Chapter 5
Generalized Linear Models

Most of the theory in the previous chapters has been based on the assumption of having similarity (or homogeneity) between the different observations. This was expressed by making an i.i.d. assumption on the observations, see, e.g., Sect. 3.3.2. In many practical applications such a homogeneity assumption is not reasonable, one may for example think of car insurance pricing where different car drivers have different driving experience and they drive different cars, or of health insurance where policyholders may have different genders and ages. Figure 5.1 shows a health insurance example where the claim sizes depend on the gender and the age of the policyholders. The most popular statistical models that are able to cope with such heterogeneous data are the *generalized linear models* (GLMs). The notion of GLMs has been introduced in the seminal work of Nelder–Wedderburn [283] in 1972. Their work has introduced a unified procedure for modeling and fitting distributions within the EDF to data having systematic differences (effects) that can be described by explanatory variables. Today, GLMs are the state-of-the-art statistical models in many applied fields including statistics, actuarial science and economics. However, the specific use of GLMs in the different fields may substantially differ. In fields like actuarial science these models are mainly used for *predictive* modeling, in other fields like economics or social sciences GLMs have become the main tool in exploring and *explaining* (hopefully) causal relations. For a discussion on “predicting” versus “explaining” we refer to Shmueli [338].

It is difficult to give a good list of references for GLMs, since GLMs and their offsprings are present in almost every statistical modeling publication and in every lecture on statistics. Classical statistical references are the books of McCullagh–Nelder [265], Fahrmeir–Tutz [123] and Dobson [107], in the actuarial literature we mention the textbooks (in alphabetical order) of Charpentier [67], De Jong–Heller [89], Denuit et al. [99–101], Frees [134] and Ohlsson–Johansson [290], but this list is far from being complete.

M. V. Wüthrich, M. Merz, *Statistical Foundations of Actuarial Learning and its Applications*, Springer Actuarial, https://doi.org/10.1007/978-3-031-12409-9_5

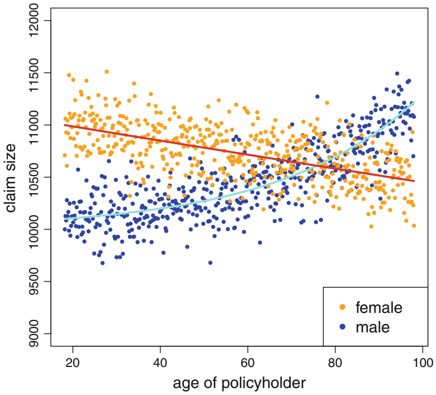


Fig. 5.1 Claim sizes in health insurance as a function of the age of the policyholder, and split by gender

In this chapter we introduce and discuss GLMs in the context of actuarial modeling. We do this in such a way that GLMs can be seen as a building block of network regression models which will be the main topic of Chap. 7 on deep learning.

5.1 Generalized Linear Models and Log-Likelihoods

5.1.1 Regression Modeling

We start by assuming of having independent random variables $Y_1, \ldots, Y_n$ which are described by a fixed member of the EDF. That is, we assume that all Y_i are independent and have densities w.r.t. a σ-finite measure ν on $\mathbb{R}$ given by

$$Y_i \sim f(y_i; \theta_i, v_i/\varphi) = \exp\left\{ \frac{y_i\theta_i - \kappa(\theta_i)}{\varphi/v_i} + a(y_i; v_i/\varphi) \right\} \qquad \text{for } 1 \le i \le n, \tag{5.1}$$

with canonical parameters $\theta_i \in \mathring{\boldsymbol{\Theta}}$, exposures $v_i > 0$ and dispersion parameter $\varphi > 0$. Throughout, we assume that the effective domain $\boldsymbol{\Theta}$ has a non-empty interior. There is a fundamental difference between (5.1) and Example 3.5. We now allow every random variable Y_i to have its own canonical parameter $\theta_i \in \mathring{\boldsymbol{\Theta}}$. We call this a *heterogeneous* situation because the observations are allowed to differ in a systematic way expressed by different canonical parameters. This is highlighted by the lines in the health insurance example of Fig. 5.1 where (expected) claim sizes differ by gender and age of policyholder.

In Sect. 4.1.2 we have introduced the *saturated model* where every observation Y_i has its own parameter θ_i. In general, if we have n observations $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^\top$ we can estimate at most n parameters. The other extreme case is the homogeneous one, meaning that $\theta_i = \theta \in \mathring{\boldsymbol{\Theta}}$ for all $1 \le i \le n$. In this latter case we have exactly one parameter to estimate, and we call this model *null model*, *intercept model* or *homogeneous model*, because all components of $\boldsymbol{Y}$ are assumed to follow the same law expressed in a single common parameter θ. Both the saturated model and the null model may behave very poorly in predicting new observations. Typically, the saturated model fully reflects the data $\boldsymbol{Y}$ including the noisy part (random component, irreducible risk, see Remarks 4.2) and, therefore, it is not useful for prediction. We also say that this model (in-sample) over-fits to the data $\boldsymbol{Y}$ and does not generalize (out-of-sample) to new data. The null model often has a poor predictive performance because if the data has systematic effects these cannot be captured by a null model. GLMs try to find a good balance between these two extreme cases, by trying to extract (only) the systematic effects from noisy data $\boldsymbol{Y}$. We therefore model the canonical parameters θ_i as a low-dimensional function of *explanatory variables* which capture the systematic effects in the data. In Fig. 5.1 gender and age of policyholder play the role of such explanatory variables.

Assume that each observation Y_i is equipped with a *feature* (explanatory variable, covariate) $\boldsymbol{x}_i$ that belongs to a fixed given *feature space* $\mathcal{X}$. These features $\boldsymbol{x}_i$ are assumed to describe the *systematic effects* in the observations Y_i, i.e., these features are assumed to be appropriate descriptions of the heterogeneity between the observations. In a nutshell, we then assume of having a suitable *regression function*

$$\theta : \mathcal{X} \to \mathring{\boldsymbol{\Theta}}, \qquad \boldsymbol{x} \mapsto \theta(\boldsymbol{x}),$$

such that we can appropriately describe the observations by

$$Y_i \overset{\text{ind.}}{\sim} f(y_i; \theta_i = \theta(\boldsymbol{x}_i), v_i/\varphi) = \exp\left\{ \frac{y_i\theta(\boldsymbol{x}_i) - \kappa(\theta(\boldsymbol{x}_i))}{\varphi/v_i} + a(y_i; v_i/\varphi) \right\}, \tag{5.2}$$

for $1 \le i \le n$. As a result we receive for the first moment of Y_i, see Corollary 2.14,

$$\mu_i = \mu(\boldsymbol{x}_i) = \mathbb{E}_{\theta(\boldsymbol{x}_i)}[Y_i] = \kappa'(\theta(\boldsymbol{x}_i)). \tag{5.3}$$

Thus, the regression function $\theta : \mathcal{X} \to \mathring{\boldsymbol{\Theta}}$ is assumed to describe the systematic differences (effects) between the random variables $Y_1, \ldots, Y_n$ being expressed by the means $\mu(\boldsymbol{x}_i)$ for features $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$. In GLMs this regression function takes a linear form after a suitable transformation, which exactly motivates the terminology *generalized linear model*.

5.1.2 Definition of Generalized Linear Models

We start with the discussion of the features $\boldsymbol{x} \in \mathcal{X}$. Features are also called explanatory variables, covariates, independent variables or regressors. Throughout, we assume that the features $\boldsymbol{x} = (x_0, x_1, \ldots, x_q)^\top$ include a first component $x_0 = 1$, and we choose feature space $\mathcal{X} \subset \{1\} \times \mathbb{R}^q$. The inclusion of this first component $x_0 = 1$ is useful in what follows. We call this first component *intercept* or *bias component* because it will be modeling an intercept of a regression model. The null model (homogeneous model) has features that only consist of this intercept component. For later purposes it will be useful to introduce the *design matrix* $\mathfrak{X}$ which collects the features $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \in \mathcal{X}$ of all responses $Y_1, \ldots, Y_n$. The design matrix is defined by

$$\mathfrak{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^\top = \begin{pmatrix} 1 & x_{1,1} & \cdots & x_{1,q} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n,1} & \cdots & x_{n,q} \end{pmatrix} \in \mathbb{R}^{n \times (q+1)}. \tag{5.4}$$

Based on these choices we assume existence of a *regression parameter* $\boldsymbol{\beta} \in \mathbb{R}^{q+1}$ and of a strictly monotone and smooth *link function* $g : \mathcal{M} \to \mathbb{R}$ such that we can express (5.3) by the following function (we drop index i)

$$\boldsymbol{x} \ \mapsto \ g(\mu(\boldsymbol{x})) = g\left(\mathbb{E}_{\theta(\boldsymbol{x})}\left[Y\right]\right) = \eta(\boldsymbol{x}) = \langle \boldsymbol{\beta}, \boldsymbol{x} \rangle = \beta_0 + \sum_{j=1}^{q} \beta_j x_j. \tag{5.5}$$

Here, $\langle \cdot, \cdot \rangle$ describes the scalar product in the Euclidean space $\mathbb{R}^{q+1}$, $\theta(\boldsymbol{x}) = h(\mu(\boldsymbol{x}))$ is the resulting canonical parameter (using canonical link $h = (\kappa')^{-1}$), and $\eta(\boldsymbol{x})$ is the so-called *linear predictor*. After applying a suitable link function g, the systematic effects of the random variable Y with features $\boldsymbol{x}$ can be described by a linear predictor $\eta(\boldsymbol{x}) = \langle \boldsymbol{\beta}, \boldsymbol{x} \rangle$, linear in the components of $\boldsymbol{x} \in \mathcal{X}$. This gives a particular functional form to (5.3), and the random variables $Y_1, \ldots, Y_n$ share a common regression parameter $\boldsymbol{\beta} \in \mathbb{R}^{q+1}$. Remark that the link function g used in (5.5) can be different from the canonical link h used to calculate $\theta(\boldsymbol{x}) = h(\mu(\boldsymbol{x}))$. We come back to this distinction below.

Summary of (5.5)

1. The independent random variables Y_i follow a fixed member of the EDF (5.1) with individual canonical parameters $\theta_i \in \mathring{\boldsymbol{\Theta}}$, for all $1 \le i \le n$.
2. The canonical parameters θ_i and the corresponding mean parameters μ_i are related by the canonical link $h = (\kappa')^{-1}$ as follows $h(\mu_i) = \theta_i$, where κ is the cumulant function of the chosen EDF, see Corollary 2.14.
3. We assume that the systematic effects in the random variables Y_i can be described by linear predictors $\eta_i = \eta(\boldsymbol{x}_i) = \langle \boldsymbol{\beta}, \boldsymbol{x}_i \rangle$ and a strictly monotone and smooth link function g such that we have $g(\mu_i) = \eta_i = \langle \boldsymbol{\beta}, \boldsymbol{x}_i \rangle$, for all $1 \le i \le n$, with common regression parameter $\boldsymbol{\beta} \in \mathbb{R}^{q+1}$.

We can either express this GLM regression structure in the dual (mean) parameter space $\mathcal{M}$ or in the effective domain $\mathring{\boldsymbol{\Theta}}$, see Remarks 2.9,

$$\boldsymbol{x} \mapsto \mu(\boldsymbol{x}) = g^{-1}(\eta(\boldsymbol{x})) = g^{-1}\langle \boldsymbol{\beta}, \boldsymbol{x} \rangle \ \in \ \mathcal{M} \qquad\qquad \text{or}$$
$$\boldsymbol{x} \mapsto \theta(\boldsymbol{x}) = (h \circ g^{-1})(\eta(\boldsymbol{x})) = (h \circ g^{-1})\langle \boldsymbol{\beta}, \boldsymbol{x} \rangle \ \in \ \mathring{\boldsymbol{\Theta}},$$

where $(h \circ g^{-1})$ is the composition of the inverse link g^{-1} and the canonical link h. For the moment, the link function g is quite general. In practice, the explicit choice needs some care. The right-hand side of (5.5) is defined on the whole real line if at least one component of $\boldsymbol{x}$ is both-sided unbounded. On the other hand, $\mathcal{M}$ and $\mathring{\boldsymbol{\Theta}}$ may be bounded sets. Therefore, the link function g may require some restrictions such that the domain and the range fulfill the necessary constraints. The dimension of $\boldsymbol{\beta}$ should satisfy $1 \leq 1 + q \leq n$, the lower bound will provide a null model and the upper bound a saturated model.

5.1.3 Link Functions and Feature Engineering

As link function we choose a strictly monotone and smooth function $g : \mathcal{M} \to \mathbb{R}$ such that we do not have any conflicts in domains and ranges. Beside these requirements, we may want further properties for the link function g and the features $\boldsymbol{x}$. From (5.5) we have

$$\mu(\boldsymbol{x}) = \mathbb{E}_{\theta(\boldsymbol{x})}[Y] = g^{-1}\langle \boldsymbol{\beta}, \boldsymbol{x} \rangle. \tag{5.6}$$

Of course, a basic requirement is that the selected features $\boldsymbol{x}$ can appropriately describe the mean of Y by the function in (5.6), see also Fig. 5.1. This may require so-called *feature engineering* of $\boldsymbol{x}$, for instance, we may want to replace the first component x_1 of the *raw features* $\boldsymbol{x}$ by, say, x_1^2 in the *pre-processed features*. For example, if this first component describes the age of the insurance policyholder, then, in some regression problems, it might be more appropriate to consider `age`^{2} instead of `age` to bring the predictive problem into structure (5.6). It may also be that we would like to enforce a certain type of *interaction* between the components of the raw features. For instance, we may include in a pre-processed feature a component x_1/x_2^2 which might correspond to `weight/height`^{2} if the policyholder has body weight x_1 and body height x_2. In fact, this pre-processed feature is exactly the body mass index of the policyholder. We will come back to feature engineering in Sect. 5.2.2, below.

Another important requirement is the ability of model interpretation. In insurance pricing problems, one often prefers additive and multiplicative effects in feature components. Choosing the identity link $g(m) = m$ we receive a model with additive effects

$$\mu(\boldsymbol{x}) = \mathbb{E}_{\theta(\boldsymbol{x})}[Y] = \langle \boldsymbol{\beta}, \boldsymbol{x} \rangle = \beta_0 + \sum_{j=1}^{q} \beta_j x_j,$$

and choosing the log-link $g(m) = \log(m)$ we receive a model with multiplicative effects

$$\mu(\boldsymbol{x}) = \mathbb{E}_{\theta(\boldsymbol{x})}[Y] = \exp\langle \boldsymbol{\beta}, \boldsymbol{x} \rangle = e^{\beta_0} \prod_{j=1}^{q} e^{\beta_j x_j}.$$

The latter is probably the most commonly used GLM in insurance pricing because it leads to explainable tariffs where feature values directly relate to price de- and increases in percentages of a base premium $\exp\{\beta_0\}$.

Another very popular choice is the canonical (natural) link, i.e., $g = h = (\kappa')^{-1}$. The canonical link substantially simplifies the analysis and it has very favorable statistical properties (as we will see below). However, in some applications practical needs overrule good statistical properties. Under the canonical link $g = h$ we have in the dual mean parameter space $\mathcal{M}$ and in the effective domain $\boldsymbol{\Theta}$, respectively,

$$\boldsymbol{x} \mapsto \mu(\boldsymbol{x}) = \kappa'(\eta(\boldsymbol{x})) = \kappa'\langle \boldsymbol{\beta}, \boldsymbol{x} \rangle \qquad \text{and} \qquad \boldsymbol{x} \mapsto \theta(\boldsymbol{x}) = \eta(\boldsymbol{x}) = \langle \boldsymbol{\beta}, \boldsymbol{x} \rangle.$$

Thus, the linear predictor η and the canonical parameter θ coincide under the canonical link choice $g = h = (\kappa')^{-1}$.

5.1.4 Log-Likelihood Function and Maximum Likelihood Estimation

After having a fully specified GLM within the EDF, there remains estimation of the regression parameter $\boldsymbol{\beta} \in \mathbb{R}^{q+1}$. This is done within the framework of MLE.

The log-likelihood function of $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^\top$ for regression parameter $\boldsymbol{\beta} \in \mathbb{R}^{q+1}$ is given by, see (5.2) and we use the independence between the Y_i's,

(continued)

$$\boldsymbol{\beta} \mapsto \ell_{\boldsymbol{Y}}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \frac{v_i}{\varphi}\Big[Y_i h(\mu(\boldsymbol{x}_i)) - \kappa\left(h(\mu(\boldsymbol{x}_i))\right)\Big] + a(Y_i; v_i/\varphi), \tag{5.7}$$

where we set $\mu(\boldsymbol{x}_i) = g^{-1}\langle\boldsymbol{\beta}, \boldsymbol{x}_i\rangle$. For the canonical link $g = h = (\kappa')^{-1}$ this simplifies to

$$\boldsymbol{\beta} \mapsto \ell_{\boldsymbol{Y}}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \frac{v_i}{\varphi}\Big[Y_i\langle\boldsymbol{\beta}, \boldsymbol{x}_i\rangle - \kappa\langle\boldsymbol{\beta}, \boldsymbol{x}_i\rangle\Big] + a(Y_i; v_i/\varphi). \tag{5.8}$$

MLE of $\boldsymbol{\beta}$ needs maximization of log-likelihoods (5.7) and (5.8), respectively; these are the GLM counterparts to the homogeneous case treated in Section 3.3.2. We calculate the score, we set $\eta_i = \langle\boldsymbol{\beta}, \boldsymbol{x}_i\rangle$ and $\mu_i = \mu(\boldsymbol{x}_i) = g^{-1}\langle\boldsymbol{\beta}, \boldsymbol{x}_i\rangle$,

$$\begin{aligned} s(\boldsymbol{\beta}, \boldsymbol{Y}) \;=\; \nabla_{\boldsymbol{\beta}}\ell_{\boldsymbol{Y}}(\boldsymbol{\beta}) &= \sum_{i=1}^{n} \frac{v_i}{\varphi}\left[Y_i - \mu_i\right] \nabla_{\boldsymbol{\beta}} h(\mu(\boldsymbol{x}_i)) \\ &= \sum_{i=1}^{n} \frac{v_i}{\varphi}\left[Y_i - \mu_i\right] \frac{\partial h(\mu_i)}{\partial \mu_i}\frac{\partial \mu_i}{\partial \eta_i}\nabla_{\boldsymbol{\beta}}\eta(\boldsymbol{x}_i) \qquad (5.9) \\ &= \sum_{i=1}^{n} \frac{v_i}{\varphi}\frac{Y_i - \mu_i}{V(\mu_i)}\left(\frac{\partial g(\mu_i)}{\partial \mu_i}\right)^{-1}\boldsymbol{x}_i, \end{aligned}$$

where we use the definition of the variance function $V(\mu) = (\kappa'' \circ h)(\mu)$, see Corollary 2.14. We define the diagonal working weight matrix, which in general depends on $\boldsymbol{\beta}$ through the means $\mu_i = g^{-1}\langle\boldsymbol{\beta}, \boldsymbol{x}_i\rangle$,

$$W(\boldsymbol{\beta}) = \operatorname{diag}\left(\left(\frac{\partial g(\mu_i)}{\partial \mu_i}\right)^{-2}\frac{v_i}{\varphi}\frac{1}{V(\mu_i)}\right)_{1\le i\le n} \in \mathbb{R}^{n\times n},$$

and the working residuals

$$\boldsymbol{R} = \boldsymbol{R}(\boldsymbol{Y}, \boldsymbol{\beta}) = \left(\frac{\partial g(\mu_i)}{\partial \mu_i}(Y_i - \mu_i)\right)^{\top}_{1\le i\le n} \in \mathbb{R}^n.$$

This allows us to write the score equations in a compact form, which provides the following proposition.

