side). The dataset variable “dist” identifies the distribution for each observation. For those observations that follow a binary distribution, the response variable option “(event = “1”)” determines which value of the binary variable is modeled as the event of interest. Since no “link” option is specified, the link is also chosen on an observation-by-observation basis as a predetermined link for the respective distribution. The following GLIMMIX commands fit this dataset with two distributions:

Table 9.45 Hospital condition and length of stay of patients

D	Patient	Age	OKstatus	y	D	Patient	Age	OKstatus	y
B	1	78	1	0	B	17	79	0	0
P	1	78	1	9	P	17	79	0	3
B	2	60	1	0	B	18	51	1	1
P	2	60	1	4	P	18	51	1	5
B	3	68	1	1	B	19	57	1	1
P	3	68	1	7	P	19	57	1	8
B	4	62	0	1	B	20	51	0	1
P	4	62	0	35	P	20	51	0	8
B	5	76	0	0	B	21	48	1	1
P	5	76	0	9	P	21	48	1	3
B	6	76	1	1	B	22	48	1	1
P	6	76	1	7	P	22	48	1	5
B	7	64	1	1	B	23	66	1	1
P	7	64	1	5	P	23	66	1	8
B	8	74	1	1	B	24	71	1	0
P	8	74	1	16	P	24	71	1	2
B	9	68	0	1	B	25	75	0	0
P	9	68	0	7	P	25	75	0	7
B	10	79	1	0	B	26	2	1	1
P	10	79	1	11	P	26	2	1	0
B	11	80	0	1	B	27	65	1	0
P	11	80	0	4	P	27	65	1	16
B	12	48	1	1	B	28	42	1	0
P	12	48	1	9	P	28	42	1	3
B	13	35	1	1	B	29	54	1	0
P	13	35	1	2	P	29	54	1	2
B	14	58	1	1	B	30	43	1	1
P	14	58	1	4	P	30	43	1	3
B	15	40	1	1	B	31	4	1	1
P	15	40	1	3	P	31	4	1	3
B	16	19	1	1	B	32	52	1	1
P	16	19	1	4	P	32	52	1	8

```

data Poi_Bin;
length dist $7;
input d$ patient age OKstatus response @@;
if d= 'B' then dist='Binary'; else dist='Poisson';
datalines;
B 1 78 1 0 P 1 78 1 9 B 2 60 1 0 P 2 60 1 4
B 3 68 1 1 P 3 68 1 7 B 4 62 0 1 P 4 62 0 35
.....
.....
.....

```

Table 9.46 Model information

Model information	
Dataset	WORK.POI_BIN
Response variable	Response
Response distribution	Multivariate
Link function	Multiple
Variance function	Default
Variance matrix blocked by	Patient
Estimation technique	Residual pseudo-likelihood (PL)
Degrees of freedom method	Containment

```

.....
.....
.....
B 29 54 1 0 P 29 54 1 2 B 30 43 1 1 1 P 30 43 1 3
B 31 4 1 1 1 P 31 4 1 3 B 32 52 1 1 1 P 32 52 1 8
;
proc glimmix data=joint;
class patient dist;
model response(event='1') = dist dist*age dist*OKstatus /
noint s dist=byobs(dist);
random int / subject=patient;
lsmeans dist/lines ilink;
run;

```

Some of the output is shown below. Table 9.46 (“Model information”) shows that the distribution of the data is multivariate and that possibly multiple link functions are involved; by default, proc. GLIMMIX uses a logit link for the binary observations and a log link for the Poisson data.

Table 9.47 shows the value of the distribution statistic Gener. chi – square/DF = 0.90, which indicates that there is no overdispersion, and also shows the estimated variance component due to patient, which is, $\hat{\sigma}_{\text{patient}}^2 = 0.299$. The fixed effects tests for the effects of age and status are shown in part (c).

In addition to the above results, the maximum likelihood estimators of the intercepts, as well as the values of the slopes of each of the variables of both probability distributions, are tabulated in Table 9.48.

Thus, to calculate the probability that a patient will experience a routine recovery, the following expression is used:

$$\hat{\pi} = \frac{1}{1 + \exp \left\{ -\hat{\beta}_0 - \hat{\beta}_1 \times \text{age} - \hat{\beta}_2 \times \text{okstatus} \right\}}$$

$$= \frac{1}{1 + \exp \left\{ -5.7783 + 0.07572 \times \text{age} + 0.4697 \times \text{okstatus} \right\}}$$

Table 9.47 Results of the analysis of variance

(a) Fit statistics				
-2 Res log pseudo-likelihood				226.71
Generalized chi-square				52.25
Gener. chi-square/DF				0.90
(b) Covariance parameter estimates				
Cov Parm	Subject	Estimate	Standard error	
Intercept	Patient	0.2990	0.1116	
(c) Type III tests of fixed effects				
Effect	Num DF	Den DF	F-value	Pr > F
Dist	2	29	2.74	0.0814
Age*dist	2	29	5.94	0.0069
OKstatus*dist	2	29	0.24	0.7909

