

# Analysis of Microdata

Rainer Winkelmann  
Stefan Boes

# Analysis of Microdata

With 38 Figures  
and 41 Tables

 Springer

Professor Dr. Rainer Winkelmann  
Dipl. Vw. Stefan Boes  
University of Zurich  
Socioeconomic Institute  
Zürichbergstrasse 14  
8032 Zurich  
Switzerland  
E-mail: winkelmann@sts.unizh.ch  
E-mail: boes@sts.unizh.ch

Cataloging-in-Publication Data

Library of Congress Control Number: 2005935030

ISBN-10 3-540-29605-0 Springer Berlin Heidelberg New York  
ISBN-13 978-3-540-29605-8 Springer Berlin Heidelberg New York

This work is subject to copyright. All rights are reserved, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilm or in any other way, and storage in data banks. Duplication of this publication or parts thereof is permitted only under the provisions of the German Copyright Law of September 9, 1965, in its current version, and permission for use must always be obtained from Springer-Verlag. Violations are liable for prosecution under the German Copyright Law.

Springer is a part of Springer Science+Business Media  
springeronline.com

© Springer-Verlag Berlin Heidelberg 2006  
Printed in Germany

The use of general descriptive names, registered names, trademarks, 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.

Cover design: Erich Kirchner  
Production: Helmut Petri  
Printing: Strauss Offsetdruck

SPIN 11573999 Printed on acid-free paper – 42/3153 – 5 4 3 2 1 0

---

## Preface

The availability of microdata has increased rapidly over the last decades, and standard statistical and econometric software packages for data analysis include ever more sophisticated modeling options. The goal of this book is to familiarize readers with a wide range of commonly used models, and thereby to enable them to become critical consumers of current empirical research, and to conduct their own empirical analyses.

The focus of the book is on regression-type models in the context of large cross-section samples. In microdata applications, dependent variables often are qualitative and discrete, while in other cases, the sample is not randomly drawn from the population of interest and the dependent variable is censored or truncated. Hence, models and methods are required that go beyond the standard linear regression model and ordinary least squares. Maximum likelihood estimation of conditional probability models and marginal probability effects are introduced here as the unifying principle for modeling, estimating and interpreting microdata relationships. We consider the limitation to maximum likelihood sensible, from a pedagogical point of view if the book is to be used in a semester-long advanced undergraduate or graduate course, and from a practical point of view because maximum likelihood estimation is used in the overwhelming majority of current microdata research.

In order to introduce and explain the models and methods, we refer to a number of illustrative applications. The main examples include the determinants of individual fertility, the intergenerational transmission of secondary school choices, and the wage elasticity of female labor supply. The models presented, while chosen with economic applications in mind, should be equally relevant for other social sciences, for example, quantitative political science and sociology, and for empirical disciplines outside of the social sciences.

The book can be used as a textbook for an advanced undergraduate, a Master's or a first-year Ph.D. course on the topic of microdata analysis. In economics and related disciplines, such a course is typically offered after a first course on linear regression analysis. Alternatively, the book can also serve as a supplementary text to an applied microeconomics field course, such as

those offered in the areas of labor economics, health economics, and the like. Finally, it is intended as a reference for graduate students, researchers as well as practitioners who encounter microdata in their work. The mathematical prerequisites are not very high. In particular, the use of linear algebra is minimal. On the other hand, some background in mathematical statistics is useful although not absolutely necessary.

The book includes numerous exercises. Most of the exercises do not require the use of a computer. Rather, they typically present specific empirical results, and the task is to assess the validity of the procedure in that particular context and to provide a correct interpretation of the estimated parameters. In addition, we encourage the reader to develop practical skills in applied data analysis by re-estimating the examples we discuss, using a software of choice. For this purpose, we have made the datasets employed available at our homepage [www.unizh.ch/sts/](http://www.unizh.ch/sts/), both in ASCII format and in Stata 7 format.

An earlier version of the manuscript was used in a course of the same name taught by us for several years at the economics department of the University of Zurich. We thank the participants for numerous suggestions for improvement. We are heavily indebted to Markus Lipp and Adrian Bruhin for careful proof-reading, to Markus in addition for creating all the figures, and to Deborah Bowen for improving our English.