Proposition 5.1 *The MLE for $\boldsymbol{\beta}$ is found by solving the score equations*

$$s(\boldsymbol{\beta}, \boldsymbol{Y}) = \nabla_{\boldsymbol{\beta}} \ell_{\boldsymbol{Y}}(\boldsymbol{\beta}) = \mathfrak{X}^\top W(\boldsymbol{\beta}) \boldsymbol{R}(\boldsymbol{Y}, \boldsymbol{\beta}) = 0.$$

For the canonical link $g = h = (\kappa')^{-1}$ the score equations simplify to

$$s(\boldsymbol{\beta}, \boldsymbol{Y}) = \nabla_{\boldsymbol{\beta}} \ell_{\boldsymbol{Y}}(\boldsymbol{\beta}) = \mathfrak{X}^\top \mathrm{diag}\left(\frac{v_i}{\varphi}\right)_{1 \le i \le n} \big(\boldsymbol{Y} - \kappa'(\mathfrak{X}\boldsymbol{\beta})\big) = 0,$$

where $\kappa'(\mathfrak{X}\boldsymbol{\beta}) \in \mathbb{R}^n$ is understood element-wise.

Remarks 5.2

- In general, the MLE of $\boldsymbol{\beta}$ is not calculated by maximizing the log-likelihood function $\ell_{\boldsymbol{Y}}(\boldsymbol{\beta})$, but rather by solving the score equations $s(\boldsymbol{\beta}, \boldsymbol{Y}) = 0$; we also refer to Remarks 3.29 on M- and Z-estimators. The score equations provide the critical points for $\boldsymbol{\beta}$, from which the global maximum of the log-likelihood function can be determined, supposed it exists.
- Existence of a MLE of $\boldsymbol{\beta}$ is not always given, similarly to Example 3.5, we may face the problem that the solution lies at the boundary of the parameter space (which itself may be an open set).
- If the log-likelihood function $\boldsymbol{\beta} \mapsto \ell_{\boldsymbol{Y}}(\boldsymbol{\beta})$ is strictly concave, then the critical point of the score equations $s(\boldsymbol{\beta}, \boldsymbol{Y}) = 0$ is unique, supposed it exists, and, henceforth, we have a unique MLE $\widehat{\boldsymbol{\beta}}^{\mathrm{MLE}}$ for $\boldsymbol{\beta}$. Below, we give cases where the strict concavity of the log-likelihood holds.
- In general, there is no closed from solution for the MLE of $\boldsymbol{\beta}$, except in the Gaussian case with canonical link, thus, we need to solve the score equations numerically.

Similarly to Remarks 3.17 we can calculate Fisher's information matrix w.r.t. $\boldsymbol{\beta}$ through the negative expected Hessian of $\ell_{\boldsymbol{Y}}(\boldsymbol{\beta})$.

We get Fisher's information matrix w.r.t. $\boldsymbol{\beta}$

$$\mathcal{I}(\boldsymbol{\beta}) = \mathbb{E}_{\boldsymbol{\beta}}\left[\nabla_{\boldsymbol{\beta}} \ell_{\boldsymbol{Y}}(\boldsymbol{\beta}) \left(\nabla_{\boldsymbol{\beta}} \ell_{\boldsymbol{Y}}(\boldsymbol{\beta})\right)^\top\right] = -\mathbb{E}_{\boldsymbol{\beta}}\left[\nabla^2_{\boldsymbol{\beta}} \ell_{\boldsymbol{Y}}(\boldsymbol{\beta})\right] = \mathfrak{X}^\top W(\boldsymbol{\beta}) \mathfrak{X}. \tag{5.10}$$

If the design matrix $\mathfrak{X} \in \mathbb{R}^{n \times (q+1)}$ has full rank $q + 1 \le n$, Fisher's information matrix $\mathcal{I}(\boldsymbol{\beta})$ is positive definite.

Dispersion parameter $\varphi > 0$ has been treated as a nuisance parameter above. Its explicit specification does not influence the MLE of $\boldsymbol{\beta}$ because it cancels in the score equations. If necessary, we can also estimate this dispersion parameter with MLE. This requires solving the additional score equation

$$\frac{\partial}{\partial \varphi}\ell_{\boldsymbol{Y}}(\boldsymbol{\beta}, \varphi) = \sum_{i=1}^{n} -\frac{v_i}{\varphi^2}\Big[Y_i h(\mu(\boldsymbol{x}_i)) - \kappa\left(h(\mu(\boldsymbol{x}_i))\right)\Big] + \frac{\partial}{\partial \varphi} a(Y_i; v_i/\varphi) = 0, \tag{5.11}$$

and we can plug in the MLE of $\boldsymbol{\beta}$ (which can be estimated independently of φ). Fisher's information matrix is in this extended framework given by

$$\mathcal{I}(\boldsymbol{\beta}, \varphi) = -\mathbb{E}_{\boldsymbol{\beta}}\left[\nabla^2_{(\boldsymbol{\beta},\varphi)}\ell_{\boldsymbol{Y}}(\boldsymbol{\beta}, \varphi)\right] = \begin{pmatrix} \mathfrak{X}^\top W(\boldsymbol{\beta})\mathfrak{X} & 0 \\ 0 & -\mathbb{E}_{\boldsymbol{\beta}}\left[\partial^2 \ell_{\boldsymbol{Y}}(\boldsymbol{\beta}, \varphi)/\partial \varphi^2\right] \end{pmatrix},$$

that is, the off-diagonal terms between $\boldsymbol{\beta}$ and φ are zero.

In view of Proposition 5.1 we need a root search algorithm to obtain the MLE of $\boldsymbol{\beta}$. Typically, one uses Fisher's scoring method or the iterative re-weighted least squares (IRLS) algorithm to solve this root search problem. This is a main result derived in the seminal work of Nelder–Wedderburn [283] and it explains the popularity of GLMs, namely, GLMs can be solved efficiently by this algorithm. Fisher's scoring method/IRLS algorithm explore the updates for $t \geq 0$ until convergence

$$\widehat{\boldsymbol{\beta}}^{(t)} \mapsto \widehat{\boldsymbol{\beta}}^{(t+1)} = \left(\mathfrak{X}^\top W(\widehat{\boldsymbol{\beta}}^{(t)})\mathfrak{X}\right)^{-1} \mathfrak{X}^\top W(\widehat{\boldsymbol{\beta}}^{(t)}) \left(\mathfrak{X}\widehat{\boldsymbol{\beta}}^{(t)} + \boldsymbol{R}(\boldsymbol{Y}, \widehat{\boldsymbol{\beta}}^{(t)})\right), \tag{5.12}$$

where all terms on the right-hand side are evaluated at algorithmic time t. If we have n observations $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^\top$ we can estimate at most n parameters. Therefore, in our GLM we assume to have a regression parameter $\boldsymbol{\beta} \in \mathbb{R}^{q+1}$ of dimension $q + 1 \leq n$. Moreover, we require that the design matrix $\mathfrak{X}$ has full rank $q + 1 \leq n$. Otherwise the regression parameter is not uniquely identifiable since linear dependence in the columns of $\mathfrak{X}$ allows us to reduce the dimension of the parameter space to a smaller representation. This is also needed to calculate the inverse matrix in (5.12). This motivates the following assumption.

Assumption 5.3 *Throughout, we assume that the design matrix* $\mathfrak{X} \in \mathbb{R}^{n \times (q+1)}$ *has full rank* $q + 1 \le n$.

Remarks 5.4 (Justification of Fisher's Scoring Method/IRLS Algorithm)

- We give a short justification of Fisher's scoring method/IRLS algorithm, for a more detailed treatment we refer to Section 2.5 in McCullagh–Nelder [265] and Section 2.2 in Fahrmeir–Tutz [123].

  The Newton–Raphson algorithm provides a numerical scheme to find solutions to the score equations. It requires to iterate for $t \ge 0$

$$\widehat{\boldsymbol{\beta}}^{(t)} \quad \mapsto \quad \widehat{\boldsymbol{\beta}}^{(t+1)} = \widehat{\boldsymbol{\beta}}^{(t)} + \widehat{\mathcal{I}}(\widehat{\boldsymbol{\beta}}^{(t)})^{-1}\, s(\widehat{\boldsymbol{\beta}}^{(t)}, \boldsymbol{Y}),$$

  where $\widehat{\mathcal{I}}(\boldsymbol{\beta}) = -\nabla^2_{\boldsymbol{\beta}} \ell_{\boldsymbol{Y}}(\boldsymbol{\beta})$ denotes the observed information matrix in $\boldsymbol{\beta} \in \mathbb{R}^{q+1}$. The calculation of the inverse of the observed information matrix $(\widehat{\mathcal{I}}(\widehat{\boldsymbol{\beta}}^{(t)}))^{-1}$ can be time consuming and unstable because we need to calculate second derivatives and the eigenvalues of the observed information matrix can be close to zero. A stable scheme is obtained by replacing the observed information matrix $\widehat{\mathcal{I}}(\boldsymbol{\beta})$ by Fisher's information matrix $\mathcal{I}(\boldsymbol{\beta}) = \mathbb{E}_{\boldsymbol{\beta}}[\widehat{\mathcal{I}}(\boldsymbol{\beta})]$ being positive definite under Assumption 5.3; this provides a quasi-Newton method. Thus, for Fisher's scoring method we iterate for $t \ge 0$

$$\widehat{\boldsymbol{\beta}}^{(t)} \quad \mapsto \quad \widehat{\boldsymbol{\beta}}^{(t+1)} = \widehat{\boldsymbol{\beta}}^{(t)} + \mathcal{I}(\widehat{\boldsymbol{\beta}}^{(t)})^{-1}\, s(\widehat{\boldsymbol{\beta}}^{(t)}, \boldsymbol{Y}), \tag{5.13}$$

  and rewriting this provides us exactly with (5.12). The latter can also be interpreted as an IRLS scheme where the response $g(Y_i)$ is replaced by an adjusted linearized version $Z_i = g(\mu_i) + \frac{\partial g(\mu_i)}{\partial \mu_i}(Y_i - \mu_i)$. This corresponds to the last bracket in (5.12), and with corresponding weights.
- Under the canonical link choice, Fisher's information matrix and the observed information matrix coincide, i.e. $\mathcal{I}(\boldsymbol{\beta}) = \widehat{\mathcal{I}}(\boldsymbol{\beta})$, and the Newton–Raphson algorithm, Fisher's scoring method and the IRLS algorithm are identical. This can easily be seen from Proposition 5.1. We receive under the canonical link choice

$$\begin{aligned} \nabla^2_{\boldsymbol{\beta}} \ell_{\boldsymbol{Y}}(\boldsymbol{\beta}) \;&=\; -\widehat{\mathcal{I}}(\boldsymbol{\beta}) = -\mathfrak{X}^\top \mathrm{diag}\left(\frac{v_i}{\varphi} V(\mu_i)\right)_{1 \le i \le n} \mathfrak{X} \\ &= -\mathfrak{X}^\top W(\boldsymbol{\beta}) \mathfrak{X} \;=\; -\mathcal{I}(\boldsymbol{\beta}). \end{aligned} \tag{5.14}$$

The full rank assumption $q + 1 \leq n$ on the design matrix $\mathfrak{X}$ implies that Fisher's information matrix $\mathcal{I}(\boldsymbol{\beta})$ is positive definite. This in turn implies that the log-likelihood function $\ell_{\boldsymbol{Y}}(\boldsymbol{\beta})$ is strictly concave, providing uniqueness of a critical point (supposed it exists). This indicates that the canonical link has very favorable properties for MLE. Examples 5.5 and 5.6 give two examples not using the canonical link, the first one is a concave maximization problem, the second one is not for $p > 2$.

Example 5.5 (Gamma Model with Log-Link) We study the gamma distribution as a single-parameter EDF model, choosing the shape parameter $\alpha = 1/\varphi$ as the inverse of the dispersion parameter, see Sect. 2.2.2. Cumulant function $\kappa(\theta) = -\log(-\theta)$ gives us the canonical link $\theta = h(\mu) = -1/\mu$. Moreover, we choose the log-link $\eta = g(\mu) = \log(\mu)$ for the GLM. This gives a canonical parameter $\theta = -\exp\{-\eta\}$. We receive the score

$$s(\boldsymbol{\beta}, \boldsymbol{Y}) \;=\; \nabla_{\boldsymbol{\beta}} \ell_{\boldsymbol{Y}}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \frac{v_i}{\varphi} \left[\frac{Y_i}{\mu_i} - 1 \right] \boldsymbol{x}_i \;=\; \mathfrak{X}^\top \mathrm{diag}\left(\frac{v_i}{\varphi} \right)_{1 \leq i \leq n} \boldsymbol{R}(\boldsymbol{Y}, \boldsymbol{\beta}).$$

Unlike in other examples with non-canonical links, we receive a favorable expression here because only one term in the square bracket depends on the regression parameter $\boldsymbol{\beta}$, or equivalently, the working weight matrix W does not dependent on $\boldsymbol{\beta}$. We calculate the negative Hessian (observed information matrix)

$$\widehat{\mathcal{I}}(\boldsymbol{\beta}) \;=\; -\nabla^2_{\boldsymbol{\beta}} \ell_{\boldsymbol{Y}}(\boldsymbol{\beta}) = \mathfrak{X}^\top \mathrm{diag}\left(\frac{v_i}{\varphi} \frac{Y_i}{\mu_i} \right)_{1 \leq i \leq n} \mathfrak{X}.$$

In the gamma model all observations Y_i are strictly positive, a.s., and under the full rank assumption $q + 1 \leq n$, the observed information matrix $\widehat{\mathcal{I}}(\boldsymbol{\beta})$ is positive definite, thus, we have a strictly concave log-likelihood function in the gamma case with log-link. ■

Example 5.6 (Tweedie's Models with Log-Link) We study Tweedie's models for power variance parameters $p > 1$ as a single-parameter EDF model, see Sect. 2.2.3. The cumulant function κ_p is given in Table 4.1. This gives us the canonical link $\theta = h_p(\mu) = \mu^{1-p}/(1-p) < 0$ for $\mu > 0$ and $p > 1$. Moreover, we choose the log-link $\eta = g(\mu) = \log(\mu)$ for the GLM. This implies $\theta = \exp\{(1-p)\eta\}/(1-p) < 0$ for $p > 1$. We receive the score

$$s(\boldsymbol{\beta}, \boldsymbol{Y}) \;=\; \nabla_{\boldsymbol{\beta}} \ell_{\boldsymbol{Y}}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \frac{v_i}{\varphi} \frac{Y_i - \mu_i}{\mu_i^{p-1}} \boldsymbol{x}_i \;=\; \mathfrak{X}^\top \mathrm{diag}\left(\frac{v_i}{\varphi} \frac{1}{\mu_i^{p-2}} \right)_{1 \leq i \leq n} \boldsymbol{R}(\boldsymbol{Y}, \boldsymbol{\beta}).$$

We calculate the negative Hessian (observed information matrix) for $\mu_i > 0$

$$\widehat{\mathcal{I}}(\boldsymbol{\beta}) \;=\; -\nabla^2_{\boldsymbol{\beta}} \ell_{\boldsymbol{Y}}(\boldsymbol{\beta}) = \mathfrak{X}^\top \mathrm{diag}\left(\frac{v_i}{\varphi}\frac{(p-1)Y_i-(p-2)\mu_i}{\mu_i^{p-1}}\right)_{1\le i\le n} \mathfrak{X}.$$

This matrix is positive definite for $p \in [1, 2]$, and for $p > 2$ it is not positive definite because $(p-1)Y_i-(p-2)\mu_i$ may have positive or negative values if we vary $\mu_i > 0$ over its domain $\mathcal{M}$. Thus, we do not have concavity of the optimization problem under the log-link choice in Tweedie's GLMs for power variance parameters $p > 2$. This in particular applies to the inverse Gaussian GLM with log-link. ■

5.1.5 Balance Property Under the Canonical Link Choice

Throughout this section we work under the canonical link choice $g = h = (\kappa')^{-1}$. This choice has very favorable statistical properties. We have already seen in Remarks 5.4 that the derivation of the MLE of $\boldsymbol{\beta}$ becomes particularly easy under the canonical link choice and the observed information matrix $\widehat{\mathcal{I}}(\boldsymbol{\beta})$ coincides with Fisher's information matrix $\mathcal{I}(\boldsymbol{\beta})$ in this case, see (5.14).

For insurance pricing, canonical links have another very remarkable property, namely, that the estimated model automatically fulfills the balance property and, henceforth, is unbiased. This is particularly important in insurance pricing because it tells us that the insurance prices (over the entire portfolio) are on the right level. We have already met the balance property in Corollary 3.19.

Corollary 5.7 (Balance Property) *Assume that* $\boldsymbol{Y}$ *has independent components being modeled by a GLM under the canonical link choice* $g = h = (\kappa')^{-1}$. *Assume that the MLE of regression parameter* $\boldsymbol{\beta} \in \mathbb{R}^{q+1}$ *exists and denote it by* $\widehat{\boldsymbol{\beta}}^{\mathrm{MLE}}$. *We have balance property on portfolio level (for constant dispersion* φ*)*

$$\sum_{i=1}^n \mathbb{E}_{\widehat{\boldsymbol{\beta}}^{\mathrm{MLE}}}\left[v_i Y_i\right] = \sum_{i=1}^n v_i \kappa'\langle \widehat{\boldsymbol{\beta}}^{\mathrm{MLE}}, \boldsymbol{x}_i\rangle = \sum_{i=1}^n v_i Y_i.$$

Proof The first column of the design matrix $\mathfrak{X}$ is identically equal to 1 representing the intercept, see (5.4). The second part of Proposition 5.1 then provides for this first column of $\mathfrak{X}$, we cancel the (constant) dispersion φ,

$$(1,\ldots,1)\,\mathrm{diag}(v_1,\ldots,v_n)\,\kappa'(\mathfrak{X}\widehat{\boldsymbol{\beta}}^{\mathrm{MLE}}) = (1,\ldots,1)\,\mathrm{diag}(v_1,\ldots,v_n)\,\boldsymbol{Y}.$$

This proves the claim. □

Remark 5.8 We mention once more that this balance property is very strong and useful, see also Remarks 3.20. In particular, the balance property holds, even though the chosen GLM might be completely misspecified. Misspecification may include an incorrect distributional model, not the right link function choice, or if we have not pre-processed features appropriately, etc. Such misspecification will imply that we have a poor model on an insurance policy level (observation level). However, the total premium charged over the entire portfolio will be on the right level (supposed that the structure of the portfolio does not change) because it matches the observations, and henceforth, we have unbiasedness for the portfolio mean.