Table 9.48 Maximum likelihood estimators for fixed effects

Solutions for fixed effects						
Effect	Dist	Estimate	Standard error	DF	t-value	Pr > t
Dist	Binary	5.7783	2.9048	29	1.99	0.0562
Dist	Poisson	0.8410	0.5696	29	1.48	0.1506
Age*dist	Binary	-0.07572	0.03791	29	-2.00	0.0552
Age*dist	Poisson	0.01875	0.007383	29	2.54	0.0167
OKstatus*dist	Binary	-0.4697	1.1251	29	-0.42	0.6794
OKstatus*dist	Poisson	-0.1856	0.3020	29	-0.61	0.5435

whereas the following expression is used to calculate the average value of the length of hospital stay after the surgery (in days):

$$\hat{\lambda} = \exp \{ \hat{\alpha}_0 + \hat{\alpha}_1 \times \text{age} + \hat{\alpha}_2 \times \text{okstatus} \} = \exp \{ 0.8410 + 0.01875 \times \text{age} - 0.1856 \times \text{okstatus} \}$$

9.11 Exercises

Exercise 9.11.1 Consider an experiment in which three treatments are compared. There are r blocks of n animals, each using grouping criteria relevant to the experiment. Within each block, one animal is randomly assigned to each treatment. A measurement was taken on animals at “week 0,” when treatments were applied, and again at weeks 4 and 12. Variables measured included weight, the presence or absence of disease symptoms, and severity of symptoms, classified as “worse,” “no change,” or “better.” The focus of this experiment was on repeated measures analysis of the last two types of data in the above list: categorical data that are binary or ordinal and ordinal responses/ratings in an experiment designed with a repeated measures and treatment factor structure. Regardless of whether the observations are

Table 9.49 Results of a repeated measures experiment with an ordinal response variable

	Week 0			Week 4			Week 12		
Response	Placebo	Trt1	Trt2	Response	Placebo	Trt1	Trt2	Response	Placebo
Bad	60	59	54	14	5	3	13	10	7
Without change	7	6	13	34	33	38	25	17	21
Better	0	0	0	15	22	17	17	28	21

normally distributed, categorical, or have some other distribution, a general approach to repeated measures analysis based on the linear mixed model uses the following general form:

$$\text{Observation} = \text{systematic between} - \text{subjects variation} + \text{random between} - \text{subjects variation} + \text{systematic within} - \text{subjects effects} + \text{random within} - \text{subjects variation}.$$

The following table shows the data from an experiment in which each cell contains the number of animals in a given treatment × week × response category combination (Table 9.49).

- (a) List all the components of the repeated measures under a multinomial GLMM.
- (b) Study and choose the best covariance structure that models this dataset. Cite the most relevant results.
- (c) Fit the multinomial cumulative logit model to these data. Perform a complete and appropriate analysis of the data, focusing on:
 - (i) An evaluation of the effects of the combination of treatments
 - (ii) Odds ratio interpretation
 - (iii) The expected probability per category for each treatment
- (d) Test whether the proportional odds assumption is viable. Cite relevant evidence to support your conclusion regarding the adequacy of the assumption.

Repeat (b) through (d), assuming a generalized multinomial logit in Exercise 9.11.1. Discuss your results.

Repeat (b) through (d) assuming a multinomial cumulative probit in Exercise 9.11.1. Discuss your results and compare with those found in (1) and (2).

Alternatively, the contingency table approach can be implemented using a log-linear model. For the previous example, 9.11.1, fit the log-linear model

$$\log(\lambda_{ijk}) = \mu + \tau_i + \varpi_j + (\tau\varpi)_{ij} + c_k + (\tau c)_{ik} + (\tau\varpi c)_{ijk}$$

where λ_{ijk} is the expected count of the treatment combination ijk by week by response category and τ , ϖ , and c refer to treatment, week, and response category effects, respectively.

Table 9.50 Nitrogen injection treatment factors study

Handling practices				Application rate			
1 = N surface applied without additional water injection				2.5 g/cm ²			
2 = N surface area applied with supplementary water injection							
3 = N injected with a number 56 nozzle (7.6 cm depth of injection)				5.0 g/cm ²			
4 = N injected with a number 53 nozzle (12.7 cm depth of injection)							
Handling practices 1				Driving practice 2			
Quality	N_1	N_2	Total	Quality	N_1	N_2	Total
Poor	14	5	19	Poor	15	8	23
Average	2	11	13	Average	1	8	9
Good	0	0	0	Good	0	0	0
Excellent	0	0	0	Excellent	0	0	0
Total	16	16	32	Total	16	16	32
Handling practices 3				Handling practices 4			
Quality	N_1	N_2	Total	Quality	N_1	N_2	Total
Poor	0	0	0	Poor	1	0	1
Average	9	2	11	Average	12	4	16
Good	7	14	21	Good	0	0	0
Excellent	0	0	0	Excellent	0	0	0
Total	16	16	32	Total	16	16	32

Exercise 9.11.2 Fertilization of turf has traditionally been accomplished through surface applications. The introduction of new equipment (Hydroject) has made it possible to place soluble materials below the surface (Table 9.50).

A study was conducted during the 1997 growing season to compare surface application and subsoil injection of nitrogen on the green color of bentgrass (*Agrostis palustris L. Huds*) 1 year after transplanting. The treatment structure was a full factorial of grass management factors (four types/levels) and the rate/level (two levels) of nitrogen application per square meter (g/m²). Eight treatment combinations were arranged in a completely randomized design with four replications. Turf color was evaluated in each experimental unit at weekly intervals of 4 weeks as poor, average, good, or excellent.