Zurich, September 2005

*Rainer Winkelmann*  
*Stefan Boes*

---

# Contents

<b>1</b>	<b>Introduction</b> . . . . .	1
1.1	What Are Microdata? . . . . .	1
1.2	Types of Microdata . . . . .	4
1.2.1	Qualitative Data . . . . .	4
1.2.2	Quantitative Data . . . . .	6
1.3	Why Not Linear Regression? . . . . .	8
1.4	Common Elements of Microdata Models . . . . .	10
1.5	Examples . . . . .	11
1.5.1	Determinants of Fertility . . . . .	11
1.5.2	Secondary School Choice . . . . .	16
1.5.3	Female Hours of Work and Wages . . . . .	17
1.6	Overview of the Book . . . . .	19
<b>2</b>	<b>From Regression to Probability Models</b> . . . . .	21
2.1	Introduction . . . . .	21
2.2	Conditional Probability Functions . . . . .	23
2.2.1	Definition . . . . .	23
2.2.2	Estimation . . . . .	24
2.2.3	Interpretation . . . . .	25
2.3	Probability and Probability Distributions . . . . .	29
2.3.1	Axioms of Probability . . . . .	29
2.3.2	Univariate Random Variables . . . . .	30
2.3.3	Multivariate Random Variables . . . . .	31
2.3.4	Conditional Probability Models . . . . .	34
2.4	Further Exercises . . . . .	39
<b>3</b>	<b>Maximum Likelihood Estimation</b> . . . . .	45
3.1	Introduction . . . . .	45
3.2	Likelihood Function . . . . .	46
3.2.1	Score Function and Hessian Matrix . . . . .	48
3.2.2	Conditional Models . . . . .	50

3.2.3	Maximization . . . . .	50
3.3	Properties of the Maximum Likelihood Estimator . . . . .	53
3.3.1	Expected Score . . . . .	54
3.3.2	Consistency . . . . .	55
3.3.3	Information Matrix Equality . . . . .	56
3.3.4	Asymptotic Distribution . . . . .	59
3.3.5	Covariance Matrix . . . . .	60
3.4	Normal Linear Model . . . . .	63
3.5	Further Aspects of Maximum Likelihood Estimation . . . . .	67
3.5.1	Invariance and Delta Method . . . . .	67
3.5.2	Numerical Optimization . . . . .	69
3.5.3	Identification . . . . .	74
3.5.4	Quasi Maximum Likelihood . . . . .	76
3.6	Testing . . . . .	76
3.6.1	Introduction . . . . .	76
3.6.2	Restricted Maximum Likelihood . . . . .	79
3.6.3	Wald Test . . . . .	81
3.6.4	Likelihood Ratio Test . . . . .	83
3.6.5	Score Test . . . . .	86
3.6.6	Model Selection . . . . .	88
3.6.7	Goodness-of-Fit . . . . .	89
3.7	Pros and Cons of Maximum Likelihood . . . . .	89
3.8	Further Exercises . . . . .	90
<b>4</b>	<b>Binary Response Models . . . . .</b>	<b>95</b>
4.1	Introduction . . . . .	95
4.2	Models for Binary Response Variables . . . . .	97
4.2.1	General Framework . . . . .	97
4.2.2	Linear Probability Model . . . . .	98
4.2.3	Probit Model . . . . .	100
4.2.4	Logit Model . . . . .	102
4.2.5	Interpretation of Parameters . . . . .	104
4.3	Discrete Choice Models . . . . .	107
4.4	Estimation . . . . .	110
4.4.1	Maximum Likelihood . . . . .	110
4.4.2	Perfect Prediction . . . . .	113
4.4.3	Properties of the Estimator . . . . .	114
4.4.4	Endogenous Regressors in Binary Response Models . . . . .	116
4.4.5	Estimation of Marginal Effects . . . . .	118
4.5	Goodness-of-Fit . . . . .	122
4.6	Non-Standard Sampling Schemes . . . . .	127
4.6.1	Stratified Sampling . . . . .	127
4.6.2	Exogenous Stratification . . . . .	127
4.6.3	Endogenous Stratification . . . . .	128
4.7	Further Exercises . . . . .	130