From the log-likelihood function (5.8) we see that under the canonical link choice we consider the statistics $S(\boldsymbol{Y}) = \mathfrak{X}^\top \text{diag}(v_i/\varphi)_{1\le i\le n}\boldsymbol{Y} \in \mathbb{R}^{q+1}$, and to prove the balance property we have used the first component of this statistics. Considering all components, $S(\boldsymbol{Y})$ is an unbiased estimator (decision rule) for

$$\mathbb{E}_{\boldsymbol{\beta}}\left[S(\boldsymbol{Y})\right] = \mathfrak{X}^\top \text{diag}(v_i/\varphi)_{1\le i\le n}\kappa'(\mathfrak{X}\boldsymbol{\beta}) = \left(\sum_{i=1}^n \frac{v_i}{\varphi}\kappa'\langle \boldsymbol{\beta}, \boldsymbol{x}_i\rangle x_{i,j}\right)^\top_{0\le j\le q}. \tag{5.15}$$

This unbiased estimator $S(\boldsymbol{Y})$ meets the Cramér–Rao information bound, hence it is UMVU: taking the partial derivatives of the previous expression gives $\nabla_{\boldsymbol{\beta}}\mathbb{E}_{\boldsymbol{\beta}}\left[S(\boldsymbol{Y})\right] = \mathcal{I}(\boldsymbol{\beta})$, the latter also being the multivariate Cramér–Rao information bound for the unbiased decision rule $S(\boldsymbol{Y})$ for (5.15). Focusing on the first component we have

$$\text{Var}_{\boldsymbol{\beta}}\left(\sum_{i=1}^n \mathbb{E}_{\widehat{\boldsymbol{\beta}}^{\text{MLE}}}\left[v_iY_i\right]\right) = \text{Var}_{\boldsymbol{\beta}}\left(\sum_{i=1}^n v_iY_i\right) = \sum_{i=1}^n \varphi v_i V(\mu_i) = \varphi^2\left(\mathcal{I}(\boldsymbol{\beta})\right)_{0,0}, \tag{5.16}$$

where the component $(0, 0)$ in the last expression is the top-left entry of Fisher's information matrix $\mathcal{I}(\boldsymbol{\beta})$ under the canonical link choice.

5.1.6 Asymptotic Normality

Formula (5.16) quantifies the uncertainty in the premium calculation of the insurance policies if we use the MLE estimated model (under the canonical link choice). That is, this quantifies the uncertainty in the dual mean parametrization in terms of the resulting variance. We could also focus on the MLE $\widehat{\boldsymbol{\beta}}^{\text{MLE}}$ itself (for general link function g). In general, this MLE is not unbiased but we have consistency and asymptotic normality similar to Theorem 3.28. Under "certain regularity conditions"[1] we have for n large

$$\widehat{\boldsymbol{\beta}}_n^{\text{MLE}} \overset{(\text{d})}{\approx} \mathcal{N}\left(\boldsymbol{\beta}, \mathcal{I}_n(\boldsymbol{\beta})^{-1}\right), \tag{5.17}$$

where $\widehat{\boldsymbol{\beta}}_n^{\text{MLE}}$ is the MLE based on the observations $\boldsymbol{Y}_n = (Y_1, \dots, Y_n)^\top$, and $\mathcal{I}_n(\boldsymbol{\beta})$ is Fisher's information matrix of $\boldsymbol{Y}_n$, which scales linearly in n in the homogeneous EF case, see Remarks 3.14, and in the homogeneous EDF case it scales as $\sum_{i=1}^n v_i$, see (3.25).

5.1.7 Maximum Likelihood Estimation and Unit Deviances

From formula (5.7) we conclude that the MLE $\widehat{\boldsymbol{\beta}}^{\text{MLE}}$ of $\boldsymbol{\beta} \in \mathbb{R}^{q+1}$ is found by the solution of (subject to existence)

$$\widehat{\boldsymbol{\beta}}^{\text{MLE}} = \underset{\boldsymbol{\beta}}{\arg\max}\, \ell_{\boldsymbol{Y}}(\boldsymbol{\beta}) = \underset{\boldsymbol{\beta}}{\arg\max} \sum_{i=1}^n \frac{v_i}{\varphi}\Big[Y_i h(\mu(\boldsymbol{x}_i)) - \kappa\left(h(\mu(\boldsymbol{x}_i))\right)\Big],$$

with $\mu_i = \mu(\boldsymbol{x}_i) = \mathbb{E}_{\theta(\boldsymbol{x}_i)}[Y] = g^{-1}\langle\boldsymbol{\beta}, \boldsymbol{x}_i\rangle$ under the link choice g. If we prefer to work with an objective function that reflects the notion of a loss function, we can work under the unit deviances $\mathfrak{d}(Y_i, \mu_i)$ studied in Sect. 4.1.2. The MLE is then obtained by, see (4.20)–(4.21),

$$\widehat{\boldsymbol{\beta}}^{\text{MLE}} = \underset{\boldsymbol{\beta}}{\arg\max}\, \ell_{\boldsymbol{Y}}(\boldsymbol{\beta}) = \underset{\boldsymbol{\beta}}{\arg\min} \sum_{i=1}^n \frac{v_i}{\varphi}\, \mathfrak{d}(Y_i, \mu_i), \tag{5.18}$$

the latter satisfying $\mathfrak{d}(Y_i, \mu_i) \geq 0$ for all $1 \leq i \leq n$, and being zero if and only if $Y_i = \mu_i$, see Lemma 2.22. Thus, using the unit deviances we have a loss function that is bounded below by zero, and we determine the regression parameter $\boldsymbol{\beta}$ such that this loss is (in-sample) minimized. This can also be interpreted in a more geometric way. Consider the $(q+1)$-dimensional manifold $\mathfrak{M} \subset \mathbb{R}^n$ spanned by the GLM function

$$\boldsymbol{\beta} \mapsto \boldsymbol{\mu}(\boldsymbol{\beta}) = g^{-1}(\mathfrak{X}\boldsymbol{\beta}) = (g^{-1}\langle\boldsymbol{\beta}, \boldsymbol{x}_1\rangle, \dots, g^{-1}\langle\boldsymbol{\beta}, \boldsymbol{x}_n\rangle)^\top \in \mathbb{R}^n. \tag{5.19}$$

[1] The regularity conditions for asymptotic normality results will depend on the particular regression problem studied, we refer to pages 43–44 in Fahrmeir–Tutz [123].

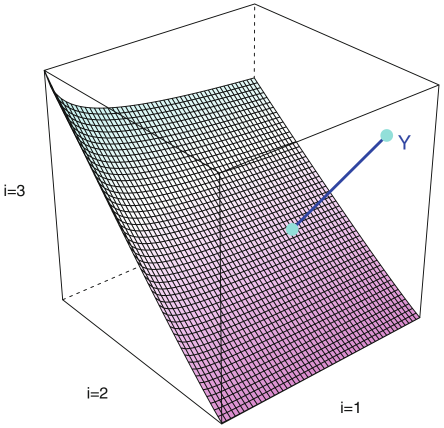


Fig. 5.2 2-dimensional manifold $\mathfrak{M} \subset \mathbb{R}^3$ for observation $\boldsymbol{Y} = (Y_1, Y_2, Y_3)^\top \in \mathbb{R}^3$, the straight line illustrates the projection (w.r.t. the unit deviance distances $\mathfrak{d}$) of $\boldsymbol{Y}$ onto $\mathfrak{M}$ which gives MLE $\widehat{\boldsymbol{\beta}}^{\text{MLE}}$ satisfying $\boldsymbol{\mu}(\widehat{\boldsymbol{\beta}}^{\text{MLE}}) \in \mathfrak{M}$

Minimization (5.18) then tries to find the point $\boldsymbol{\mu}(\boldsymbol{\beta})$ in this manifold $\mathfrak{M} \subset \mathbb{R}^n$ that minimizes simultaneously all unit deviances $\mathfrak{d}(Y_i, \cdot)$ w.r.t. the observation $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^\top \in \mathbb{R}^n$. Or in other words, the optimal parameter $\boldsymbol{\beta}$ is obtained by "projecting" observation $\boldsymbol{Y}$ onto this manifold $\mathfrak{M}$, where "projection" is understood as a simultaneous minimization of loss function $\sum_{i=1}^n \frac{v_i}{\varphi} \mathfrak{d}(Y_i, \mu_i)$, see Fig. 5.2. In the un-weighted Gaussian case, this corresponds to the usual orthogonal projection as the next example shows, and in the non-Gaussian case it is understood in the KL divergence minimization sense as displayed in formula (4.11).

Example 5.9 (Gaussian Case) Assume we have the Gaussian EDF case $\kappa(\theta) = \theta^2/2$ with canonical link $g(\mu) = h(\mu) = \mu$. In this case, the manifold (5.19) is the linear space spanned by the columns of the design matrix $\mathfrak{X}$

$$\boldsymbol{\beta} \;\mapsto\; \boldsymbol{\mu}(\boldsymbol{\beta}) = \mathfrak{X}\boldsymbol{\beta} = (\langle \boldsymbol{\beta}, \boldsymbol{x}_1 \rangle, \ldots, \langle \boldsymbol{\beta}, \boldsymbol{x}_n \rangle)^\top \;\in\; \mathbb{R}^n.$$

If additionally we assume $v_i/\varphi = c > 0$ for all $1 \le i \le n$, the minimization problem (5.18) reads as

$$\widehat{\boldsymbol{\beta}}^{\text{MLE}} \;=\; \underset{\boldsymbol{\beta}}{\arg\min} \sum_{i=1}^n \frac{v_i}{\varphi}\, \mathfrak{d}(Y_i, \mu_i) \;=\; \underset{\boldsymbol{\beta}}{\arg\min} \, \|\boldsymbol{Y} - \mathfrak{X}\boldsymbol{\beta}\|_2^2,$$

where we have used that the unit deviances in the Gaussian case are given by the square loss function, see Example 4.12. As a consequence, the MLE $\widehat{\boldsymbol{\beta}}^{\text{MLE}}$ is found by orthogonally projecting $\boldsymbol{Y}$ onto $\mathfrak{M} = \{\mathfrak{X}\boldsymbol{\beta} | \, \boldsymbol{\beta} \in \mathbb{R}^{q+1}\} \subset \mathbb{R}^n$, and this orthogonal projection is given by $\mathfrak{X}\widehat{\boldsymbol{\beta}}^{\text{MLE}} \in \mathfrak{M}$. ■

5.2 Actuarial Applications of Generalized Linear Models

The purpose of this section is to illustrate how the concept of GLMs is used in actuarial modeling. We therefore explore the typical actuarial examples of claim counts and claim size modeling.

5.2.1 Selection of a Generalized Linear Model

The selection of a predictive model within GLMs for solving an applied actuarial problem requires the following choices.

Choice of the Member of the EDF Select a member of the EDF that fits the modeling problem. In a first step, we should try to understand the properties of the data $\boldsymbol{Y}$ before doing this selection, for instance, do we have count data, do we have a classification problem, do we have continuous observations?

All members of the EDF are light-tailed because the moment generating function exists around the origin, see Corollary 2.14, and the EDF is not suited to model heavy-tailed data, for instance, having a regularly varying tail. Therefore, a datum Y is sometimes first transformed before being modeled by a member of the EDF. A popular transformation is the logarithm for positive observations. After this transformation a member of the EDF can be chosen to model $\log(Y)$. For instance, if we choose the Gaussian distribution for $\log(Y)$, then Y will be log-normally distributed, or if we choose the exponential distribution for $\log(Y)$, then Y will be Pareto distributed, see Sect. 2.2.5. One can then model the transformed datum with a GLM. Often this provides very accurate models, say, on the log scale for the log-transformed data. There is one issue with this approach, namely, if a model is unbiased on the transformed scale then it is typically biased on the original observation scale; if the transformation is concave this easily follows from Jensen's inequality. The problematic part now is that the bias correction itself often has systematic effects which means that the transformation (or the involved nuisance parameters) should be modeled with a regression model, too, see Sect. 5.3.9. In many cases this will not easily work, unfortunately. Therefore, if possible, clear preference should be given to modeling the data on the original observation scale (if unbiasedness is a central requirement).

Choice of Link Function From a statistical point of view we should choose the canonical link $g = h$ to connect the mean μ of the model to the linear predictor η because this implies many favorable mathematical properties. However, as seen, sometimes we have different needs. Practical reasons may require that we have a model with additive or multiplicative effects, which favors the identity or the log-link, respectively. Another requirement is that the resulting canonical parameter $\theta = (h \circ g^{-1})(\eta)$ needs to be within the effective domain $\boldsymbol{\Theta}$. If this effective domain is bounded, for instance, if it covers the negative real line as for the gamma model, a (transformation of the) log-link might be more suitable than the canonical link because $g^{-1}(\cdot) = -\exp(\cdot)$ has a strictly negative range, see Example 5.5.

Choice of Features and Feature Engineering Assume we have selected the member of the EDF and the link function g. This gives us the relationship between the mean μ and the linear predictor η, see (5.5),

$$\mu(\boldsymbol{x}) = \mathbb{E}_{\theta(\boldsymbol{x})}[Y] = g^{-1}(\eta(\boldsymbol{x})) = g^{-1}\langle \boldsymbol{\beta}, \boldsymbol{x}\rangle. \tag{5.20}$$

Thus, the features $\boldsymbol{x} \in \mathcal{X} \subset \mathbb{R}^{q+1}$ need to be in the right functional form so that they can appropriately describe the systematic effect via the function (5.20). We distinguish the following feature types:

- *Continuous real-valued feature components*, examples are age of policyholder, weight of car, body mass index, etc.
- *Ordinal categorical feature components*, examples are ratings like good-medium-bad or A-B-C-D-E.
- *Nominal categorical feature components*, examples are vehicle brands, occupation of policyholders, provinces of living places of policyholders, etc. The values that the categorical feature components can take are called *levels*.
- *Binary feature components* are special categorical features that only have two levels, e.g. female-male, open-closed. Because binary variables often play a distinguished role in modeling they are separated from categorical variables which are typically assumed to have more than two levels.

All these components need to be brought into a suitable form so that they can be used in a linear predictor $\eta(\boldsymbol{x}) = \langle \boldsymbol{\beta}, \boldsymbol{x}\rangle$, see (5.20). This requires the consideration of the following points (1) transformation of continuous components so that they can describe the systematic effects in a linear form, (2) transformation of categorical components to real-valued components, (3) interaction of components beyond an additive structure in the linear predictor, and (4) the resulting design matrix $\mathfrak{X}$ should have full rank $q + 1 \le n$. We are going to describe these points (1)–(4) in the next section.

5.2.2 *Feature Engineering*

Categorical Feature Components: Dummy Coding

Categorical variables need to be embedded into a Euclidean space. This embedding needs to be done such that the resulting design matrix $\mathfrak{X}$ has full rank $q + 1 \le n$. There are many different ways to do so, and the particular choice depends on the modeling purpose. The most popular way is *dummy coding*. We only describe dummy coding here because it is sufficient for our purposes, but we mention that

Table 5.1 Dummy coding example that maps the $K = 11$ levels (colors) to the unit vectors of the 10-dimensional Euclidean space $\mathbb{R}^{10}$ selecting the last level a_{11} (brown color) as reference level, and showing the resulting dummy vectors $\boldsymbol{x}_j^\top$ as row vectors

a_1 = white	1	0	0	0	0	0	0	0	0	0
a_2 = yellow	0	1	0	0	0	0	0	0	0	0
a_3 = orange	0	0	1	0	0	0	0	0	0	0
a_4 = red	0	0	0	1	0	0	0	0	0	0
a_5 = magenta	0	0	0	0	1	0	0	0	0	0
a_6 = violet	0	0	0	0	0	1	0	0	0	0
a_7 = blue	0	0	0	0	0	0	1	0	0	0
a_8 = cyan	0	0	0	0	0	0	0	1	0	0
a_9 = green	0	0	0	0	0	0	0	0	1	0
a_{10} = beige	0	0	0	0	0	0	0	0	0	1
a_{11} = brown	0	0	0	0	0	0	0	0	0	0

there are also other codings like effects coding or Helmert's contrast coding.[2] The choice of the coding will not influence the predictive model (if we work with a full rank design matrix), but it may influence parameter selection, parameter reduction and model interpretation. For instance, the choice of the coding is (more) important in medical studies where one tries to understand the effects between certain therapies.

Assume that the raw feature component $\widetilde{x}_j$ is a categorical variable taking K different levels $\{a_1, \ldots, a_K\}$. For dummy coding we declare one level, say a_K, to be the reference level and all other levels are described relative to that reference level. Formally, this can be described by an embedding map

$$\widetilde{x}_j \;\mapsto\; \boldsymbol{x}_j = (\mathbb{1}_{\{\widetilde{x}_j=a_1\}}, \ldots, \mathbb{1}_{\{\widetilde{x}_j=a_{K-1}\}})^\top \in \mathbb{R}^{K-1}. \tag{5.21}$$

This is closely related to the categorical distribution in Sect. 2.1.4. An explicit example is given in Table 5.1.

Example 5.10 (Multiplicative Model) If we choose the log-link function $\eta = g(\mu) = \log(\mu)$, we receive the regression function for the categorical example of Table 5.1

$$\widetilde{x}_j \;\mapsto\; \exp\langle \boldsymbol{\beta}, \boldsymbol{x}_j\rangle = \exp\{\beta_0\} \prod_{k=1}^{K-1} \exp\left\{\beta_k \mathbb{1}_{\{\widetilde{x}_j=a_k\}}\right\}, \tag{5.22}$$

including an intercept component. Thus, the base value $\exp\{\beta_0\}$ is determined by the reference level a_{11} = brown, and any color different from brown has a deviation from the base value described by the multiplicative correction term $\exp\{\beta_k \mathbb{1}_{\{\widetilde{x}_j=a_k\}}\}$. ■

[2] There is an example of Helmert's contrast coding in Remarks 2.7 of lecture notes [392], and for more examples we refer to the UCLA statistical consulting website: https://stats.idre.ucla.edu/r/library/r-library-contrast-coding-systems-for-categorical-variables/.

Remarks 5.11

- Importantly, dummy coding leads to full rank design matrices $\mathfrak{X}$ and, henceforth, Assumption 5.3 is fulfilled.
- Dummy coding is different from one-hot encoding which is going to be introduced in Sect. 7.3.1, below.
- Dummy coding needs some care if we have categorical feature components with many levels, for instance, considering car brands and car models we can get hundreds of levels. In that case we will have sparsity in the resulting design matrix. This may cause computational issues, and, as the following example will show, it may lead to high uncertainty in parameter estimation. In particular, the columns of the design matrix $\mathfrak{X}$ of very rare levels will be almost collinear which implies that we do not receive very well-conditioned matrices in Fisher's scoring method (5.12). For this reason, it is recommended to merge levels to bigger classes. In Sect. 7.3.1, below, we are going to present a different treatment. Categorical variables are embedded into low-dimensional spaces, so that proximity in these spaces has a reasonable meaning for the regression task at hand.

Example 5.12 (Balance Property and Dummy Coding) A main argument for the use of the canonical link function has been the fulfillment of the balance property, see Corollary 5.7. If we have categorical feature components and if we apply dummy coding to those, then the balance property is projected down to the individual levels of that categorical variable. Assume that columns 2 to K of design matrix $\mathfrak{X}$ are used to model a raw categorical feature $\widetilde{x}_1$ with K levels according to (5.21). In that case, columns $2 \le k \le K$ will indicate all observations Y_i which belong to levels a_{k-1}. Analogously to the proof of Corollary 5.7, we receive (summation i runs over the different instances/policies)

$$\sum_{i:\, \widetilde{x}_{i,1}=a_{k-1}} \mathbb{E}_{\widehat{\boldsymbol{\beta}}^{\mathrm{MLE}}}\left[v_i Y_i\right] = \sum_{i=1}^{n} x_{i,k} \mathbb{E}_{\widehat{\boldsymbol{\beta}}^{\mathrm{MLE}}}\left[v_i Y_i\right] = \sum_{i=1}^{n} x_{i,k} v_i Y_i = \sum_{i:\, \widetilde{x}_{i,1}=a_{k-1}} v_i Y_i. \tag{5.23}$$

Thus, we receive the balance property for all policies $1 \le i \le n$ that belong to level a_{k-1}.