Of particular interest was the determination of the water injection effect, the subsurface effect, and the comparison of injection versus surface applications. These are contrasts between the levels of factor management practice and their primary objective, which was to determine whether the factor interacts with the rate of application.

- (a) List all the GLMM components of this experiment.
- (b) Fit the multinomial cumulative logit proportional odds model to these data. Perform a complete and appropriate analysis of the data, focusing on:
 - (i) An evaluation of the effects of the combination of treatments
 - (ii) Interpretation of the odds ratios
 - (iii) The expected probability per category for each treatment

- (c) Test whether the proportional odds assumption is viable. Cite relevant evidence to support your conclusion regarding the adequacy of the assumption.

Exercise 9.11.3 Refer to Exercise 9.11.1.

- (a) Fit the multinomial generalized logit proportional odds model to these data.
 (b) List all the components of the GLMM of this experiment.
 (c) Perform a complete and appropriate analysis of the data, focusing on:
- (i) An evaluation of the effects of the combination of treatments
 - (ii) Interpretation of odds ratios
 - (iii) The expected probability per category for each treatment
- (d) Test whether the proportional odds assumption is viable. Cite relevant evidence to support your conclusion regarding the adequacy of the assumption.

Exercise 9.11.4 Refer to Exercise 9.11.1.

- (a) List all the components of the GLMM of this experiment.
 (b) Fit the multinomial cumulative probit proportional odds model to these data. Perform a complete and appropriate analysis of the data, focusing on:
- (i) An evaluation of the effects of the combination of treatments
 - (ii) Interpretation of the odds ratios
 - (iii) The expected probability per category for each treatment
- (c) Test whether the proportional odds assumption is viable. Cite relevant evidence to support your conclusion regarding the adequacy of the assumption.

Appendix

Data: Feeding line experiment

Tray	Feeding	Run	Proportion	Tray	Feeding	Run	Proportion
1	H	1	0.18217	1	H	3	0.06818
2	H	1	0.15493	2	H	3	0.05874
3	H	1	0.15906	3	H	3	0.05757
4	H	1	0.15869	4	H	3	0.10349
5	H	1	0.14891	5	H	3	0.08564
6	H	1	0.17654	6	H	3	0.09359
7	H	1	0.12915	7	H	3	0.09706
8	H	1	0.12895	8	H	3	0.13188
9	H	1	0.16688	9	H	3	0.18477
10	H	1	0.11965	10	H	3	0.10966
11	H	1	0.21719	11	H	3	0.18069
12	H	1	0.20797	12	H	3	0.18182

(continued)

Data: Feeding line experiment

Tray	Feeding	Run	Proportion	Tray	Feeding	Run	Proportion
1	L	1	0.70601	1	L	3	0.75524
2	L	1	0.68817	2	L	3	0.77249
3	L	1	0.68317	3	L	3	.
4	L	1	0.77805	4	L	3	0.84204
5	L	1	0.76692	5	L	3	0.81572
6	L	1	0.79127	6	L	3	0.79161
7	L	1	0.73653	7	L	3	0.81234
8	L	1	0.74939	8	L	3	0.81795
9	L	1	0.78773	9	L	3	0.8225
10	L	1	0.7381	10	L	3	0.79384
11	L	1	0.88486	11	L	3	0.8135
12	L	1	0.90401	12	L	3	0.83965
1	H	2	0.07547	1	H	4	0.07105
2	H	2	0.05801	2	H	4	0.05511
3	H	2	0.0565	3	H	4	0.05217
4	H	2	0.09579	4	H	4	0.10567
5	H	2	0.10954	5	H	4	0.0755
6	H	2	0.12154	6	H	4	0.0853
7	H	2	0.1144	7	H	4	0.09363
8	H	2	0.13728	8	H	4	0.11154
9	H	2	0.15012	9	H	4	0.16264
10	H	2	0.12113	10	H	4	0.09215
11	H	2	0.17633	11	H	4	0.1834
12	H	2	0.16408	12	H	4	0.21016
1	L	2	0.78318	1	L	4	0.76556
2	L	2	0.78418	2	L	4	0.78307
3	L	2	0.78589	3	L	4	0.76486
4	L	2	0.78867	4	L	4	0.82391
5	L	2	0.81988	5	L	4	0.8044
6	L	2	0.82793	6	L	4	0.81178
7	L	2	0.81384	7	L	4	0.84339
8	L	2	0.81037	8	L	4	0.78833
9	L	2	0.77528	9	L	4	0.79804
10	L	2	0.78916	10	L	4	0.82236
11	L	2	0.87109	11	L	4	0.83807
12	L	2	0.84704	12	L	4	0.85532