<b>5</b>	<b>Multinomial Response Models</b> . . . . .	137
5.1	Introduction . . . . .	137
5.2	Multinomial Logit Model . . . . .	139
5.2.1	Basic Model . . . . .	139
5.2.2	Estimation . . . . .	140
5.2.3	Interpretation of Parameters . . . . .	144
5.3	Conditional Logit Model . . . . .	150
5.3.1	Introduction . . . . .	150
5.3.2	General Model of Choice . . . . .	151
5.3.3	Modeling Conditional Logits . . . . .	152
5.3.4	Interpretation of Parameters . . . . .	155
5.3.5	Independence of Irrelevant Alternatives . . . . .	159
5.4	Generalized Multinomial Response Models . . . . .	160
5.4.1	Multinomial Probit Model . . . . .	161
5.4.2	Mixed Logit Models . . . . .	163
5.4.3	Nested Logit Models . . . . .	164
5.5	Further Exercises . . . . .	166
<b>6</b>	<b>Ordered Response Models</b> . . . . .	171
6.1	Introduction . . . . .	171
6.2	Standard Ordered Response Models . . . . .	174
6.2.1	General Framework . . . . .	174
6.2.2	Ordered Probit Model . . . . .	176
6.2.3	Ordered Logit Model . . . . .	177
6.2.4	Estimation . . . . .	179
6.2.5	Interpretation of Parameters . . . . .	179
6.2.6	Single Indices and Parallel Regression . . . . .	186
6.3	Generalized Threshold Models . . . . .	188
6.3.1	Generalized Ordered Logit and Probit Models . . . . .	188
6.3.2	Interpretation of Parameters . . . . .	189
6.4	Sequential Models . . . . .	194
6.4.1	Modeling Conditional Transitions . . . . .	194
6.4.2	Generalized Conditional Transition Probabilities . . . . .	197
6.4.3	Marginal Effects . . . . .	197
6.4.4	Estimation . . . . .	198
6.5	Interval Data . . . . .	200
6.6	Further Exercises . . . . .	202
<b>7</b>	<b>Limited Dependent Variables</b> . . . . .	207
7.1	Introduction . . . . .	207
7.1.1	Corner Solution Outcomes . . . . .	208
7.1.2	Sample Selection Models . . . . .	209
7.1.3	Treatment Effect Models . . . . .	210
7.2	Tobin's Corner Solution Model . . . . .	211
7.2.1	Introduction . . . . .	211



7.2.2	Tobit Model	212
7.2.3	Truncated Normal Distribution	214
7.2.4	Inverse Mills Ratio and its Properties	215
7.2.5	Interpretation of the Tobit Model	218
7.2.6	Comparing Tobit and OLS	221
7.2.7	Further Specification Issues	223
7.3	Sample Selection Models	224
7.3.1	Introduction	224
7.3.2	Censored Regression Model	226
7.3.3	Estimation of the Censored Regression Model	228
7.3.4	Truncated Regression Model	230
7.3.5	Incidental Censoring	231
7.3.6	Example: Estimating a Labor Supply Model	237
7.4	Treatment Effect Models	239
7.4.1	Introduction	239
7.4.2	Endogenous Binary Variable	242
7.4.3	Switching Regression Model	243
7.5	Appendix: Bivariate Normal Distribution	246
7.6	Further Exercises	247
<b>8</b>	<b>Event History Models</b>	<b>251</b>
8.1	Introduction	251
8.2	Duration Models	254
8.2.1	Introduction	254
8.2.2	Basic Concepts	254
8.2.3	Discrete Time Duration Models	259
8.2.4	Continuous Time Duration Models	262
8.2.5	Key Element: Hazard Function	265
8.2.6	Duration Dependence	267
8.2.7	Unobserved Heterogeneity	271
8.3	Count Data Models	279
8.3.1	The Poisson Regression Model	279
8.3.2	Unobserved Heterogeneity	284
8.3.3	Efficient versus Robust Estimation	289
8.3.4	Censoring and Truncation	289
8.3.5	Hurdle and Zero-Inflated Count Data Models	291
8.4	Further Exercises	294
	<b>List of Figures</b>	<b>297</b>
	<b>List of Tables</b>	<b>299</b>
	<b>References</b>	<b>301</b>
	<b>Index</b>	<b>309</b>