If we have many levels, then it will happen that some levels have only very few observations, and the above summation (5.23) only runs over very few insurance policies with $\widetilde{x}_{i,1} = a_{k-1}$. Suppose additionally the volumes v_i are small. This can lead to considerable estimation uncertainty, because the estimated prices on the left-hand side of (5.23) will be based too much on individual observations Y_i having the corresponding level, and we are not in the regime of a law of large numbers that balances these observations.

Thus, this balance property from dummy coding is a natural property under the canonical link choice. Actuarial pricing is very familiar with such a property. Early distribution-free approaches have postulated this property resulting in the method of the total marginal sums, see Bailey and Jung [22, 206], where the balance property is enforced for marginal sums of all categorical levels in parameter estimation. However, if we have scarce levels in categorical variables, this approach needs careful consideration. ■

Binary Feature Components

Binary feature components do not need a treatment different from the categorical ones, they are Bernoulli variables which can be encoded as 0 or 1. This is exactly dummy coding for $K = 2$ levels.

Continuous Feature Components

Continuous feature components are already real-valued. Therefore, from the viewpoint of 'variable types', the continuous feature components do not need any pre-processing because they are already in the right format to be included in scalar products.

Nevertheless, in many cases, also continuous feature components need feature engineering because only in rare cases they directly fit the functional form (5.20). We give an example. Consider car drivers that have different driving experience and different driving skills. To explain experience and skills we typically choose the age of driver as explanatory variable. Modeling the claim frequency as a function of the age of driver, we often observe a U-shaped function, thus, a function that is non-monotone in the age of driver variable. Since the link function g needs to be strictly monotone, this regression problem cannot be modeled by (5.20), directly including the age of driver as a feature because this leads to monotonicity of the regression function in the age of driver variable.

Typically, in such situations, the continuous variable is discretized to categorical classes. In the driver's age example, we build age classes. These age classes are then treated as categorical variables using dummy coding (5.21). We will give examples below. These age classes should fulfill the requirement of being sufficiently homogeneous in the sense that insurance policies that fall into the same class should have a similar propensity to claims. This implies that we would like to have many small homogeneous classes. However, the classes should be sufficiently large, otherwise parameter estimation involves high uncertainty, see also Example 5.12. Thus, there is a trade-off to sufficiently meet both of these two requirements.

A disadvantage of this discretization approach is that neighboring age classes will not be recognized by the regression function because, per se, dummy coding is based on nominal variables not having any topology. This is also illustrated by the fact, that all categorical levels (excluding the reference level) have, in view of embedding (5.21), the same mutual Euclidean distance. Therefore, in some applications, one prefers a different approach by rather trying to find an appropriate functional form. For instance, we can pre-process a strictly positive raw feature component $\widetilde{x}_l$ to a higher-dimensional functional form

$$\widetilde{x}_l \;\mapsto\; \beta_1\widetilde{x}_l + \beta_2\widetilde{x}_l^2 + \beta_3\widetilde{x}_l^3 + \beta_4\log(\widetilde{x}_l), \tag{5.24}$$

with regression parameter $(\beta_1, \ldots, \beta_4)^\top$, i.e., we have a polynomial function of degree 3 plus a logarithmic term in this choice. If one does not want to choose a specific functional form, one often chooses natural cubic splines. This, together with regularization, leads to the framework of generalized additive models (GAMs), which is popular family of regression models besides GLMs; for literature on GAMs we refer to Hastie–Tibshirani [182], Wood [384], Ohlsson–Johansson [290], Denuit et al. [99] and Wüthrich–Buser [392]. In these notes we will not further pursue GAMs.

Example 5.13 (Multiplicative Model) If we choose the log-link function $\eta = g(\mu) = \log(\mu)$ we receive a multiplicative regression function

$$\boldsymbol{x} \;\mapsto\; \mu(\boldsymbol{x}) = \exp\langle\boldsymbol{\beta}, \boldsymbol{x}\rangle = \exp\{\beta_0\}\prod_{j=1}^{q}\exp\left\{\beta_j x_j\right\}.$$

That is, all feature components x_j enter the regression function in an exponential form. In general insurance, one may have specific variables for which it is explicitly known that they should enter the regression function as a power function. Having a raw feature $\widetilde{x}_l$ we can pre-process it as $\widetilde{x}_l \mapsto x_l = \log(\widetilde{x}_l)$. This implies

$$\mu(\boldsymbol{x}) = \exp\langle\boldsymbol{\beta}, \boldsymbol{x}\rangle = \exp\{\beta_0\}\,\widetilde{x}_l^{\beta_l}\prod_{j=1, j\neq l}^{q}\exp\left\{\beta_j x_j\right\},$$

which gives a power term of order β_l. The GLM estimates in this case the power parameter that should be used for $\widetilde{x}_l$. If the power parameter is known, then one can even include this component as an offset; offsets are discussed in Sect. 5.2.3, below. ■

Interactions

Naturally, GLMs only allow for an additive structure in the linear predictor. Similar to continuous feature components, such an additive structure may not always be suitable and one wants to model more complex interaction terms. Such interactions need to be added manually by the modeler, for instance, if we have two raw feature components $\widetilde{x}_l$ and $\widetilde{x}_k$, we may want to consider a functional form

$$(\widetilde{x}_l, \widetilde{x}_k) \mapsto \beta_1 \widetilde{x}_l + \beta_2 \widetilde{x}_k + \beta_3 \widetilde{x}_l \widetilde{x}_k + \beta_4 \widetilde{x}_l^2 \widetilde{x}_k,$$

with regression parameter $(\beta_1, \ldots, \beta_4)^\top$.

More generally, this manual feature engineering of adding interactions and of specifying functional forms (5.24) can be understood as a new representation of raw features. Representation learning in relation to deep learning is going to be discussed in Sect. 7.1, and this discussion is also related to Mercer's kernels.

5.2.3 Offsets

In many heterogeneous portfolio problems with observations $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^\top$, there are known prior differences between the individual risks Y_i, for instance, the time exposure varies between the different policies i. Such known prior differences can be integrated into the predictors, and this integration typically does not involve any additional model parameters. A simple way is to use an *offset* (constant) in the linear predictor of a GLM. Assume that each observation Y_i is equipped with a feature $\boldsymbol{x}_i \in \mathcal{X}$ and a known offset $o_i \in \mathbb{R}$ such that the linear predictor η_i takes the form

$$(\boldsymbol{x}_i, o_i) \mapsto g(\mu_i) = \eta_i = \eta(\boldsymbol{x}_i, o_i) = o_i + \langle \boldsymbol{\beta}, \boldsymbol{x}_i \rangle, \tag{5.25}$$

for all $1 \le i \le n$. An offset o_i does not change anything from a structural viewpoint, in fact, it could be integrated into the feature $\boldsymbol{x}_i$ with a regression parameter that is identically equal to 1.

Offsets are frequently used in Poisson models with the (canonical) log-link choice to model multiplicative time exposures in claim frequency modeling. Under the log-link choice we receive from (5.25) the following mean function

$$(\boldsymbol{x}_i, o_i) \mapsto \mu(\boldsymbol{x}_i, o_i) = \exp\{\eta(\boldsymbol{x}_i, o_i)\} = \exp\{o_i + \langle \boldsymbol{\beta}, \boldsymbol{x}_i \rangle\} = \exp\{o_i\} \exp\langle \boldsymbol{\beta}, \boldsymbol{x}_i \rangle.$$

In this version, the offset o_i provides us with an exposure $\exp\{o_i\}$ that acts multiplicatively on the regression function. If $w_i = \exp\{o_i\}$ measures time, then w_i is a so-called pro-rata temporis (proportional in time) exposure.

Remark 5.14 (Boosting) A popular machine learning technique in statistical modeling is boosting. Boosting tries to step-wise adaptively improve a regression model. Offsets (5.25) are a simple way of constructing boosted models. Assume we have constructed a predictive model using any statistical model, and denote the resulting estimated means of Y_i by $\widehat{\mu}_i^{(0)}$. The idea of boosting is that we select another statistical model and we try to see whether this second model can still find systematic structure in the data which has not been found by the first model. In view of (5.25), we include the first model into the offset and we build a second model around this offset, that is, we may explore a GLM

$$\widehat{\mu}_i^{(1)} = g^{-1}\left(g(\widehat{\mu}_i^{(0)}) + \langle \boldsymbol{\beta}, \boldsymbol{x}_i \rangle\right).$$

If the first model is perfect we come up with a regression parameter $\boldsymbol{\beta} = 0$, otherwise the linear predictor $\langle \boldsymbol{\beta}, \boldsymbol{x}_i \rangle$ of the second model starts to compensate for weaknesses in $\widehat{\mu}_i^{(0)}$. Of course, this boosting procedure can then be iterated and one should stop boosting before the resulting model starts to over-fit to the data. Typically, this approach is applied to regression trees instead of GLMs, see Ferrario–Hämmerli [125], Section 7.4 in Wüthrich–Buser [392], Lee–Lin [241] and Denuit et al. [100].

5.2.4 Lab: Poisson GLM for Car Insurance Frequencies

We present a first GLM example. This example is based on French motor third party liability (MTPL) insurance claim counts data. The data is described in detail in Chap. 13.1; an excerpt of the available MTPL data is given in Listing 13.2. For the moment we only consider claim frequency modeling. We use the following data: N_i describes the number of claims, $v_i \in (0, 1]$ describes the duration of the insurance policy, and $\widetilde{\boldsymbol{x}}_i$ describes the available raw feature information of insurance policy i, see Listing 13.2.

We are going to model the claim counts N_i with a Poisson GLM using the canonical link function of the Poisson model. In the Poisson approach there are two different ways to account for the duration of the insurance policy. Either we model $Y_i = N_i/v_i$ with the Poisson model of the EDF, see Sect. 2.2.2 and Remarks 2.13 (reproductive form), or we directly model N_i with the Poisson distribution from the EF and treat the log-duration as an offset variable $o_i = \log v_i$. In the first approach we have for the log-link choice $g(\cdot) = h(\cdot) = \log(\cdot)$ and dispersion $\varphi = 1$

$$Y_i = N_i/v_i \;\sim\; f(y_i; \theta_i, v_i) = \exp\left\{ \frac{y_i \langle \boldsymbol{\beta}, \boldsymbol{x}_i \rangle - e^{\langle \boldsymbol{\beta}, \boldsymbol{x}_i \rangle}}{1/v_i} + a(y_i; v_i) \right\}, \qquad (5.26)$$

where $\boldsymbol{x}_i \in \mathcal{X}$ is the suitably pre-processed feature information of insurance policy i, and with canonical parameter $\theta_i = \eta(\boldsymbol{x}_i) = \langle \boldsymbol{\beta}, \boldsymbol{x}_i \rangle$. In the second approach we include the log-duration as offset into the regression function and model N_i with the Poisson distribution from the EF. Using notation (2.2) this gives us

$$N_i \sim f(n_i;\theta_i) = \exp\Big\{ n_i\left(\log v_i + \langle \boldsymbol{\beta}, \boldsymbol{x}_i\rangle\right) - e^{\log v_i + \langle \boldsymbol{\beta}, \boldsymbol{x}_i\rangle} + a(n_i)\Big\} \qquad (5.27)$$

$$= \exp\left\{ \frac{\frac{n_i}{v_i}\langle \boldsymbol{\beta}, \boldsymbol{x}_i\rangle - e^{\langle \boldsymbol{\beta}, \boldsymbol{x}_i\rangle}}{1/v_i} + a(n_i) + n_i \log v_i \right\},$$

with canonical parameter $\theta_i = \eta(\boldsymbol{x}_i, o_i) = o_i + \langle \boldsymbol{\beta}, \boldsymbol{x}_i\rangle = \log v_i + \langle \boldsymbol{\beta}, \boldsymbol{x}_i\rangle$ for observation $n_i = v_i y_i$. That is, we receive the same model in both cases (5.26) and (5.27) under the canonical log-link choice for the Poisson GLM.

Finally, we make the assumption that all observations N_i are independent. There remains the pre-processing of the raw features $\widetilde{\boldsymbol{x}}_i$ to features $\boldsymbol{x}_i$ so that they can be used in a sensible way in the linear predictors $\eta_i = \eta(\boldsymbol{x}_i, o_i) = o_i + \langle \boldsymbol{\beta}, \boldsymbol{x}_i\rangle$.

Feature Engineering

Categorical and Binary Variables: Dummy Coding

For categorical and binary variables we use dummy coding as described in Sect. 5.2.2. We have two categorical variables `VehBrand` and `Region`, as well as a binary variable `VehGas`, see Listing 13.2. We choose the first level as reference level, and the remaining levels are characterized by $(K-1)$-dimensional embeddings (5.21). This provides us with $K-1 = 10$ parameters for `VehBrand`, $K-1 = 21$ parameters for `Region` and $K-1 = 1$ parameter for `VehGas`.

Figure 5.3 shows the empirical marginal frequencies $\bar{\lambda} = \sum N_i / \sum v_i$ on all levels of the categorical feature components `VehBrand`, `Region` and `VehGas`. Moreover, the blue areas (in the colored version) give confidence bounds of $\pm 2\sqrt{\bar{\lambda}/\sum v_i}$ (under a Poisson assumption), see Example 3.22. The more narrow these confidence bounds, the bigger the volumes $\sum v_i$ behind these empirical marginal estimates.

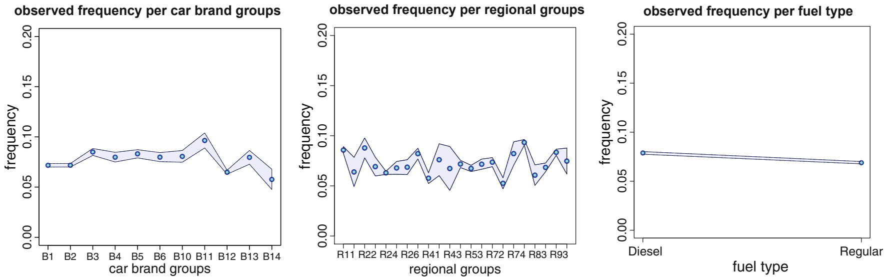


Fig. 5.3 Empirical marginal frequencies on each level of the categorical variables (lhs) `VehBrand`, (middle) `Region`, and (rhs) `VehGas`

Continuous Variables

We consider feature engineering of the continuous variables `Area`, `VehPower`, `VehAge`, `DrivAge`, `BonusMalus` and log-`Density` (`Density` on the log scale); note that we map the `Area` codes $(A, \ldots, F) \mapsto (1, \ldots, 6)$. Some of these variables do not show any monotonicity nor log-linearity in the empirical marginal frequency plots, see Fig. 5.4.

These non-monotonicity and non-log-linearity suggest in a first step to build homogeneous classes for these feature components and use dummy coding for the resulting classes. We make the following choices here (motivated by the marginal graphs of Fig. 5.4):

- `Area`: continuous log-linear feature component for $\{\mathrm{A}, \ldots, \mathrm{F}\} \mapsto \{1, \ldots, 6\}$;
- `VehPower`: discretize into categorical classes where we merge vehicle power groups bigger and equal to 9 (totally $K = 6$ levels);
- `VehAge`: we build categorical classes $[0, 6)$, $[6, 13)$, $[13, \infty)$ (totally $K = 3$ levels);
- `DrivAge`: we build categorical classes $[18, 21)$, $[21, 26)$, $[26, 31)$, $[31, 41)$, $[41, 51)$, $[51, 71)$, $[71, \infty)$ (totally $K = 7$ levels);
- `BonusMalus`: continuous log-linear feature component (we censor at 150);
- `Density`: log-density is chosen as continuous log-linear feature component.

This encoding is slightly different from Noll et al. [287] because of different data cleaning. The discretization has been chosen quite ad-hoc by just looking at the empirical plots; as illustrated in Section 6.1.6 of Wüthrich–Buser [392] regression trees may provide an algorithmic way of choosing homogeneous classes of sufficient volume. This provides us with a feature space (the initial component stands for the intercept $x_{i,0} = 1$ and the order of the terms is the same as in Listing 13.2)

$$\mathcal{X} \subset \{1\} \times \mathbb{R} \times \{0, 1\}^5 \times \{0, 1\}^2 \times \{0, 1\}^6 \times \mathbb{R} \times \{0, 1\}^{10} \times \{0, 1\} \times \mathbb{R} \times \{0, 1\}^{21},$$

of dimension $q+1 = 1+1+5+2+6+1+10+1+1+21 = 49$. The R code [307] for this pre-processing of continuous variables is shown in Listing 5.1, categorical variables do not need any special treatment because variables of `factor` type are consider internally in R by dummy coding; we call this model Poisson GLM1.

Choice of Learning and Test Samples

To measure predictive performance we follow the generalization approach as proposed in Chap. 4. This requires that we partition our entire data into learning sample $\mathcal{L}$ and test sample $\mathcal{T}$, see Fig. 4.1. Model selection and model fitting will be done on the learning sample $\mathcal{L}$, only, and the test sample $\mathcal{T}$ is used to analyze the generalization of the fitted models to unseen data. We partition the data at random (non-stratified) in a ratio of $9 : 1$, and we are going to hold on to the same partitioning throughout this monograph whenever we study this example. The R code used is given in Listing 5.2.