Data: Nitrous oxide emission

A	T	T	F	pH	HE	TE	tx	tm	Hx	Hm	A	T	t	F	pH	HE	TE	tx	tm	Hx	Hm
3	1	1	108	8	85	20	17	16	69	68	2	1	1	4.13	6.9	87	20	19	19	67	66
2	2	1	15	4.7	86	20	17	16	69	68	3	2	1	82.4	7	87	22	19	19	67	66
2	3	1	33	5.5	85	20	17	16	69	68	4	3	1	51.4	6.4	87	20	19	19	67	66
4	4	1	23	6.4	84	21	17	16	69	68	1	4	1	704	95	20	19	19	67	66	170
3	1	2	-58	8	85	23	34	34	54	51	3	2	2	14	7	87	22	27	27	68	67
2	2	2	-45	4.7	86	24	34	34	54	51	4	3	2	130	6.4	87	24	27	27	68	67
1	3	2	-97	5.5	85	29	34	34	54	51	1	4	2	537	95	24	27	27	68	67	170
4	4	2	185	6.4	84	28	34	34	54	51	3	2	3	1.02	7	87	25	28	27	63	57
3	1	3	47	8	85	32	35	35	38	30	4	3	3	41.9	6.4	87	24	28	27	63	57
2	2	3	26	4.7	86	32	35	35	38	30	1	4	3	824	95	24	28	27	63	57	170
1	3	3	41	5.5	85	34	35	35	38	30	3	2	4	53.9	7	87	20	22	22	61	60
4	4	3	19	6.4	84	34	35	35	38	30	4	3	4	9.88	6.4	87	20	22	22	61	60
3	1	4	311	8	85	28	30	30	40	38	1	4	4	745	95	20	22	22	61	60	170
2	2	4	-29	4.7	86	27	30	30	40	38	3	2	5	92.7	7	87	20	22	22	65	62
1	3	4	37	5.5	85	28	30	30	40	38	4	3	5	187	6.4	87	20	22	22	65	62
4	4	4	204	6.4	84	27	30	30	40	38	1	4	5	591	95	20	22	22	65	62	170
3	1	5	6.8	8	85	22	27	27	51	43	3	2	1	7	6.6	85	20	20	18	80	74
2	2	5	-18	4.7	86	22	27	27	51	43	4	3	1	8.39	6.8	87	20	20	18	80	74
1	3	5	68	5.5	85	22	27	27	51	43	1	4	1	1367	86	20	20	18	80	74	170
4	4	5	91	6.4	84	22	27	27	51	43	3	2	5	-49	6.6	85	19	18	18	89	89
3	1	1	135	4.6	84	20	18	18	60	59	4	3	5	-91	6.8	87	19	18	18	89	89
2	2	1	1.4	4.3	85	20	18	18	60	59	1	4	5	711	86	19	18	18	89	89	170
1	3	1	55	6.5	85	20	18	18	60	59	3	2	1	108	7.1	86	20	17	16	87	87
4	4	1	18	6.1	85	21	18	18	60	59	4	3	1	15	6.8	87	20	17	16	87	87
3	1	5	5	4.6	84	22	24	24	61	59	1	4	1	621	86	20	17	16	87	87	170

2	2	5	12	4.3	85	22	24	24	61	59	3	2	5	6.19	7.1	86	26	24	24	65	64
1	3	5	121	6.5	85	22	24	24	61	59	4	3	5	1.36	6.8	87	24	24	24	65	64
4	4	5	51	6.1	85	23	24	24	61	59	1	4	5	656	86	25	24	24	65	64	170
3	1	1	21	4.3	85	19	18	17	61	58	3	2	1	18.2	7.3	86	20	18	17	77	76
2	2	1	87	4.6	86	19	18	17	61	58	4	3	1	55.9	7	87	20	18	17	77	76
1	3	1	21	6.5	85	19	18	17	61	58	1	4	1	731	90	19	18	17	77	76	170
4	4	1	28	6.1	85	19	18	17	61	58	3	2	5	-77	7.3	86	26	25	25	65	63
3	1	5	31	4.3	85	23	25	25	57	55	4	3	5	33.6	7	87	25	25	25	65	63
2	2	5	101	4.6	86	23	25	25	57	55	1	4	5	880	90	26	25	25	65	63	163
1	3	5	-37	6.5	85	23	25	25	57	55	1	2	1	29.4	7.5	88	20	20	21	60	59
4	4	5	136	6.1	85	23	25	25	57	55	2	3	1	49.3	7.1	88	20	20	21	60	59
3	1	1	26	4.4	84	19	19	19	61	60	3	4	1	69.7	6.7	85	20	20	21	60	59
2	2	1	16	4.8	85	19	19	19	61	60	4	1	2	36.1	6.8	90	24	42	29	51	20
1	3	1	92	5.5	87	19	19	19	61	60	1	2	2	-100	7.5	88	24	42	29	51	20
4	4	1	-82	6	85	19	19	19	61	60	2	3	2	123	7.1	88	24	42	29	51	20
3	1	5	35	4.4	84	22	24	22	67	62	3	4	2	45.3	6.7	85	24	42	29	51	20
2	2	5	-10	4.8	85	25	24	22	67	62	4	1	3	18.2	6.8	90	30	39	34	45	29
1	3	5	41	5.5	87	23	24	22	67	62	1	2	3	-216	7.5	88	30	39	34	45	29
4	4	5	19	6	85	23	24	22	67	62	2	3	3	-71	7.1	88	30	39	34	45	29
1	1	1	16	54	160	161	4	2	2	1	3	4	3	3.47	6.7	85	30	39	34	45	29
3	3	1	83	5.8	85	21	16	19	50	54	4	1	4	57.6	6.8	90	26	31	28	26	23
2	4	1	28	7.2	84	20	16	19	50	54	1	2	4	59.1	7.5	88	26	31	28	26	23
1	1	2	24	55	160	161	4	2	2	2	2	3	4	38.2	7.1	88	26	31	28	26	23
3	3	2	29	5.8	85	23	24	23	58	55	3	4	4	74.4	6.7	85	26	31	28	26	23
2	4	2	9.4	7.2	84	22	24	23	58	55	4	1	5	26	6.8	90	24	26	25	43	31
1	1	3	29	44	160	161	4	2	2	3	1	2	5	67.5	7.5	88	24	26	25	43	31
3	3	3	38	5.8	85	31	29	25	51	44	2	3	5	-77	7.1	88	24	26	25	43	31

(continued)

Data: Nitrous oxide emission

A	T	T	F	pH	HE	TE	tx	tm	Hx	Hm	A	T	t	F	pH	HE	TE	tx	tm	Hx	Hm
2	4	3	1.2	7.2	84	31	29	25	51	44	3	4	5	40.8	6.7	85	24	26	25	43	31
1	1	4	28	35	160	161	4	2	2	4	4	1	1	37.5	7.2	89	22	18	17	61	48
3	3	4	81	5.8	85	27	28	27	35	35	1	2	1	88.4	8.1	88	21	18	17	61	48
2	4	4	17	7.2	84	29	28	27	35	35	2	3	1	63.4	7.6	90	22	18	17	61	48
1	1	5	26	39	160	161	4	2	2	5	3	4	1	-72	7.4	85	22	18	17	61	48
3	3	5	99	5.8	85	20	26	26	41	39	4	1	5	-83	7.2	89	26	29	27	48	45
2	4	5	31	7.2	84	20	26	26	41	39	1	2	5	95.1	8.1	88	26	30	28	49	46
1	1	1	20	50	160	161	4	2	2	1	2	3	5	72.8	7.6	90	26	31	29	50	47
3	3	1	-45	6.4	85	20	20	20	54	50	3	4	5	12.2	7.4	85	26	32	30	51	48
2	4	1	108	7.4	84	22	20	20	54	50	4	1	1	27	6.9	88	21	23	22	63	58
1	1	5	27	47	160	161	4	2	2	5	1	2	1	23.2	8.1	90	22	23	22	63	58
3	3	5	2.3	6.4	85	28	27	26	50	47	2	3	1	-0.7	7.4	88	22	23	22	63	58
2	4	5	39	7.4	84	28	27	26	50	47	3	4	1	61.1	7.3	86	24	23	22	63	58
1	1	1	24	53	160	161	4	2	2	1	4	1	5	73.8	6.9	88	19	18	18	76	65
3	3	1	13	6.4	85	20	24	22	54	53	1	2	5	62.1	8.1	90	20	18	18	76	65
2	4	1	148	7.1	84	20	24	22	54	53	2	3	5	-96	7.4	88	20	18	18	76	65
1	1	5	24	57	160	161	4	2	2	5	3	4	5	40.2	7.3	86	20	18	18	76	65
3	3	5	9	6.4	85	28	24	23	60	57	4	1	1	7.22	7.3	90	17	17	17	73	67
2	4	5	13	7.1	84	28	24	23	60	57	1	2	1	5.53	7.9	88	17	17	17	73	67
1	1	1	22	69	160	161	4	2	2	1	2	3	1	74.5	7.4	86	18	17	17	73	67
3	3	1	-36	6.5	83	20	22	21	73	69	3	4	1	-23	7.4	82	19	17	17	73	67
2	4	1	45	6.8	84	20	22	21	73	69	4	1	5	90.3	7.3	90	20	21	20	70	69
1	1	5	23	70	160	161	4	2	2	5	1	2	5	-21	7.9	88	20	21	20	70	69
3	3	5	35	6.5	83	22	23	22	71	70	2	3	5	63.9	7.4	86	20	21	20	70	69
2	4	5	-17	6.8	84	22	23	22	71	70	3	4	5	-16	7.4	82	20	21	20	70	69

Data: Percentage inhibition (*Bio* bioassay, *Con* concentration, *Rep* repetition, *Por* percentage inhibition)