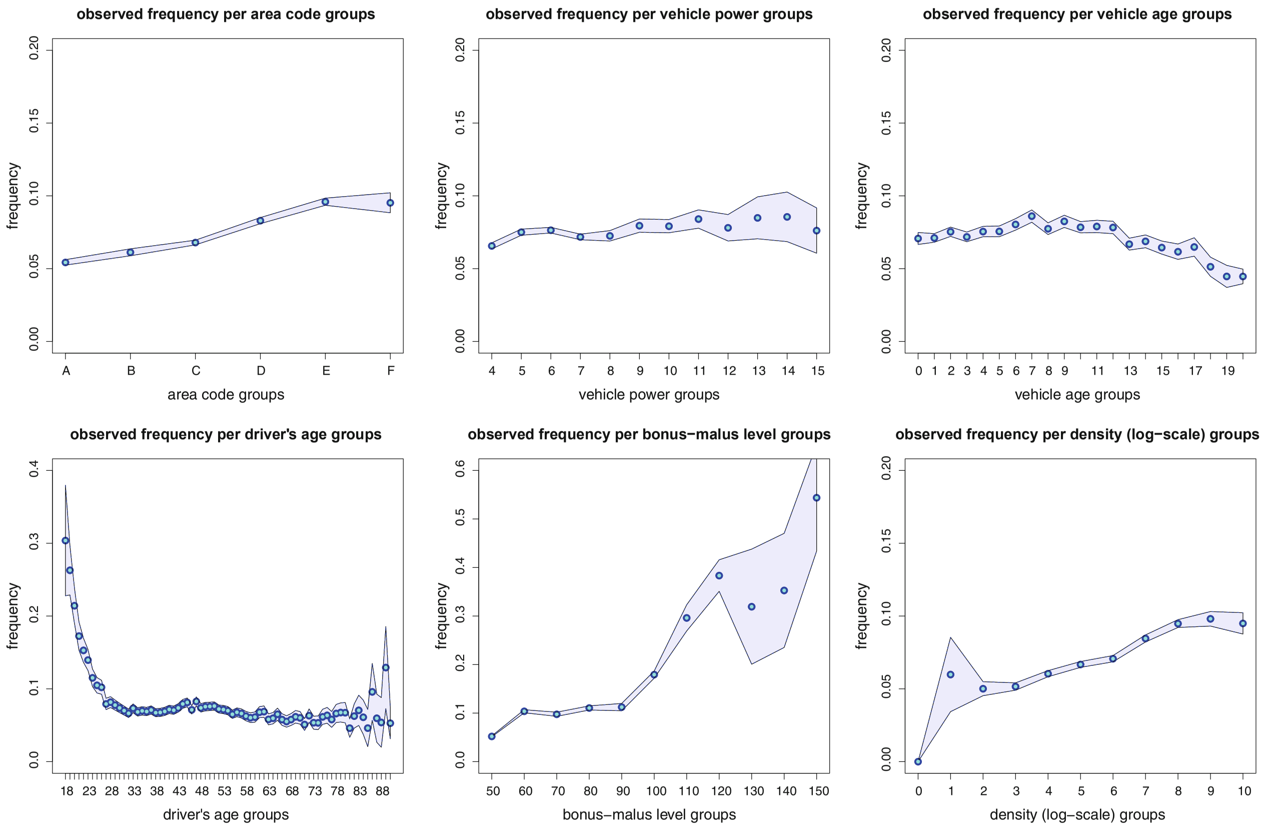


Fig. 5.4 Empirical marginal frequencies of the continuous variables: top row (lhs) `Area`, (middle) `VehPower`, (rhs) `VehAge`, and bottom row (lhs) `DrivAge`, (middle) `BonusMalus`, (rhs) log-`Density`, i.e., `Density` on the log scale; note that `DrivAge` and `BonusMalus` have a different y-scale in these plots

Listing 5.1 Pre-processing of features for model Poisson GLM1 in R

```
dat$AreaGLM     <- as.integer(dat$Area)
dat$VehPowerGLM <- as.factor(pmin(dat$VehPower, 9))
dat$VehAgeGLM   <- as.factor(cut(dat$VehAge, c(0,5,12,101),
                     labels = c("0-5","6-12","12+"),
                     include.lowest = TRUE))
dat$DrivAgeGLM  <- as.factor(cut(dat$DrivAge, c(18,20,25,30,40,50,70,101),
                     labels = c("18-20","21-25","26-30","31-40","41-50",
                                "51-70","71+"), include.lowest = TRUE))
dat$BonusMalusGLM <- pmin(dat$BonusMalus, 150)
dat$DensityGLM  <- log(dat$Density)
```

Table 5.2 shows the summary of the chosen partition into learning and test samples

$$\mathcal{L} = \left\{(Y_i = N_i/v_i, \boldsymbol{x}_i, v_i) : \; i = 1, \ldots, n = 610'206\right\},$$

and

$$\mathcal{T} = \left\{(Y_t^\dagger = N_t^\dagger/v_t^\dagger, \boldsymbol{x}_t^\dagger, v_t^\dagger) : \; t = 1, \ldots, T = 67'801\right\}.$$

In contrast to Sect. 4.2 we also include feature information and exposure information to $\mathcal{L}$ and $\mathcal{T}$.

Listing 5.2 Partition of the data to learning sample $\mathcal{L}$ and test sample $\mathcal{T}$

```
RNGversion("3.5.0")    # we use R version 3.5.0 for this partition
set.seed(500)
ll    <- sample(c(1:nrow(dat)), round(0.9*nrow(dat)), replace = FALSE)
learn <- dat[ll,]
test  <- dat[-ll,]
```

Table 5.2 Choice of learning data set $\mathcal{L}$ and test data set $\mathcal{T}$; the empirical frequency on both data sets is similar (last column), and the split of the policies w.r.t. the numbers of claims is also rather similar

	Numbers of observed claims						Empirical
	0	1	2	3	4	5	frequency
Learning sample $\mathcal{L}$	96.32%	3.47%	0.19%	0.01%	0.0006%	0.0002%	7.36%
Test sample $\mathcal{T}$	96.31%	3.50%	0.18%	0.01%	0.0015%	0.0015%	7.35%

Maximum-Likelihood Estimation and Results

The remaining step is to perform MLE to estimate regression parameter $\boldsymbol{\beta} \in \mathbb{R}^{q+1}$. This can be done either by maximizing the Poisson log-likelihood function or by minimizing the Poisson deviance loss. In view of (4.9) and Example 4.27, the Poisson deviance loss on the learning data $\mathcal{L}$ is given by

$$\boldsymbol{\beta} \mapsto \mathfrak{D}(\mathcal{L}, \boldsymbol{\beta}) = \frac{2}{n} \sum_{i=1}^{n} v_i \left(\mu(\boldsymbol{x}_i) - Y_i - Y_i \log\left(\frac{\mu(\boldsymbol{x}_i)}{Y_i} \right) \right) \geq 0, \tag{5.28}$$

where the terms under the summation are set equal to $v_i \mu(\boldsymbol{x}_i)$ for $Y_i = 0$, see (4.8), and we have GLM regression function

$$\boldsymbol{x} \mapsto \mu(\boldsymbol{x}) = \mu_{\boldsymbol{\beta}}(\boldsymbol{x}) = \exp\langle \boldsymbol{\beta}, \boldsymbol{x} \rangle.$$

That is, we work under the canonical link with the canonical parameter being equal to the linear predictor. The MLE of $\boldsymbol{\beta}$ is found by minimizing (5.28). This is done with Fisher's scoring method. In order to receive a non-degenerate solution we need to ensure that we have sufficiently many claims $Y_i > 0$, otherwise it might happen that the MLE provides a (degenerate) solution at the boundary of the effective domain $\boldsymbol{\Theta}$. We denote the MLE by $\widehat{\boldsymbol{\beta}}_{\mathcal{L}}^{\text{MLE}} = \widehat{\boldsymbol{\beta}}^{\text{MLE}}$, because it has been estimated on the learning data $\mathcal{L}$, only. This gives us estimated regression function

$$\boldsymbol{x} \mapsto \widehat{\mu}(\boldsymbol{x}) = \mu_{\widehat{\boldsymbol{\beta}}_{\mathcal{L}}^{\text{MLE}}}(\boldsymbol{x}) = \exp\langle \widehat{\boldsymbol{\beta}}_{\mathcal{L}}^{\text{MLE}}, \boldsymbol{x} \rangle.$$

We emphasize that we only use the learning data $\mathcal{L}$ for this model fitting. In view of Definition 4.24 we receive in-sample and out-of-sample Poisson deviance losses

$$\mathfrak{D}(\mathcal{L}, \widehat{\boldsymbol{\beta}}_{\mathcal{L}}^{\text{MLE}}) = \frac{2}{n} \sum_{i=1}^{n} v_i \left(\widehat{\mu}(\boldsymbol{x}_i) - Y_i - Y_i \log\left(\frac{\widehat{\mu}(\boldsymbol{x}_i)}{Y_i} \right) \right) \geq 0,$$

$$\mathfrak{D}(\mathcal{T}, \widehat{\boldsymbol{\beta}}_{\mathcal{L}}^{\text{MLE}}) = \frac{2}{T} \sum_{t=1}^{T} v_t^\dagger \left(\widehat{\mu}(\boldsymbol{x}_t^\dagger) - Y_t^\dagger - Y_t^\dagger \log\left(\frac{\widehat{\mu}(\boldsymbol{x}_t^\dagger)}{Y_t^\dagger} \right) \right) \geq 0.$$

We implement this GLM on the data of Listing 5.1 (and including the categorical features) in R using the function `glm` [307], a short overview of the results is presented in Listing 5.3. This overview presents the regression model implemented, an excerpt of the parameter estimates $\widehat{\boldsymbol{\beta}}_{\mathcal{L}}^{\text{MLE}}$, standard errors which are received from the square-rooted diagonal entries of the inverse of the estimated Fisher's information matrix $\mathcal{I}_n(\widehat{\boldsymbol{\beta}}_{\mathcal{L}}^{\text{MLE}})$, see (5.17); the remaining columns will be described in Sect. 5.3.2 on the Wald test (5.33). The bottom line of the output says that Fisher's scoring algorithm has converged in 6 iterations, it gives the in-sample deviance loss $n\mathfrak{D}(\mathcal{L}, \widehat{\boldsymbol{\beta}}_{\mathcal{L}}^{\text{MLE}})$ called `Residual deviance` (not being scaled by the number of

Listing 5.3 Results in model Poisson GLM1 using the R command `glm`

```
1  Call:
2  glm(formula = ClaimNb ~ VehPowerGLM + VehAgeGLM + DrivAgeGLM +
3                BonusMalusGLM + VehBrand + VehGas + DensityGLM + Region +
4                AreaGLM, family = poisson(), data = learn, offset = log(Exposure))
5
6  Deviance Residuals:
7      Min       1Q   Median       3Q      Max
8  -1.4728  -0.3256  -0.2456  -0.1383   7.7971
9
10 Coefficients:
11                   Estimate Std. Error z value Pr(>!z!)
12 (Intercept)     -4.8175439  0.0579296 -83.162  < 2e-16 ***
13 VehPowerGLM5     0.0604293  0.0229841   2.629 0.008559 **
14 VehPowerGLM6     0.0868252  0.0225509   3.850 0.000118 ***
15 .                       .                       .
16 .                       .                       .
17 RegionR93        0.1388160  0.0294901   4.707 2.51e-06 ***
18 RegionR94        0.1918538  0.0938250   2.045 0.040874 *
19 AreaGLM          0.0407973  0.0200818   2.032 0.042199 *
20 ---
21 Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
22
23 (Dispersion parameter for poisson family taken to be 1)
24
25     Null deviance: 153852  on 610205  degrees of freedom
26 Residual deviance: 147069  on 610157  degrees of freedom
27 AIC: 192818
28
29 Number of Fisher Scoring iterations: 6
```

Table 5.3 Run times, number of parameters, AICs, in-sample and out-of-sample deviance losses, tenfold cross-validation losses with empirical standard deviation in brackets, see also (4.36), (units are in 10^{-2}) and the in-sample average frequency of the null model (Poisson intercept model, see Example 4.27) and of model Poisson GLM1

	Run time	# Param.	AIC	In-sample loss on $\mathcal{L}$	Out-of-sample loss on $\mathcal{T}$	Tenfold CV loss $\widehat{\mathfrak{D}}^{\mathrm{CV}}$	Aver. freq.
Poisson null	–	1	199’506	25.213	25.445	25.213(0.234)	7.36%
Poisson GLM1	16 s	49	192’818	24.101	24.146	24.121(0.245)	7.36%

observations), as well as Akaike’s Information Criterion (AIC), see Sect. 4.2.3 for AIC. Note that we have implemented Poisson version (5.27) with the exposures entering the offset, see lines 2–4 of Listing 5.3; this is important for understanding AIC being calculated on the (unscaled) claim counts N_i.

Table 5.3 summarizes the results of model Poisson GLM1 and it compares the figures to the null model (only having an intercept β_0); the null model has already been introduced in Example 4.27. We present the run time needed to fit the model,[3] the number of regression parameters $q + 1$ in $\boldsymbol{\beta} \in \mathbb{R}^{q+1}$, AIC, in-sample and out-of-sample deviance losses, as well as tenfold cross-validation losses on the

[3] All run times are measured on a personal laptop Intel(R) Core(TM) i7-8550U CPU @ 1.80 GHz 1.99 GHz with 16 GB RAM, and they only correspond to fitting the model (or the corresponding step) once, i.e., they do not account for multiple runs, for instance, for K-fold cross-validation.

learning data $\mathcal{L}$. For tenfold cross-validation we always use the same (non-stratified) partition of $\mathcal{L}$ (in all examples in this monograph), and in bracket we show the empirical standard deviation received by (4.36). Tenfold cross-validation would not be necessary in this case because we have test data $\mathcal{T}$ on which we can evaluate the out-of-sample deviance GL. We present both figures to back-test whether tenfold cross-validation works properly in our example. We observe that the out-of-sample deviance losses $\mathfrak{D}(\mathcal{T}, \widehat{\boldsymbol{\beta}}_{\mathcal{L}}^{\mathrm{MLE}})$ are within one empirical standard deviation of the tenfold cross-validation losses $\widehat{\mathfrak{D}}^{\mathrm{CV}}$, which supports this methodology of model comparison.

From Table 5.3 we conclude that we should prefer model Poisson GLM1 over the null model, this decision is supported by a smaller AIC, a smaller out-of-sample deviance loss $\mathfrak{D}(\mathcal{T}, \widehat{\boldsymbol{\beta}}_{\mathcal{L}}^{\mathrm{MLE}})$ as well as a smaller cross-validation loss $\widehat{\mathfrak{D}}^{\mathrm{CV}}$. The last column of Table 5.3 confirms that the estimated model meets the balance property (we work with the canonical link here). Note that this balance property should be fulfilled for two reasons. Firstly, we would like to have the overall portfolio price on the right level, and secondly, deviance losses should only be compared on the same overall frequency, see Example 4.10.

Before we continue to introduce more models to challenge model Poisson GLM1, we are going to discuss statistical tools for model evaluation. Of course, we would like to know whether model Poisson GLM1 is a good model for this data or whether it is just the better model of two bad options.

Remark 5.15 (Prior and Posterior Information) Pricing literature distinguishes between prior feature information and posterior feature information, see Verschuren [372]. Prior feature information is available at the inception of the (new) insurance contract before having any claims history. This includes, for instance, age of driver, vehicle brand, etc. For policy renewals, past claims history is available and prices of policy renewals can also be based on such posterior information. Past claims history has led to the development of so-called bonus-malus systems (BMS) which often are in the form of multiplicative factors to the base premium to reward and punish good and bad past experience, respectively. One stream of literature studies optimal designs of BMS, we refer to Loimaranta [255], De Pril [91], Lemaire [245], Denuit et al. [102], Brouhns et al. [57] Pinquet [304], Pinquet et al. [305], Tzougas et al. [360] or Ágoston–Gyetvai [4]. Another stream of literature studies how one can optimally extract predictive information from an existing BMS, see Boucher–Inoussa [46], Boucher–Pigeon [47] and Verschuren [372].

The latter is basically what we also do in the above example: note that we include the variable `BonusMalus` into the feature information and, thus, we use past claims information to predict future claims. For new policies, the bonus-malus level is at 100%, and our information does not allow to clearly distinguish between new policies and policy renewals for drivers that have posterior information reflected by a bonus-malus level of 100%. Since young drivers are more likely new customers we expect interactions between the driver's age variable and the bonus-malus level, this intuition is supported by Fig. 13.12 (lhs). In order to improve our model, we would require more detailed information about past claims history. Remark that we do not strictly distinguish between prior and posterior information, here. If we go over to a time-series consideration, where more and more claims experience becomes available of an individual driver, we should clearly distinguish the different sets of information, because otherwise it may happen that in prior and posterior pricing factors we correct twice for the same factor; an interesting paper is Corradin et al. [82].

We also mention that a new source of posterior information is emerging through the collection of telematics car driving data. Telematics car driving data leads to a completely new way of posterior information rate making (experience rating), we refer to Ayuso et al. [17–19], Boucher et al. [42], Lemaire et al. [246] and Denuit et al. [98]. We mention the papers of Gao et al. [152, 154] and Meng et al. [271] who directly extract posterior feature information from telematics car driving data in order to improve rate making. This approach combines a Poisson GLM with a network extractor for the telematics car driving data.

5.3 Model Validation

One of the purposes of Chap. 4 has been to describe measures to analyze how well a fitted model generalizes to unseen data. In a proper generalization analysis this requires learning data $\mathcal{L}$ for in-sample model fitting and a test sample $\mathcal{T}$ for an out-of-sample generalization analysis. In many cases, one is not in the comfortable situation of having a test sample. In such situations one can use AIC that tries to correct the in-sample figure for model complexity or, alternatively, K-fold cross-validation as used in Table 5.3.

The purpose of this section is to introduce diagnostic tools for fitted models; these are often based on unit deviances $\mathfrak{d}(Y_i, \mu_i)$, which play the role of squared residuals in classical linear regression. Moreover, we discuss parameter and model selection, for instance, by step-wise backward elimination or forward selection using the analysis of variance (ANOVA) or the likelihood ratio test (LRT).

5.3.1 Residuals and Dispersion

Within the EDF we distinguish two different types of residuals. The first type of residuals are based on the unit deviances $\mathfrak{d}(Y_i, \mu_i)$ studied in (4.7). The *deviance* *residuals* are given by

$$r_i^{\mathrm{D}} = \mathrm{sign}(Y_i - \mu_i)\sqrt{\frac{v_i}{\varphi}\,\mathfrak{d}\,(Y_i, \mu_i)}.$$

Secondly, *Pearson's residuals* are given by, see also (4.12),

$$r_i^{\mathrm{P}} = \sqrt{\frac{v_i}{\varphi}}\frac{Y_i - \mu_i}{\sqrt{V(\mu_i)}}.$$

In the Gaussian case the two residuals coincide. This indicates that Pearson's residuals are most appropriate in the Gaussian case because they respect the distributional properties in that case. For other distributions, Pearson's residuals can be markedly skewed, as stated in Section 2.4.2 of McCullagh–Nelder [265], and therefore may fail to have properties similar to Gaussian residuals. An other issue occurs in Pearson's residuals when the denominator involves an estimated standard deviation $\sqrt{V(\widehat{\mu}_i)}$, for instance, if we work in a small frequency Poisson problem. Estimation uncertainty in small denominators of Pearson's residuals may substantially distort the estimated residuals. For this reason, we typically work with (the more robust) deviance residuals; this is related to the discussion in Chap. 4 on MSEPs versus expected deviance GLs, see Remarks 4.6.

The squared residuals provide unit deviance and weighted square loss, respectively,

$$(r_i^{\mathrm{D}})^2 = \frac{v_i}{\varphi}\,\mathfrak{d}\,(Y_i, \mu_i) \qquad \text{and} \qquad (r_i^{\mathrm{P}})^2 = \frac{v_i}{\varphi}\frac{(Y_i - \mu_i)^2}{V(\mu_i)},$$

the latter corresponds to Pearson's χ^2-statistic, see (4.12).