Bio	Day	Con	Rep	Por	Bio	Day	Con	Rep	Por
1	1	0	3	5.2632	1	6	2000	4	35.1724
1	1	0	4	5.2632	2	1	0	2	0.0016
1	1	500	1	15.7895	2	1	0	3	14.2857
1	1	500	2	26.3158	2	1	500	1	42.8571
1	1	500	3	15.7895	2	1	500	2	42.8571
1	1	500	4	15.7895	2	1	500	3	42.8571
1	1	1000	1	36.8421	2	1	500	4	42.8571
1	1	1000	2	36.8421	2	1	1000	1	7.1429
1	1	1000	3	36.8421	2	1	1000	2	42.8571
1	1	1000	4	36.8421	2	1	1000	3	42.8571
1	1	2000	1	15.7895	2	1	1000	4	42.8571
1	1	2000	2	36.8421	2	2	0	1	1.3699
1	1	2000	3	36.8421	2	2	0	2	1.3699
1	1	2000	4	36.8421	2	2	0	4	1.3699
1	2	0	2	1.9355	2	2	500	1	34.2466
1	2	0	3	4.5161	2	2	500	2	31.5068
1	2	0	4	1.9355	2	2	500	3	42.4658
1	2	500	1	43.2258	2	2	500	4	36.9863
1	2	500	2	48.3871	2	2	1000	1	34.2466
1	2	500	3	40.6452	2	2	1000	2	47.9452
1	2	500	4	40.6452	2	2	1000	3	45.2055
1	2	1000	1	35.4839	2	2	1000	4	45.2055
1	2	1000	2	45.8065	2	2	2000	1	47.9452
1	2	1000	3	43.2258	2	2	2000	2	53.4247
1	2	1000	4	32.9032	2	2	2000	3	50.6849
1	2	2000	1	58.7097	2	2	2000	4	56.1644
1	2	2000	2	53.5484	2	3	0	1	4.2735
1	2	2000	3	53.5484	2	3	0	4	14.5299
1	2	2000	4	58.7097	2	3	500	1	28.2051
1	3	0	2	1.2346	2	3	500	2	28.2051
1	3	0	3	3.7037	2	3	500	3	35.0427
1	3	500	1	25.9259	2	3	500	4	24.7863
1	3	500	2	23.4568	2	3	1000	1	24.7863
1	3	500	3	23.4568	2	3	1000	2	35.0427
1	3	500	4	24.6914	2	3	1000	3	24.7863
1	3	1000	1	30.8642	2	3	1000	4	26.4957
1	3	1000	2	32.0988	2	3	2000	1	40.1709
1	3	1000	3	28.3951	2	3	2000	2	38.4615
1	3	1000	4	25.9259	2	3	2000	3	47.0085
1	3	2000	1	53.0864	2	3	2000	4	41.8803
1	3	2000	2	49.3827	2	4	0	2	1.5015
1	3	2000	3	49.3827	2	4	0	3	1.5015

(continued)

Data: Percentage inhibition (*Bio* bioassay, *Con* concentration, *Rep* repetition, *Por* percentage inhibition)

Bio	Day	Con	Rep	Por	Bio	Day	Con	Rep	Por
1	3	2000	4	51.8519	2	4	0	4	1.5015
1	4	0	3	4.6729	2	4	500	1	20.7207
1	4	500	1	19.6262	2	4	500	2	23.1231
1	4	500	2	20.5607	2	4	500	3	27.9279
1	4	500	3	22.4299	2	4	500	4	20.7207
1	4	500	4	20.5607	2	4	1000	1	35.1351
1	4	1000	1	21.4953	2	4	1000	2	26.7267
1	4	1000	2	21.4953	2	4	1000	3	26.7267
1	4	1000	3	23.3645	2	4	1000	4	32.7327
1	4	1000	4	20.5607	2	4	2000	1	33.9339
1	4	2000	1	42.0561	2	4	2000	2	37.5375
1	4	2000	2	36.4486	2	4	2000	3	44.7447
1	4	2000	3	32.7103	2	4	2000	4	38.7387
1	4	2000	4	40.1869	2	5	0	2	2.008
1	5	0	3	4.065	2	5	0	4	0.4016
1	5	0	4	4.065	2	5	500	1	13.253
1	5	500	1	21.1382	2	5	500	2	21.2851
1	5	500	2	24.3902	2	5	500	3	21.2851
1	5	500	3	17.0732	2	5	500	4	18.0723
1	5	500	4	17.0732	2	5	1000	1	21.2851
1	5	1000	1	18.6992	2	5	1000	2	18.0723
1	5	1000	2	18.6992	2	5	1000	3	16.4659
1	5	1000	3	20.3252	2	5	1000	4	16.4659
1	5	1000	4	17.8862	2	5	2000	1	35.743
1	5	2000	1	41.4634	2	5	2000	2	34.1365
1	5	2000	2	38.2114	2	5	2000	3	29.3173
1	5	2000	3	34.1463	2	5	2000	4	30.9237
1	5	2000	4	33.3333	2	6	0	2	4.2159
1	6	0	3	4.8276	2	6	0	4	0.1686
1	6	0	4	2.069	2	6	500	1	18.3811
1	6	500	1	17.2414	2	6	500	2	20.4047
1	6	500	2	18.6207	2	6	500	3	22.4283
1	6	500	3	16.5517	2	6	500	4	20.4047
1	6	500	4	13.7931	2	6	1000	1	21.0793
1	6	1000	1	15.8621	2	6	1000	2	17.7066
1	6	1000	2	16.5517	2	6	1000	3	17.7066
1	6	1000	3	15.8621	2	6	1000	4	20.4047
1	6	1000	4	18.6207	2	6	2000	1	31.1973
1	6	2000	1	32.4138	2	6	2000	2	29.1737
1	6	2000	2	29.6552	2	6	2000	3	29.8482
1	6	2000	3	31.7241	2	6	2000	4	30.5228

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



References

- Agresti A (2013) Introduction to categorical data analysis, 3rd edn. Wiley, Hoboken
- Aitchison J, Silvey S (1957) The generalization of probit analysis to the case of multiple responses. *Biometrika* 44:131–140
- Akaike H (1973) Information theory and an extension of the maximum likelihood principle. In: Petrov BN, Casake F (eds) Second international symposium on information theory. Akademiai kiado, Budapest, pp 267–281
- Akaike H (1974) A new look at the statistical model identification. *IEEE Trans Automat Control* 19: 716–723
- Amoah S, Wilkinson M, Dunwell J, King GJ (2008) Understanding the relationship between DNA methylation and phenotypic plasticity in crop plants. *Comp Biochem Physiol* 150(Suppl 1): S145
- Bekele A, Bultosa G, Belete K (2012) The effect of germination time on malt quality of six sorghum (sorghum bicolor) varieties grown at Melkassa, Ethiopia. *J Brew* 118(1):76–81
- Bilgili S, Hess J, Blake J, Macklin K, Saenmahayak B, Sibley J (2009) Influence of bedding material on footpad dermatitis in broiler chickens. *J Appl Poult Res* 18(3):583–589
- Bliss CL (1934) Methods of probits. *Science* 79:38–39
- Bliss CI (1935) The calculation of the dosage-mortality curve. *Ann Appl Biol* 22(1):134–167
- Breslow NE (2004) Whither PQL? In: Lin DY, Heagerty PJ (eds) Proceedings of the second Seattle symposium in biostatistics: analysis of correlated data. Springer, pp 1–22
- Breslow NE, Clayton DG (1993) Approximate inference in generalized linear mixed models. *J Am Stat Assoc* 88:9–25
- Casella G, Berger RL (2002) Statistical inference, 2nd edn. Duxbury, Pacific Grove
- Collett D (2002) Modelling binary data, 2nd edn. Chapman & Hall/CRC Press, Boca Raton. 387 pp.
- De Jong IC, Guémené D (2012) Major welfare issues in broiler breeders. *Worlds Poult Sci J* 67:73–82
- Engel B, te Brake J (1993) Analysis of embryonic development with a model for under-or overdispersion relative binomial variation. *Biometrics* 49:269–279
- Fisher RA (1925) Statistical methods for research workers. Oliver and Boyd, Edinburgh
- Garcia RG, Almeida PICL, Caldara FR, Naas IA, Pereira DF et al (2010) Effect of litter material on water quality in broiler production. *Braz J Poult Sci* 12:165–169
- Gbur EE, Stroup WW, McCarter KS, Durham S, Young LJ, Christman M, West M, Kramer M (2012) Analysis of generalized linear mixed models in the agricultural and natural resources sciences. ASA, CSSA, SSSA, Madison
- Gilks WR et al (1996) Introducing Markov chain Monte Carlo. In: Gilks WR (ed) Markov chain Monte Carlo in practice. Chapman and Hall, pp 1–19

- Goldstein H, Rasbash J (1996) Improved approximations for multilevel models with binary responses. *J R Stat Soc Ser A Stat Soc* 159:505–513
- Hand DJ, Daly F, Lunn AD, McConway KJ, Ostrowski E (1994) *A handbook of small data sets*. Chapman and Hall, London
- Heindel J, Price C, Field E, Marr M, Myers C, Morrissey R, Schwetz B (1992) Developmental toxicity of boric acid in mice and rats. *Toxicol Sci* 18(2):266–277
- Henderson CR (1950) Estimation of genetic parameters. *Ann Math Stat* 21:309–310
- Henderson CR (1984) Applications of linear models in animal breeding. In *Univ. of Guelph, Guelph*
- Hosmer DW, Lemeshow S (2000) *Applied logistic regression*, 2nd edn. Wiley, New York
- Immer RF, Hayes HK, Powers LR (1934) Statistical determination of barley varietal adaptation. *J Am Soc Agron* 26:403–419
- Jermann R, Toumiat M, Imfeld D (2001) Development of an in vitro efficacy test for self-tanning formulations. *Int J Cosmet Sci* 24(1):35–42
- Johnson NL, Kotz S, Balakrishnan N (1995) *Continuous univariate distributions*, vol 2. Wiley, New York
- Kaplan EL, Meier P (1958) Nonparametric estimation from incomplete observations. *J Am Stat Assoc* 53(282):457–481
- Lee Y, Nelder JA (2001) Hierarchical generalised linear models: a synthesis of generalised linear models, random-effect models and structured dispersions. *Biometrika* 88:987–1006
- Lee Y, Nelder JA (2004) Conditional and marginal models: another view. *Stat Inference* 19(2): 219–238
- Lew M (2007) Good statistical practice in pharmacology. *Br J Pharmacol* 152(3):299–303
- Littell RC, Milliken GA, Stroup WW, Wolfinger RD (1996) *SAS for mixed models*. SAS Institute, Inc., Cary
- Littell RC et al (2006) *SAS for Mixed Models*, 2nd edn. SAS Publishing
- Limpert E, Stahel WA, Abbt M (2001) Log-normal Distributions across the Sciences: Keys and Clues. *BioScience* 51(5)
- Logan M (2010) *Biostatistical design and analysis using R: a practical guide*. Wiley
- Madden L, Hughes G (1995) Plant disease incidence: distributions, heterogeneity, and temporal analysis. *Annu Rev Phytopathol* 33:529–564
- Margolin BH, Kaplan N, Zeiger E (1981) Statistical analysis of the Ames Salmonella/microsome test. *Proc Natl Acad Sci USA* 76:3779–3783
- Martrenchar A, Boilletot E, Huonnic D, Pol F (2002) Risk factors for foot-pad dermatitis in chicken and Turkey broilers in France. *Prev Vet Med* 52(3–4):213–226
- McCullagh P (1980) Regression models for ordinal data. *J R Stat Soc Series B Methodol* 42:109–142
- McCullagh P (1983) Quasi-likelihood functions. *Ann Stat* 11(1):59–67
- McCullagh P, Nelder JA (1989) *Generalized linear models*, 2nd edn. Chapman and Hall, London
- Mead J, Curnow R, Hasted A (1993) *Statistical methods in agriculture and experimental biology*, 2nd edn. Chapman and Hall, London, p 325
- Mosteller F, Tukey JW (1977) *Data analysis and regression*. Addison-Wesley, Reading
- Myers R, Montgomery D, Vining G (2002) *Generalized linear models with applications in engineering and the sciences*. Wiley, New York
- Nadia Hernández-Tapia, Josafhat Salinas-Ruiz, Vinisa Saynes-Santillán, Julio M. Ayala-Rodríguez, Francisco Hernández-Rosas y Joel Velasco-Velasco (2019). N₂O, CO₂ and NH₃ emission from dung of bovine with different percentage of crude protein in diet. *Rev. Int. Contam. Ambie* 35 (3) 597–608. DOI: 10.20937/RICA.2019.35.03.07
- Nelder JA, Wedderburn RWM (1972) Generalized linear models. *J R Stat Soc Ser A* 135:370–384
- Pinheiro JC, Bates DM (2000) *Mixed-effects models in S and SPLUS*. Springer
- Pinheiro JC, Chao EC (2006) Efficient Laplacian and adaptive Gaussian quadrature algorithms for multilevel generalized linear mixed models. *J Comput Graph Stat* 15:58–81