Example 5.16 (Residuals in the Poisson Case) In the Poisson case, Pearson's χ^2-statistic is for $v_i = \varphi = 1$ given by

$$(r_i^{\mathrm{P}})^2 = \frac{(Y_i - \mu_i)^2}{\mu_i},$$

because we have variance function $V(\mu) = \mu$. A second order Taylor expansion around Y_i on the scale $\mu_i^{1/3}$ (for μ_i) provides approximation to the unit deviances in the Poisson case, see formula (6.4) and Figure 6.2 in McCullagh–Nelder [265],

$$\mathfrak{d}\,(Y_i, \mu_i) \;\approx\; 9Y_i^{1/3}\left(Y_i^{1/3} - \mu_i^{1/3}\right)^2. \tag{5.29}$$

This emphasizes the different behaviors around the observation Y_i of the two types of residuals in the Poisson case. The scale $\mu_i^{1/3}$ has been motivated in McCullagh–

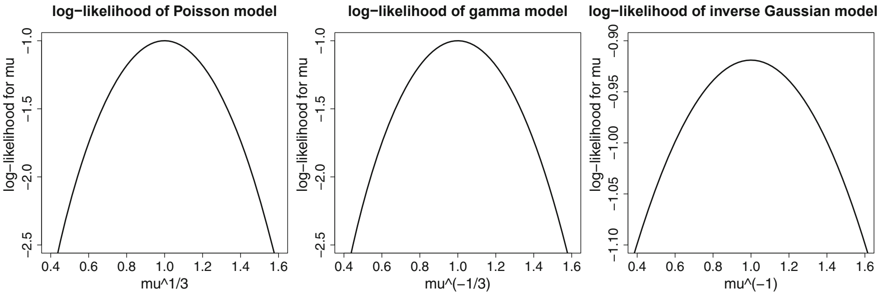


Fig. 5.5 Log-likelihoods $\ell_Y(\mu)$ in $Y = 1$ as a function of μ plotted against (lhs) $\mu^{1/3}$ in the Poisson case, (middle) $\mu^{-1/3}$ in the gamma case with shape parameter $\alpha = 1$, and (rhs) μ^{-1} in the inverse Gaussian case with $\alpha = 1$

Nelder [265] by providing a symmetric behavior around the mode in $Y_i = 1$ of the resulting log-likelihood function, see Fig. 5.5 (lhs).

■

The explicit calculation of the residuals requires knowledge of the dispersion parameter $\varphi > 0$. In the Poisson Example 5.16 this dispersion parameter has been set equal to 1 because the Poisson model does neither allow for under- nor for over-dispersion. Typically, this is not the case for other models, and this requires determination of the dispersion parameter if we want to simulate from these other models. So far, this dispersion parameter has been treated as a nuisance parameter and, in fact, it canceled in MLE (because it was assumed to be constant), see Proposition 5.1.

If we need to estimate the dispersion parameter, we can either do this within MLE, see Remarks 5.2, or we can use Pearson's or the deviance estimates, respectively,

$$\widehat{\varphi}^{\mathrm{P}} = \frac{1}{n-(q+1)} \sum_{i=1}^{n} \frac{(Y_i - \widehat{\mu}_i)^2}{V(\widehat{\mu}_i)/v_i} \quad \text{and} \quad \widehat{\varphi}^{\mathrm{D}} = \frac{1}{n-(q+1)} \sum_{i=1}^{n} v_i \mathfrak{d}\left(Y_i, \widehat{\mu}_i\right), \tag{5.30}$$

where $\widehat{\mu}_i = \widehat{\mu}(\boldsymbol{x}_i)$ are the MLE estimated means involving $q + 1$ estimated parameters $\widehat{\boldsymbol{\beta}}^{\mathrm{MLE}} \in \mathbb{R}^{q+1}$. We briefly motivate these choices. Firstly, Pearson's estimate $\widehat{\varphi}^{\mathrm{P}}$ is consistent for φ. Note that in the Gaussian case this is just the standard estimate for the variance parameter. Justification of the deviance dispersion estimate is more challenging. Consider the unscaled deviance with $\widehat{\boldsymbol{\mu}}_n = (\widehat{\mu}_1, \ldots, \widehat{\mu}_n)^\top$, see (4.9),

$$n\varphi\mathfrak{D}(\boldsymbol{Y}_n, \widehat{\boldsymbol{\mu}}_n) = \sum_{i=1}^{n} v_i \mathfrak{d}\left(Y_i, \widehat{\mu}_i\right).$$

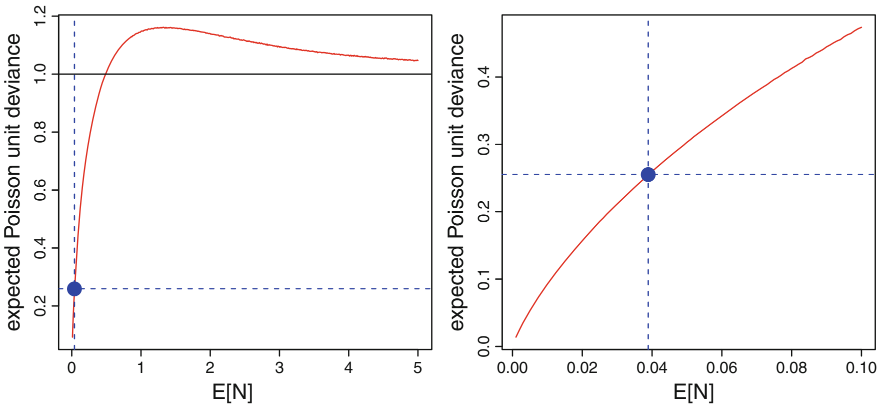


Fig. 5.6 Expected unit deviance $v\mathbb{E}_\mu[\mathfrak{d}(Y,\mu)]$ in the Poisson case as a function of $\mathbb{E}[N] = \mathbb{E}[vY] = v\mu$; the two plots only differ in the scale on the x-axis

This statistic is under *certain* assumptions asymptotically $\varphi\chi^2_{n-(q+1)}$-distributed, where $\chi^2_{n-(q+1)}$ denotes a χ^2-distribution with $n-(q+1)$ degrees of freedom. Thus, this approximation gives us an expected value of $\varphi(n-(q+1))$. This exactly justifies the deviance dispersion estimate (5.30) in these cases. However, as stated in the last paragraph of Section 2.3 of McCullagh–Nelder [265], often a χ^2-approximation is not suitable even as $n \to \infty$. We give an example.

Example 5.17 (Poisson Unit Deviances) The deviance statistics in the Poisson model with means $\boldsymbol{\mu}_n = (\mu_1, \ldots, \mu_n)^\top$ is given by

$$\mathfrak{D}(\boldsymbol{Y}_n, \boldsymbol{\mu}_n) = \frac{1}{n}\sum_{i=1}^n v_i \mathfrak{d}\left(Y_i, \mu_i\right) = \frac{1}{n}\sum_{i=1}^n 2v_i \left(\mu_i - Y_i - Y_i \log\left(\frac{\mu_i}{Y_i}\right)\right),$$

note that in the Poisson model we have (by definition) $\varphi = 1$. We evaluate the expected value of this deviance statistics. It is given by

$$\mathbb{E}_{\boldsymbol{\mu}_n}\left[\mathfrak{D}(\boldsymbol{Y}_n, \boldsymbol{\mu}_n)\right] = \frac{1}{n}\sum_{i=1}^n 2v_i \mathbb{E}_{\mu_i}\left[\mu_i - Y_i - Y_i \log\left(\frac{\mu_i}{Y_i}\right)\right] = \frac{1}{n}\sum_{i=1}^n 2\mathbb{E}_{\mu_i}\left[N_i \log\left(\frac{N_i}{v_i\mu_i}\right)\right],$$

with $N_i \overset{\text{ind.}}{\sim} \text{Poi}(v_i\mu_i)$.

In Fig. 5.6 we plot the expected unit deviance $v\mu \mapsto v\mathbb{E}_\mu[\mathfrak{d}(Y,\mu)]$ in the Poisson model. In our example of Table 5.3, we have $\mathbb{E}_\mu[vY] = v\mu \approx 3.89\%$, which results in an expected unit deviance of $v\mathbb{E}_\mu[\mathfrak{d}(Y,\mu)] \approx 25.52 \cdot 10^{-2} < 1$. This is in line with the losses in Table 5.3. Thus, the expected deviance $n\mathbb{E}_{\boldsymbol{\mu}_n}\left[\mathfrak{D}(\boldsymbol{Y}_n, \boldsymbol{\mu}_n)\right] \approx n/4 < n$. Therefore it is substantially smaller than n. But this implies that $n\mathfrak{D}(\boldsymbol{Y}_n, \boldsymbol{\mu}_n)$ cannot be asymptotically $\chi^2_{n-(q+1)}$-distributed because the latter has an expected of value $n-(q+1) \approx n$ for $n \to \infty$. In fact, the deviance dispersion estimate is not consistent in this example, and for a consistent estimate one should rely on Pearson's deviance estimate.

In order to have an asymptotic χ^2-distribution we need to have large volumes v because then a saddlepoint approximation holds that allows to approximate the (scaled) unit deviances by χ^2-distributions, see Sect. 5.5.2, below. ■

5.3.2 Hypothesis Testing

Consider a sub-vector $\boldsymbol{\beta}_r \in \mathbb{R}^r$ of the GLM parameter $\boldsymbol{\beta} \in \mathbb{R}^{q+1}$, for $r < q+1$. We would like to understand if we can set this sub-vector $\boldsymbol{\beta}_r = 0$, and at the same time we do not lose any generalization power. Thus, we investigate whether there is a simpler *nested* GLM that provides a similar prediction accuracy. If this is the case, preference should be given to the simpler model because the bigger model seems over-parametrized (has redundancy, is not parsimonious). This section is based on Section 2.2.2 of Fahrmeir–Tutz [123].

Geometric Interpretation We begin by giving a geometric interpretation. We start from the full model being expressed by the design matrix $\mathfrak{X} \in \mathbb{R}^{n\times(q+1)}$. This design matrix together with the link function g generates a $(q+1)$-dimensional manifold $\mathfrak{M} \subset \mathbb{R}^n$ given by, see (5.19) and Fig. 5.2,

$$\mathfrak{M} = \left\{ \boldsymbol{\mu} = g^{-1}(\mathfrak{X}\boldsymbol{\beta}) = (g^{-1}\langle\boldsymbol{\beta}, \boldsymbol{x}_1\rangle, \ldots, g^{-1}\langle\boldsymbol{\beta}, \boldsymbol{x}_n\rangle)^\top \in \mathbb{R}^n \middle|\; \boldsymbol{\beta} \in \mathbb{R}^{q+1} \right\} \subset \mathbb{R}^n.$$

The MLE $\widehat{\boldsymbol{\beta}}^{\mathrm{MLE}}$ is determined by the point in $\mathfrak{M}$ that minimizes the distance to $\boldsymbol{Y}$, where distance between $\boldsymbol{Y}$ and $\mathfrak{M}$ is measured component-wise by $\frac{v_i}{\varphi}\mathfrak{d}(Y_i, \mu_i)$ with $\boldsymbol{\mu} \in \mathfrak{M}$, i.e., w.r.t. the KL divergence.

Assume, now, that we want to drop the components $\boldsymbol{\beta}_r$ in $\boldsymbol{\beta}$, i.e., we want to drop these columns from the design matrix resulting in a smaller design matrix $\mathfrak{X}_r \in \mathbb{R}^{n\times(q+1-r)}$. This generates a $(q+1-r)$-dimensional *nested* manifold $\mathfrak{M}_r \subset \mathfrak{M}$ described by

$$\mathfrak{M}_r = \left\{ \boldsymbol{\mu} = g^{-1}(\mathfrak{X}_r\boldsymbol{\beta}) \in \mathbb{R}^n \middle|\; \boldsymbol{\beta} \in \mathbb{R}^{q+1-r} \right\} \subset \mathfrak{M}.$$

If the distance of $\boldsymbol{Y}$ to $\mathfrak{M}_r$ and $\mathfrak{M}$ is roughly the same, we should go for the smaller model. In the Gaussian case of Example 5.9 this can be explained by the Pythagorean theorem applied to successive orthogonal projections. In the general unit deviance case, this has to be studied in terms of information geometry considering the KL divergence, see Sect. 2.3.

Likelihood Ratio Test (LRT) We consider the testing problem of the null hypothesis H_0 against the alternative hypothesis H_1

$$H_0 : \boldsymbol{\beta}_r = 0 \qquad \text{against} \qquad H_1 : \boldsymbol{\beta}_r \neq 0. \tag{5.31}$$

Denote by $\widehat{\boldsymbol{\beta}}^{\text{MLE}}$ the MLE under the full model and by $\widehat{\boldsymbol{\beta}}^{\text{MLE}}_{(-r)}$ the MLE under the null hypothesis model. Define the (log-)*likelihood ratio test (LRT) statistics*

$$\Lambda = -2\left(\ell_{\boldsymbol{Y}}(\widehat{\boldsymbol{\beta}}^{\text{MLE}}_{(-r)}) - \ell_{\boldsymbol{Y}}(\widehat{\boldsymbol{\beta}}^{\text{MLE}})\right) \;\geq\; 0.$$

The inequality holds because the null hypothesis model is nested in the full model, henceforth, the latter needs to have a bigger log-likelihood value in the MLE. If the LRT statistics Λ is large, the null hypothesis should be rejected because the reduced model is not competitive compared to the full model. More mathematically, under similar conditions as for the asymptotic normality results of the MLE of $\boldsymbol{\beta}$ in (5.17), we have that under the null hypothesis H_0 the LRT statistics Λ is asymptotically χ^2-distributed with r degrees of freedom. Therefore, we should reject the null hypothesis in favor of the full model if the resulting p-value of Λ under the χ^2_r-distribution is too small. These results remain true if the unknown dispersion parameter φ is replaced by a consistent estimator $\widehat{\varphi}$, e.g., Pearson's dispersion estimate $\widehat{\varphi}^{\text{P}}$ (from the bigger model).

The LRT statistics Λ may not be properly defined in over-dispersed situations where the distributional assumptions are not fully specified, for instance, in an over-dispersed Poisson model. In such situations, one usually divides the log-likelihood (of the Poisson model) by the estimated over-dispersion and then uses the resulting scaled LRT statistics Λ as an approximation to the unspecified model.

Wald Test Alternatively, we can use the Wald statistics. The Wald statistics uses a second order approximation to the log-likelihood and, therefore, is only based on the first two moments (and not on the entire distribution). Define the matrix $I_r \in \mathbb{R}^{r\times(q+1)}$ such that $\boldsymbol{\beta}_r = I_r\boldsymbol{\beta}$, i.e., matrix I_r selects exactly the components of $\boldsymbol{\beta}$ that are included in $\boldsymbol{\beta}_r$ (and which are set to 0 under the null hypothesis H_0 given in (5.31)).

Asymptotic normality (5.17) motivates consideration of the Wald statistics

$$W = (I_r\widehat{\boldsymbol{\beta}}^{\text{MLE}} - 0)^\top \left(I_r\mathcal{I}(\widehat{\boldsymbol{\beta}}^{\text{MLE}})^{-1}I_r^\top\right)^{-1}(I_r\widehat{\boldsymbol{\beta}}^{\text{MLE}} - 0). \tag{5.32}$$

The Wald statistics measures the distance between the MLE in the full model $I_r\widehat{\boldsymbol{\beta}}^{\text{MLE}}$ restricted to the components of $\boldsymbol{\beta}_r$ and the null hypothesis H_0 (being $\boldsymbol{\beta}_r = 0$). The estimated Fisher's information matrix $\mathcal{I}(\widehat{\boldsymbol{\beta}}^{\text{MLE}})$ is used to bring all components onto the same unit scale (and to account for collinearity). The Wald statistics W is asymptotically χ^2_r-distributed under the same assumptions as for (5.17) to hold. Thus, the null hypothesis H_0 should be rejected if the resulting p-

value of W under the χ_r^2-distribution is too small. Note that this test does not require calculation of the MLE in the null hypothesis model, i.e., this test is computationally more attractive than the LRT because we only need to fit one model. Again, an unknown dispersion parameter φ in Fisher's information matrix $\mathcal{I}(\boldsymbol{\beta})$ is replaced by a consistent estimator $\widehat{\varphi}$ (from the bigger model).

In the special case of considering only one component of $\boldsymbol{\beta}$, i.e., if $\boldsymbol{\beta}_r = \beta_k$ with $r = 1$ and for one selected component $0 \le k \le q$, the Wald statistics reduces to

$$W_k = \frac{(\widehat{\beta}_k^{\mathrm{MLE}})^2}{\widehat{\sigma}_k^2} \qquad \text{or} \qquad T_k = W_k^{1/2} = \frac{\widehat{\beta}_k^{\mathrm{MLE}}}{\widehat{\sigma}_k}, \tag{5.33}$$

with diagonal entries of the inverse of the estimated Fisher's information matrix given by $\widehat{\sigma}_k^2 = (\mathcal{I}(\widehat{\boldsymbol{\beta}}^{\mathrm{MLE}})^{-1})_{k,k}$, $0 \le k \le q$. The square-roots of these estimates are provided in column `Std. Error` of the R output in Listing 5.3.

In this case the Wald statistics W_k is equal to the square of the t-statistics T_k; this t-statistics is provided in column `z value` of the R output of Listing 5.3. Remark that Fisher's information matrix involves the dispersion parameter φ. If this dispersion parameter is estimated with a consistent estimator $\widehat{\varphi}$ we have a t-statistics. For known dispersion parameter the t-statistics reduces to a z-statistics, i.e., the corresponding p-values can be calculated from a normal distribution instead of a t-distribution. In the Poisson case, the dispersion $\varphi = 1$ is known, and for this reason, we perform a z-test (and not a t-test) in the last column of Listing 5.3; and we call T_k a z-statistics in that case.

5.3.3 Analysis of Variance

In the previous section, we have presented tests that allow for model selection in the case of nested models. More generally, if we have a full model, say, based on regression parameter $\boldsymbol{\beta} \in \mathbb{R}^{q+1}$ we would like to select the "best" sub-model according to some selection criterion. In most cases, it is computationally not feasible to fit all sub-models if q is large, therefore, this is not a practical solution. For large models and data sets step-wise procedures are a feasible tool. *Backward elimination* starts from the full model, and then recursively drops feature components which have high p-values in the corresponding Wald statistics (5.32) and (5.33). Performing this recursively will provide us with hierarchy of nested models. *Forward selection* works just in the opposite direction, that is, we start with the null model and we include feature components one after the other that have a low p-value in the corresponding Wald statistics.

Remarks 5.18

- The order of the inclusion/exclusion of the feature components matters in this selection algorithms because we do not have additivity in this selection process. For this reason, often backward elimination and forward selection is combined in an alternating way.
- This process as well as the tests from Sect. 5.3.2 are based on a fixed pre-processing of features. If the feature pre-processing is done differently, all analysis needs to be repeated for this new model. Moreover, between two different models we need to apply different tools for model selection (if they are not nested), for instance, AIC, cross-validation or an out-of-sample generalization analysis.
- For categorical variables with dummy coding we should apply the forward selection or the backward elimination simultaneously on the entire dummy coded vector of a categorical variable. This will include or exclude this variable; if we only apply the Wald test to one component of the dummy vector, then we test whether this level should be merged with the reference level.

Typically, in practice, a so-called analysis of variance (ANOVA) table is studied. The ANOVA table is mainly motivated by the Gaussian model with orthogonal data. The Gaussian assumption implies that the deviance loss is equal to the square loss and the orthogonality implies that the square loss decouples in an additive way w.r.t. the feature components. This implies that one can explicitly study the contribution of each feature component to the decrease in square loss; an example is given in Section 2.3.2 of McCullagh–Nelder [265]. In non-Gaussian and non-orthogonal situations one loses this additivity property and, as mentioned in Remarks 5.18, the order of inclusion matters. Therefore, for the ANOVA table we pre-specify the order in which the components are included and then we analyze the decrease of deviance loss by the inclusion of additional components.

Example 5.19 (Poisson GLM1, Revisited) We revisit the MTPL claim frequency example of Sect. 5.2.4 to illustrate the variable selection procedures. Based on the model presented in Listing 5.3 we run an ANOVA analysis using the R command `anova`, the results are presented in Listing 5.4.

Listing 5.4 shows the hierarchy of models starting from the null model by sequentially including feature components one by one. The column `Df` gives the number of regression parameters involved and the column `Deviance` the decrease of deviance loss by the inclusion of this feature component. The biggest model improvements are provided by the bonus-malus level and driver's age, this is not surprising in view of the empirical analysis in Figs. 5.3 and 5.4, and in Chap. 13.1. At the other end we have the `Area` code which only seems to improve the model marginally. However, this does not imply, yet, that this variable should be dropped. There are two points that need to be considered: (1) maybe feature pre-processing of `Area` has not been done in an appropriate way and the variable is not in the right functional form for the chosen link function; and (2) `Area` is the last variable included in the model in Listing 5.4 and, maybe, there are already other variables

Listing 5.4 ANOVA table of model Poisson GLM1

```
Analysis of Deviance Table

Model: poisson, link: log

Response: ClaimNb

Terms added sequentially (first to last)


              Df Deviance Resid. Df Resid. Dev
NULL                         610205     153852
VehPowerGLM    5     73.7    610200     153779
VehAgeGLM      2    179.7    610198     153599
DrivAgeGLM     6   1199.4    610192     152400
BonusMalusGLM  1   4300.6    610191     148099
VehBrand      10    240.3    610181     147859
VehGas         1     82.4    610180     147776
DensityGLM     1    512.1    610179     147264
Region        21    191.3    610158     147073
AreaGLM        1      4.1    610157     147069
```

that take over the role of `Area` in smaller models which is possible if we have correlations between the feature components. In our data, `Area` and `Density` are highly correlated. For this reason, we exchange the order of these two components and run the same analysis again, we call this model Poisson GLM1B (which of course provides the same predictive model as Poisson GLM1).

Listing 5.5 ANOVA table of model Poisson GLM1B

```
Analysis of Deviance Table

Model: poisson, link: log

Response: ClaimNb

Terms added sequentially (first to last)


              Df Deviance Resid. Df Resid. Dev
NULL                         610205     153852
VehPowerGLM    5     73.7    610200     153779
VehAgeGLM      2    179.7    610198     153599
DrivAgeGLM     6   1199.4    610192     152400
BonusMalusGLM  1   4300.6    610191     148099
VehBrand      10    240.3    610181     147859
VehGas         1     82.4    610180     147776
AreaGLM        1    505.0    610179     147271
Region        21    192.4    610158     147079
DensityGLM     1     10.1    610157     147069
```

Listing 5.5 shows the ANOVA table if we exchange the order of these two variables. We observe that the magnitudes of the decrease of the deviance loss has switched between the two variables. Overall, `Density` seems slightly more predictive, and we may consider dropping `Area` from the model, also because the correlation between `Density` and `Area` is very high.

If we want to perform backward elimination (sequentially drop one variable after the other) we can use the R command `drop1`. For small models this is doable, for larger models it is computationally demanding.

Listing 5.6 `drop1` analysis of model Poisson GLM1

```
Single term deletions

Model:
ClaimNb ~ VehPowerGLM + VehAgeGLM + DrivAgeGLM + BonusMalusGLM +
    VehBrand + VehGas + DensityGLM + Region + AreaGLM
              Df Deviance    AIC    LRT  Pr(>Chi)
<none>              147069 192818
VehPowerGLM    5   147152 192892   83.4 < 2.2e-16 ***
VehAgeGLM      2   147283 193028  214.1 < 2.2e-16 ***
DrivAgeGLM     6   147603 193341  534.5 < 2.2e-16 ***
BonusMalusGLM  1   150970 196718 3901.5 < 2.2e-16 ***
VehBrand      10   147298 193027  228.9 < 2.2e-16 ***
VehGas         1   147213 192961  144.5 < 2.2e-16 ***
DensityGLM     1   147079 192826   10.1  0.001459 **
Region        21   147259 192967  190.7 < 2.2e-16 ***
AreaGLM        1   147073 192820    4.1  0.042180 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

In Listing 5.6 we present the results of this `drop1` analysis. Both, according to AIC and according to the LRT, we should keep all variables in the model. Again, `Area` and `Density` provide the smallest LRT statistics Λ which illustrates the high collinearity between these two variables (note that the values in Listing 5.6 are identical to the ones in Listings 5.4 and 5.5, respectively).

We conclude that in model Poisson GLM1 we should keep all feature components, and a model improvement can only be obtained by a different feature pre-processing, by a different regression function or by a different distributional model. ■

5.3.4 Lab: Poisson GLM for Car Insurance Frequencies, Revisited

Continuous Coding of Non-monotone Feature Components

We revisit model Poisson GLM1 studied in Sect. 5.2.4 for MTPL claim frequency modeling, and we consider additional competing models by using different feature pre-processing. From Example 5.19, above, we conclude that we should keep all variables in the model if we work with model Poisson GLM1.

Table 5.4 Contingency table of observed number of policies against predicted number of policies with given claim counts ClaimNb

	Numbers of claims ClaimNb					
	0	1	2	3	4	5
Observed number of policies	587'772	21'198	1'174	57	4	1
Predicted number of policies	587'325	22'064	779	34	3	0.3

We calculate Pearson's dispersion estimate which provides $\widehat{\varphi}^{\mathrm{P}} = 1.6697 > 1$. This indicates that the model is not fully suitable for our data because in a Poisson model the dispersion parameter should be equal to 1. There may be two reasons for this over-dispersion: (1) the Poisson assumption is not appropriate because, for instance, the tail of the observations is more heavy-tailed, or (2) the Poisson assumption is appropriate but the regression function has not been chosen in a fully suitable way (maybe also due to missing feature information).

We believe that in our example the observed over-dispersion is a mixture of the two reasons (1) and (2). Surely, the regression structure can be improved since our feature pre-processing is non-optimal and since the chosen regression function only considers multiplicative interactions between the feature components (we have chosen the log-link regression function without adding interaction terms to the regression function).

Table 5.4 gives a contingency table. We observe that we have much more policies with more than 1 claim compared to what is predicted by the fitted model. As a result, a χ^2-test rejects this Poisson model because the resulting p-value is close to 0.

In our data, we have a rather large number of policies with short exposures v_i, and further analysis suggests that these short exposures are not suitably modeled. We will not invest more time into improving the exposure modeling. As mentioned in the appendix, there seem to be a couple of issues how the exposures are displayed and how policy renewals are accounted for in this data. However, it is difficult (almost impossible) to clean the data for better exposure measures without more detailed information about the data collection process.

Our next aim is to model continuous feature components differently, if their raw form does not match the linear predictor assumption. In Poisson GLM1 we have categorized such components and then used dummy coding for the resulting classes, see Sect. 5.2.4. Alternatively, we can use different functional forms, for instance, we can use for DrivAge the following pre-processing

$$\texttt{DrivAge} \mapsto \beta_l \, \texttt{DrivAge} + \beta_{l+1} \log(\texttt{DrivAge}) + \sum_{j=2}^{4} \beta_{l+j} (\texttt{DrivAge})^j. \tag{5.34}$$

Table 5.5 Run times, number of parameters, AICs, in-sample and out-of-sample deviance losses, tenfold cross-validation losses (units are in 10^{-2}) and in-sample average frequency of the null model (intercept model) and of different Poisson GLMs

	Run time	# Param.	AIC	In-sample loss on $\mathcal{L}$	Out-of-sample loss on $\mathcal{T}$	Tenfold CV loss $\widehat{\mathfrak{D}}^{\text{CV}}$	Aver. freq.
Poisson null	–	1	199'506	25.213	25.445	25.213	7.36%
Poisson GLM1	16s	49	192'818	24.101	24.146	24.121	7.36%
Poisson GLM2	15s	48	192'753	24.091	24.113	24.110	7.36%
Poisson GLM3	15s	50	192'716	24.084	24.102	24.104	7.36%

This replaces the $K = 7$ categorical age classes of model Poisson GLM1 by 5 continuous functions of the variable `DrivAge`, and the number of regression parameters is reduced from $K - 1 = 6$ to 5. We call this model Poisson GLM2.

Besides improving the modeling of the feature components we can also start to add interactions beyond the multiplicative ones. For instance, Fig. 13.12 in Chap. 13 may indicate that there is an interaction term between `BonusMalus` and `DrivAge`. New young drivers enter the bonus-malus system at level 100, and it takes some years free of accidents to reach the lowest bonus-malus level of 50. Whereas for senior drivers a bonus-malus level of 100 may indicate that they have had a bad claim experience because otherwise they would be on the lowest bonus-malus level, see also Remark 5.15. We are adding the following interaction to Poisson GLM2 and we call the resulting model Poisson GLM3

$$\beta_{l'}\, \texttt{BonusMalus} \cdot \texttt{DrivAge} + \beta_{l'+1} \texttt{BonusMalus} \cdot (\texttt{DrivAge})^2. \tag{5.35}$$

From Table 5.5 we observe that this leads to a further small model improvement. We mention that this model improvement can also be observed in a decrease of Pearson's dispersion estimate to $\widehat{\varphi}^P = 1.6644$. Noteworthy, all model selection criteria AIC, out-of-sample generalization loss and cross-validation come to the same conclusion in this example.

The tedious task of the modeler now is to find all these systematic effects and bring them in an appropriate form into the model. Here, this is still possible because we have a comparably small model. However, if we have hundreds of feature components, such a manual analysis becomes intractable. Other regression models such as network regression models should be preferred, or at least should be used to find systematic effects. But, one should also keep in mind that the (final) chosen model should be as simple as possible (parsimonious).

Remarks 5.20

- An advantage of GLMs is that these regression models can deal with collinearity in feature components. Nevertheless, the results should be carefully checked if the collinearity in feature components is very high. If we have a high collinearity between two feature components then we may observe large values with opposite signs in the corresponding regression parameters compensating each other. The

Listing 5.7 `drop1` analysis of model Poisson GLM2

```
Single term deletions

Model:
ClaimNb ~ VehPowerGLM + VehAgeGLM + DrivAge + log(DrivAge) +
    I(DrivAge^2) + I(DrivAge^3) + I(DrivAge^4) + BonusMalusGLM +
    VehBrand + VehGas + DensityGLM + Region + AreaGLM
              Df Deviance    AIC    LRT  Pr(>Chi)
<none>            147005 192753
VehPowerGLM    5   147087 192825   82.4 2.671e-16 ***
VehAgeGLM      2   147225 192969  220.3 < 2.2e-16 ***
DrivAge        1   147157 192902  151.9 < 2.2e-16 ***
log(DrivAge)   1   147190 192935  184.8 < 2.2e-16 ***
I(DrivAge^2)   1   147123 192869  118.1 < 2.2e-16 ***
I(DrivAge^3)   1   147094 192840   89.0 < 2.2e-16 ***
I(DrivAge^4)   1   147071 192816   65.5 5.687e-16 ***
BonusMalusGLM  1   150907 196653 3902.0 < 2.2e-16 ***
VehBrand      10   147232 192959  226.5 < 2.2e-16 ***
VehGas         1   147148 192893  142.8 < 2.2e-16 ***
DensityGLM     1   147015 192761   10.1  0.001498 **
Region        21   147193 192899  188.0 < 2.2e-16 ***
AreaGLM        1   147009 192755    4.1  0.043123 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

resulting GLM will not be very robust, and a slight change in the observations may change these regression parameters completely. In this case one should drop one of the two highly collinear feature components. This problem may also occur if we include too many terms in functional forms like in (5.34).

- A tool to find suitable functional forms of regression functions in continuous feature components are the partial residual plots of Cook–Croos-Dabrera [80]. If we want to analyze the first feature component x_1 of $\boldsymbol{x}$, we can fit a GLM to the data using the entire feature vector $\boldsymbol{x}$. The partial residuals for component x_1 are defined by, see formula (8) in Cook–Croos-Dabrera [80],

$$r_i^{\text{partial}} = (Y_i - \mu(\boldsymbol{x}_i))g'(\mu(\boldsymbol{x}_i)) + \beta_1 x_{i,1} \qquad \text{for } 1 \le i \le n,$$

  where g is the chosen link function and $g(\mu(\boldsymbol{x}_i)) = \langle \boldsymbol{\beta}, \boldsymbol{x}_i \rangle$. These partial residuals offset the effect of feature component x_1. The partial residual plot shows r_i^{partial} against $x_{i,1}$. If this plot shows a linear structure then including x_1 linearly is justified, and any other functional form may be detected from that plot.

Under-Sampling and Over-Sampling

Often run times are an issue in model fitting, in particular, if we want to experiment with different models, different feature codings, etc. Under-sampling is an interesting approach that can be applied in imbalanced situations (like in our claim frequency data situation) to speed up calculations, and still receiving accurate approximations. We briefly describe under-sampling in this subsection.

Under-sampling is based on the idea that we do not need to consider all $n = 610'206$ insurance policies for model fitting, and we can still receive accurate results. For this we select all insurance policies that have at least 1 claim; in our data these are 22'434 insurance policies, we call this data set $\mathcal{L}^*_{\geq 1}$. The motivation for selecting these insurance policies is that these are exactly the policies that have information about the drivers causing claims. These selected insurance policies need to be complemented with policies that do not cause any claims. We select at random (under-sample) 22'434 insurance policies of drivers without claims, we call this data set $\mathcal{L}^*_0$. Merging the two sets we receive data $\mathcal{L}^* = \mathcal{L}^*_0 \cup \mathcal{L}^*_{\geq 1}$ comprising 44'868 insurance policies. This data is balanced from the viewpoint of claim causing policies because exactly half of the policies in $\mathcal{L}^*$ suffers a claim and the other half does not. The idea now is to fit a GLM only on this learning data $\mathcal{L}^*$, and because we only consider 44'868 insurance policies the fitting should be fast.

There is still one point to be considered, namely, in the new learning data $\mathcal{L}^*$ policies with claims are over-represented (because we work in a low frequency problem). This motivates that we adjust the time exposures v_i in $\mathcal{L}^*_0$ accordingly by multiplying as follows

$$v_i \;\mapsto\; v_i^* = v_i \, \frac{\sum_{j=1}^n v_j \mathbb{1}_{\{N_j=0\}}}{\sum_{v_j \in \mathcal{L}^*_0} v_j}.$$

Thus, we stretch the exposures of the policies without claims in $\mathcal{L}^*$; for our data this factor is 26.17. This then provides us with an empirical frequency on $\mathcal{L}^*$ of 7.36% which is identical to the observed frequency on the entire learning data $\mathcal{L}$.

We fit model Poisson GLM3 on this reduced (and exposure adjusted) learning data $\mathcal{L}^*$, the results are presented on the last line of Table 5.6. This model can be fitted in 1s, and by construction it fulfills the balance property. The resulting in-sample and out-of-sample losses (evaluated on the entire data $\mathcal{L}$ and $\mathcal{T}$) are very close to model Poisson GLM3 which verifies that the model fitted only on the learning data $\mathcal{L}^*$ gives a good approximation. We do not provide AIC because the data used is not identical to the data used to fit the other models. The tenfold cross-

Table 5.6 Run times, number of parameters, AICs, in-sample and out-of-sample deviance losses, tenfold cross-validation losses (units are in 10^{-2}) and in-sample average frequency of the null model (intercept model) and of different Poisson GLMs, the last row uses under-sampling in model Poisson GLM3

	Run time	# param.	AIC	In-sample loss on $\mathcal{L}$	Out-of-sample loss on $\mathcal{T}$	Tenfold CV loss $\widehat{\mathfrak{D}}^{\mathrm{CV}}$	Aver. freq.
Poisson null	–	1	199'506	25.213	25.445	25.213	7.36%
Poisson GLM1	16 s	49	192'818	24.101	24.146	24.121	7.36%
Poisson GLM2	15 s	48	192'753	24.091	24.113	24.110	7.36%
Poisson GLM3	15 s	50	192'716	24.084	24.102	24.104	7.36%
under-sampling	1 s	50	–	24.098	24.108	24.120	7.36%

validation loss is a little bit bigger which seems to be a consequence of applying the non-stratified version to only 44’868 insurance policies, i.e., this higher cross-validation loss shows that we fit the model on less data which provides higher uncertainty in model fitting. This finishes this example.

The presented method is called under-sampling because we under-sample from the insurance policies without claims to make both classes (policies with claims and policies without claims) equally large. Alternatively, to achieve a class balance we could also over-sample from the minority class by duplicating policies. This has a similar effect, but it increases run times. Importantly, if we under- or over-sample we *have* to adjust the exposures correspondingly. Otherwise we obtain a biased model that is not useful for pricing, the same applies to methods such as the synthetic minority oversampling technique (SMOTE) and similar techniques.

Alternatively, to under-sampling we could also fit a so-called zero-truncated Poisson (ZTP) model to the data by only fitting a model on the insurance policies that suffer at least one claim, and adjusting the distribution to the observations $N_i|_{\{N_i \geq 1\}}$. This is rather similar to a hurdle Poisson model and we come back to this in Example 6.19, below.

5.3.5 Over-Dispersion in Claim Counts Modeling

Mixed Poisson Distribution

In the previous example we have seen that the considered Poisson GLMs do not fully fit our data, at least not with the chosen feature engineering, because there is over-dispersion in the data (relative to the chosen models). This may give rise to consider models that allow for over-dispersion. Typically, such over-dispersed models are constructed starting from the Poisson model, because the Poisson model enjoys many nice properties as we have seen above. A natural extension is to introduce the family of mixed Poisson models, where the frequency is not modeled with a single parameter but rather with a whole family of parameters described by an underlying mixing distribution.

In the dual mean parametrization the Poisson distribution for $Y = N/v$ reads as

$$Y \sim f(y; \lambda, v) = e^{-v\lambda} \frac{(v\lambda)^{vy}}{(vy)!} \qquad \text{for } y \in \mathbb{N}_0/v,$$

where the mean parameter is given by $\lambda = \kappa'(\theta) = \exp\{\theta\}$, and θ denotes the canonical parameter; on purpose we use for the mean notation λ instead of μ, here, the reason will become clear below. This model satisfies for the first two moments of $N = vY$

$$\mathbb{E}_\lambda[N] = v\kappa'(\theta) = v\lambda \qquad \text{and} \qquad \mathrm{Var}_\lambda(N) = v\kappa''(\theta) = v\lambda = \mathbb{E}_\lambda[N],$$

with dispersion parameter $\varphi = 1$. A mixed Poisson distribution is obtained by mixing/integrating over different frequency parameters $\lambda > 0$. We choose a distribution π on $\mathbb{R}_+$ (strictly positively supported), and define the new distribution

$$Y = N/v \;\sim\; f_\pi(y; v) = \int_{\mathbb{R}_+} f(y; \lambda, v)\, d\pi(\lambda) = \int_{\mathbb{R}_+} e^{-v\lambda} \frac{(v\lambda)^{vy}}{(vy)!}\, d\pi(\lambda). \tag{5.36}$$

If π is not concentrated in a single point, the tower property immediately implies

$$\mathbb{E}_\pi[N] < \mathrm{Var}_\pi(N), \tag{5.37}$$

supposed that the moments exist, we refer to Lemma 2.18 in Wüthrich [387]. Hence, mixing over different frequency parameters allows us to receive over-dispersion. Of course, this concept can also be applied to mixing over the canonical parameter θ in the EF (instead of the mean parameter).

This leads to the framework of Bayesian credibility models which are widely used and studied in actuarial science, we refer to the textbook of Bühlmann–Gisler [58]. We have already met this idea in the Bayesian decision rule of Example 3.3 which has led to the Bayesian estimator in Definition 3.6.