- Quinn GP, Keough MJ (2002) *Experimental design and data analysis for biologists*. Cambridge University Press, New York
- Raudenbush SW et al (2000) Maximum likelihood for generalized linear models with nested random effects via high-order, multivariate Laplace approximation. *J Comput Graph Stat* 9: 141–157
- Robinson GK (1991) That BLUP is a good thing: The estimation of random effects (with discussion), *Statist Sci* 6:15–51
- Rodriguez G, Goldman N (2001) Improved estimation procedures for multilevel models with binary response: a case-study. *J R Stat Soc Ser A Stat Soc* 164:339–355
- Schall R (1991) Estimation in generalized linear models with random effects. *Biometrika* 78:719–727
- Schwarz G (1978) Estimating the dimension of a model. *Ann Stat* 6:461–464
- Shinozaki K, Kira T (1956) Intraspecific competition among higher plants. VII. Logistic theory of the C-D effect. *J Inst Polytech* 12:69–82
- Smith T, Mlambo V, Sikosana JLN, Maphosa V, Mueller-Harvey I, Owen E (2005) *Dichrostachys cinerea* and *Acacia nilotica* fruits as dry season feed supplements for goats in a semi-arid environment. *Anim Feed Sci Technol* 122(1–2):149–157
- Spurgeon J (1978) The correlation of animal response data with the yields of selected thermal decomposition products for typical aircraft interior materials. U.S. D.O.T. Report No. FAA-RD-78-131
- Stanley VG (1981) The effect of stocking density on commercial broiler performance. *Poult Sci* 60: 1737–1738
- Stephens PA et al (2005) Information theory and hypothesis testing: a call for pluralism. *J Appl Ecol* 42:4–12
- Stroup W (2012) *Generalized linear mixed models*, 1st edn. Chapman & Hall/CRC
- Stroup W (2013) *Generalized linear mixed models*. CRC Press, Boca Raton
- Taira K, Nagai T, Obi T, Takase K (2014) Effect of litter moisture on the development of footpad dermatitis in broiler chickens. *J Vet Med Sci* 76:583–586
- Wagner JG, Aghajanian GK, Bing OHL (1968) Correlation of performance test scores with "tissue concentration" of lysergic acid diethylamide in human subjects. *Clin Pharmacol Ther* 9(5): 635–638
- Wallsten TS, Budescu DV (1981) Adaptivity and nonadditivity in judging MMPI profiles. *J Exp Psychol Hum Percept Perform* 7:1096–1109
- Walters KJ, Hosfield GL, Uebersax MA, Kelly JD (1997) Navy bean canning quality: correlations, heritability estimates and randomly amplified polymorphic DNA markers associated with component traits. *J Am Soc Hortic Sci* 122(3):338–343
- Wolfinger R, O'Connell M (1993) Generalized linear mixed models: a pseudo-likelihood approach. *J Stat Comput Simul* 48:233–243
- Yates F (1935) Complex experiments, with discussion. *J R Stat Soc Ser B* 2:181–223
- Zeger SL, Liang K-Y, Albert PS (1988) Models for longitudinal data: a generalized estimating equation approach. *Biometrics* 44(4):1049
- Zuur AF, Leno EN, Walker N, Saveliev AA, Smith GM (2009) *Mixed effects models and extensions in ecology with R*. Springer, New York
- Zuur AF, Hilbe JM, Leno EN (2013) *A Beginner's guide to GLM and GLMM with R: a frequentist. and Bayesian perspective for ecologists*. Highland Statistics Ltd., Newburgh