Negative-Binomial Model

In the case of the Poisson model, the gamma distribution is a particularly attractive mixing distribution for λ because it allows for a closed-form solution in (5.36), and $f_{\pi=\Gamma}(y; v)$ will be a negative-binomial distribution.[4] One can choose different parametrizations of this mixing distribution, and they will provide different scalings in the resulting negative-binomial distribution. We choose the following parametrization $\pi(\lambda) \overset{(d)}{=} \Gamma(v\alpha, v\alpha/\mu)$ for mean parameter $\mu > 0$ and shape parameter $v\alpha > 0$. This implies, see (5.36),

$$\begin{aligned} f_{\mathrm{NB}}(y; \mu, v, \alpha) &= \int_{\mathbb{R}_+} e^{-v\lambda} \frac{(v\lambda)^{vy}}{(vy)!} \frac{(v\alpha/\mu)^{v\alpha}}{\Gamma(v\alpha)} \lambda^{v\alpha-1} e^{-v\alpha\lambda/\mu} d\lambda \\ &= \frac{\Gamma(vy + v\alpha)}{(vy)!\Gamma(v\alpha)} \frac{v^{vy}(v\alpha/\mu)^{v\alpha}}{(v + v\alpha/\mu)^{vy+v\alpha}} \\ &= \binom{vy + v\alpha - 1}{vy} \left(e^{\theta}\right)^{vy} \left(1 - e^{\theta}\right)^{v\alpha}, \end{aligned}$$

[4] The gamma distribution is the conjugate prior to the Poisson distribution. As a result, the posterior distribution, given observations, will again be a gamma distribution with posterior parameters, see Section 8.1 of Wüthrich [387]. This Bayesian model has been introduced to the actuarial literature by Bichsel [32].

setting for canonical parameter $\theta = \log(\mu/(\mu + \alpha)) < 0$. This is the negative-binomial distribution we have already met in (2.5). A single-parameter linear EDF representation is given by, we set unit dispersion parameter $\varphi = 1$,

$$Y \sim f_{\rm NB}(y; \theta, v, \alpha) = \exp\left\{ \frac{y\theta + \alpha \log(1 - e^{\theta})}{1/v} + \log \binom{vy + v\alpha - 1}{vy} \right\}, \tag{5.38}$$

where this is a density w.r.t. the counting measure on $\mathbb{N}_0/v$. The cumulant function and the canonical link, respectively, are given by

$$\kappa(\theta) = -\alpha \log(1 - e^{\theta}) \quad \text{and} \quad \theta = h(\mu) = \log\left(\frac{\mu}{\mu + \alpha}\right) \in \boldsymbol{\Theta} = (-\infty, 0).$$

Note that $\alpha > 0$ is treated as nuisance parameter (which is a fixed part of the cumulant function, here). The first two moments of the claim count $N = vY$ are given by

$$v\mu = \mathbb{E}_{\theta}[N] \; = \; v\alpha \frac{e^{\theta}}{1 - e^{\theta}}, \tag{5.39}$$

$$\mathrm{Var}_{\theta}(N) = \mathbb{E}_{\theta}[N] \left(1 + \frac{e^{\theta}}{1 - e^{\theta}}\right) = \mathbb{E}_{\theta}[N] \left(1 + \frac{\mu}{\alpha}\right) \; > \; \mathbb{E}_{\theta}[N]. \tag{5.40}$$

This shows that we receive a fixed over-dispersion of size μ/α, which (in this parametrization) does not depend on the exposure v; this is the reason for choosing a mixing distribution $\pi(\lambda) \overset{(d)}{=} \Gamma(v\alpha, v\alpha/\mu)$. This parametrization is called NB2 parametrization.

Remarks 5.21

- We emphasize that the effective domain $\boldsymbol{\Theta} = (-\infty, 0)$ is one-sided bounded. Therefore, the canonical link for the linear predictor will not work in general because the linear predictor $\boldsymbol{x} \mapsto \eta(\boldsymbol{x})$ can be both-sided unbounded in a GLM setting. Instead, we use the log-link for $g(\cdot)$ in our example below, with the downside that one loses the balance property.
- The unit deviance in this negative-binomial EDF model is given by

$$(y, \mu) \; \mapsto \; \mathfrak{d}(y, \mu) = 2\left[y \log\left(\frac{y}{\mu}\right) - (y + \alpha) \log\left(\frac{y + \alpha}{\mu + \alpha}\right) \right],$$

  we also refer to Table 4.1 for $\alpha = 1$. We emphasize that this is the unit deviance in a single-parameter linear EDF, and we only aim at estimating canonical parameter $\theta \in \boldsymbol{\Theta}$ and mean parameter $\mu \in \mathcal{M}$, respectively, whereas $\alpha > 0$ is treated as a given nuisance parameter. This is important because the unit deviance relies on the saturated model which, in general, estimates a one-dimensional parameter θ and μ, respectively, from the one-dimensional observation Y. The nuisance parameter is not affected by the consideration of the saturated model, and it is treated as a fixed part of the cumulant function, which is not estimated at this stage. An important consequence of this is that model comparison using deviance residuals only works for identical nuisance parameters.

- We mention that we receive over-dispersion in (5.40) though we have dispersion parameter $\varphi = 1$ in (5.38). Alternatively, we could do the duality transformation $y \mapsto \widetilde{y} = y/\alpha$ for nuisance parameter $\alpha > 0$; this gives the reproductive form of the negative-binomial model NB2, see also Remarks 2.13. This provides us with a density on $\mathbb{N}_0/(v\alpha)$, set $\widetilde{\varphi} = 1/\alpha$,

$$\widetilde{Y} \sim f_{\mathrm{NB}}(\widetilde{y}; \theta, v/\widetilde{\varphi}) = \exp\left\{ \frac{\widetilde{y}\theta + \log(1-e^{\theta})}{1/(v\alpha)} + \log\binom{v\alpha\widetilde{y} + v\alpha - 1}{v\alpha\widetilde{y}} \right\}.$$

The cumulant function and the canonical link, respectively, are now given by

$$\kappa(\theta) = -\log(1-e^{\theta}) \quad \text{and} \quad \theta = h(\widetilde{\mu}) = \log\left(\frac{\widetilde{\mu}}{\widetilde{\mu}+1}\right) \in \boldsymbol{\Theta} = (-\infty, 0).$$

The first two moments are for $\theta \in \boldsymbol{\Theta}$ given by

$$\widetilde{\mu} = \mathbb{E}_{\theta}[\widetilde{Y}] = \frac{e^{\theta}}{1-e^{\theta}},$$

$$\mathrm{Var}_{\theta}(\widetilde{Y}) = \frac{\widetilde{\varphi}}{v}\kappa''(\theta) = \frac{1}{v\alpha}\widetilde{\mu}(1+\widetilde{\mu}).$$

Thus, we receive the reproductive EDF representation with dispersion parameter $\widetilde{\varphi} = 1/\alpha$ and variance function $V(\widetilde{\mu}) = \widetilde{\mu}(1+\widetilde{\mu})$. Moreover, $N = vY = v\alpha\widetilde{Y}$.

- The negative-binomial model with the NB1 parametrization uses the mixing distribution $\pi(\lambda) \overset{(d)}{=} \Gamma(\mu v/\alpha, v/\alpha)$. This leads to mean $\mathbb{E}_{\theta}[N] = v\mu$ and variance $\mathrm{Var}_{\theta}(N) = \mathbb{E}_{\theta}[N](1+\alpha)$. In this parametrization, μ enters the gamma function as $\Gamma(\mu v/\alpha)$ in the gamma density which does not allow for an EDF representation. This parametrization has been called NB1 by Cameron–Trivedi [63] because both terms in the variance $\mathrm{Var}_{\theta}(N) = v\mu + v\mu\alpha$ are linear in μ. In contrast, in the NB2 parametrization the second term has a square $v\mu^2/\alpha$ in μ, see (5.40). Further discussion is provided in Greene [171].

Nuisance Parameter Estimation

All previous statements have been based on the assumption that $\alpha > 0$ is a *given* nuisance parameter. If α needs to be estimated too, then, we drop out of the EF. In this case, an iterative estimation procedure is applied to the EDF representation (5.38). One starts with a fixed nuisance parameter $\alpha^{(0)}$ and fits the negative-binomial GLM with MLE which provides a first set of MLE $\widehat{\boldsymbol{\beta}}^{(1)} = \widehat{\boldsymbol{\beta}}^{(1)}(\alpha^{(0)})$. Based on this estimate the nuisance parameter is updated $\alpha^{(0)} \mapsto \alpha^{(1)}$ by maximizing the log-likelihood in α for given $\widehat{\boldsymbol{\beta}}^{(1)}$. Iteration of this procedure then leads to a joint estimation of regression parameter $\boldsymbol{\beta}$ and nuisance parameter α. Both MLE steps in this algorithm increase the joint log-likelihood.

Remark 5.22 (Implementation of the Negative-Binomial GLM in R*)* Implementation of the negative-binomial model needs some care. There are two R procedures `glm` and `glm.nb` that can be used to fit negative-binomial GLMs, the latter being built on the former. The procedure `glm` is just the classical R procedure [307] that is usually used to fit GLMs within the EDF, it requires to set

```
family=negative.binomial(theta, link="log").
```

This parametrization considers the single-parameter linear EF on $\mathbb{N}$ (for mean $\mu \in \mathcal{M}$)

$$f_{\mathrm{NB}}(n; \mu, \mathtt{theta}) = \binom{n + \mathtt{theta} - 1}{n} \left(\frac{\mu}{\mu + \mathtt{theta}} \right)^{n} \left(1 - \frac{\mu}{\mu + \mathtt{theta}} \right)^{\mathtt{theta}},$$

where $\mathtt{theta} > 0$ denotes the nuisance parameter. The tricky part now is that we have to bring in the different exposures v_i of all policies $1 \leq i \leq n$. That is, we would like to have for claim counts $n_i = v_i y_i$, see (5.38),

$$\begin{aligned} f_{\mathrm{NB}}(y_i; \mu_i, v_i, \alpha) &= \binom{v_i y_i + v_i \alpha - 1}{v_i y_i} \left(\frac{v_i \mu_i}{v_i \mu_i + v_i \alpha} \right)^{v_i y_i} \left(1 - \frac{v_i \mu_i}{v_i \mu_i + v_i \alpha} \right)^{v_i \alpha} \\ &= \binom{v_i y_i + v_i \alpha - 1}{v_i y_i} \left[\left(\frac{\mu_i}{\mu_i + \alpha} \right)^{y_i} \left(1 - \frac{\mu_i}{\mu_i + \alpha} \right)^{\alpha} \right]^{v_i}. \end{aligned}$$

The square bracket can be implemented in `glm` as a scaled and weighted regression problem, see Listing 5.8 with $\mathtt{theta} = \alpha$. This approach provides the correct GLM parameter estimates $\widehat{\boldsymbol{\beta}}^{\mathrm{MLE}}$ for given α, however, the outputted AIC values cannot be compared to the Poisson case. Note that the Poisson case of Table 5.5 considers observations N_i whereas Listing 5.8 uses $Y_i = N_i / v_i$. For this reason we calculate the log-likelihood and AIC by an own implementation.

The same remark applies to `glm.nb`, and also nuisance parameter estimation cannot be performed by that routine under different exposures v_i. Therefore, we have implemented an iterative estimation algorithm ourselves, alternating `glm` of Listing 5.8 for given α and a maximization routine `optimize` to find the optimal α for given $\boldsymbol{\beta}$ using (5.38). We have applied this iteration in Example 5.23, below, and it has converged in 5 iterations.

Example 5.23 (Negative-Binomial Distribution for Claim Counts) We revisit the MTPL claim frequency GLM example of Sect. 5.3.4, but we replace the Poisson distribution by the negative-binomial one. We start with the negative-binomial (NB)

Listing 5.8 Implementation of model NB GLM3

```
d.glmnb <- glm(ClaimNb/Exposure ~ VehPowerGLM + VehAgeGLM
                        + log(DrivAge) + I(DrivAge^3) + I(DrivAge^4)
                        + BonusMalusGLM*DrivAge + BonusMalusGLM*I(DrivAge^2)
                        + VehBrand + VehGas + DensityGLM + Region + AreaGLM,
                        data=learn, weights=Exposure,
                        family=negative.binomial(alpha, link="log"))
```

Table 5.7 Run times, number of parameters, AICs, in-sample and out-of-sample deviance losses (units are in 10^{-2}) and in-sample average frequency of the null models (Poisson and negative-binomial) and the Poisson and negative-binomial GLMs. The optimal model is highlighted in boldface

	Run time	# Param.	AIC	In-sample loss on $\mathcal{L}$	Out-of-sample loss on $\mathcal{T}$	Aver. freq.
Poisson null	–	1	199'506	25.213	25.445	7.36%
Poisson GLM3	15 s	50	192'716	24.084	24.102	7.36%
NB null $\widehat{\alpha}^{\rm MLE}_{\rm null} = 1.059$	–	2	198'466	20.357	20.489	7.36%
NB null $\widehat{\alpha}^{\rm MLE}_{\rm NB} = 1.810$	–	1	198'564	21.796	21.948	7.36%
NB GLM3 $\widehat{\alpha}^{\rm MLE}_{\rm NB} = 1.810$	85s	51	**192'113**	20.722	20.674	7.38%

null model. The NB null model has two parameters, the homogeneous (overall) frequency and the nuisance parameter. MLE of the homogeneous overall frequency is identical to the one in the Poisson null model, and MLE of the nuisance parameter provides $\widehat{\alpha}^{\rm MLE}_{\rm null} = 1.059$. This is substantially smaller than infinity and suggests over-dispersion. The results are presented on the third line of Table 5.7. We observe a smaller AIC of the NB null model against the Poisson null model which says that we should allow for over-dispersion.

We now focus on the NB GLM. The feature pre-processing is done exactly as in model Poisson GLM3, and we choose the log-link for g. We call this model NB GLM3. The iterative estimation procedure outlined above provides a nuisance parameter estimate $\widehat{\alpha}^{\rm MLE}_{\rm NB} = 1.810$. This is bigger than in the NB null model because the regression structure explains some part of the over-dispersion, however, it is still substantially smaller than infinity which justifies the inclusion of this over-dispersion parameter.

The last line of Table 5.7 gives the result of model NB GLM3. From AIC we conclude that we favor the negative-binomial GLM over the Poisson GLM since AIC decreases from 192'716 to 192'113. The in-sample and out-of-sample deviance losses can only be compared within the same models, i.e., the models that have the same cumulant function. This also applies to the negative-binomial models which have cumulant function $\kappa(\theta) = -\alpha \log(1 - e^{\theta})$. Thus, to compare the NB null model and model NB GLM3, we need to choose the same nuisance parameter α. For this reason we added this second NB null model to Table 5.7. This second NB null model no longer uses the MLE $\widehat{\alpha}^{\rm MLE}_{\rm null}$, therefore, the corresponding AIC only includes one estimated parameter.

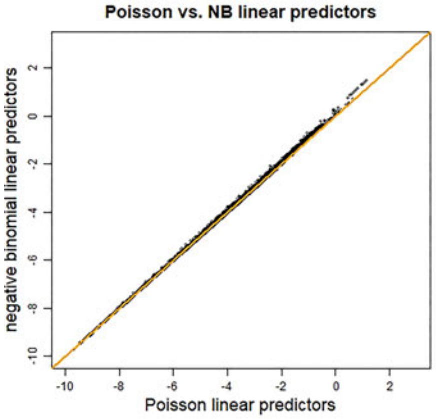


Fig. 5.7 Poisson logged predictors vs. negative-binomial logged predictors

Table 5.8 Out-of-sample deviance losses: forecast dominance. The optimal model is highlighted in boldface

Model	Poisson deviance	NB deviance $\widehat{\alpha}^{\rm MLE}_{\rm null} = 1.059$	NB deviance $\widehat{\alpha}^{\rm MLE}_{\rm NB} = 1.810$
Null model	25.445	20.489	21.948
Poisson GLM3	24.102	19.266	20.678
NB GLM3 $\widehat{\alpha}^{\rm MLE}_{\rm NB} = 1.810$	**24.100**	**19.262**	**20.674**

As mentioned above, deviance losses can only be compared under exactly the same cumulant function (including the same nuisance parameters). If we want to have a more robust model selection, we can consider forecast dominance according to Definition 4.20. Being less ambitious, here, we consider forecast dominance only for the three considered cumulant functions Poisson, negative-binomial with $\widehat{\alpha}^{\rm MLE}_{\rm null} = 1.059$ and negative-binomial with $\widehat{\alpha}^{\rm MLE}_{\rm NB} = 1.810$. The out-of-sample deviance losses are given in Table 5.8 in the different columns. According to this forecast dominance analysis we also give preference to model NB GLM3, but model Poisson GLM3 is pretty close.

Figure 5.7 compares the logged predictors $\log(\widehat{\mu}_i)$, $1 \le i \le n$, of the models Poisson GLM3 and NB GLM3. We see a huge similarity in these predictors, only high frequency policies are judged slightly differently by the NB model compared to the Poisson model.

Table 5.9 gives the predicted number of claims against the observed ones. We observe that model NB GLM3 predicts more accurately the number of policies with 2 or less claims, but it over-estimates the number of policies with more than 2 claims. This may also be related to the fact that the estimated in-sample frequency has a

Table 5.9 Contingency table of observed number of policies against predicted number of policies with given claim counts `ClaimNb`

	Numbers of claims `ClaimNb`					
	0	1	2	3	4	5
Observed number of policies	587’772	21’198	1’174	57	4	1
Poisson predicted number of policies	587’325	22’064	779	34	3	0.3
NB predicted number of policies	587’902	20’982	1’200	100	15	4

positive bias in model NB GLM3, see Table 5.7. That is, since we do not work with the canonical link, we do not have the balance property.

Listing 5.9 `drop1` analysis of model NB GLM3

```
Single term deletions

Model:
ClaimNb/Exposure ~ VehPowerGLM + VehAgeGLM + DrivAge + log(DrivAge) +
    I(DrivAge^2) + I(DrivAge^3) + I(DrivAge^4) + BonusMalusGLM *
    DrivAge + BonusMalusGLM * I(DrivAge^2) + BonusMalusGLM +
    VehBrand + VehGas + DensityGLM + Region + AreaGLM
                           Df Deviance    AIC scaled dev.  Pr(>Chi)
<none>                          126446 171064
VehPowerGLM                 5   126524 171102      48.266 3.134e-09 ***
VehAgeGLM                   2   126655 171190     130.070 < 2.2e-16 ***
log(DrivAge)                1   126592 171153      91.057 < 2.2e-16 ***
I(DrivAge^3)                1   126527 171112      50.483 1.202e-12 ***
I(DrivAge^4)                1   126508 171100      38.381 5.820e-10 ***
VehBrand                   10   126658 171176     132.098 < 2.2e-16 ***
VehGas                      1   126583 171147      85.232 < 2.2e-16 ***
DensityGLM                  1   126456 171068       6.137   0.01324 *
Region                     21   126622 171132     109.838 5.042e-14 ***
AreaGLM                     1   126450 171064       2.411   0.12049
DrivAge:BonusMalusGLM       1   126484 171085      23.481 1.262e-06 ***
I(DrivAge^2):BonusMalusGLM  1   126490 171089      27.199 1.836e-07 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

We close this example by providing the `drop1` analysis in Listing 5.9. From this analysis we conclude that the feature component `Area` should be dropped. Of course, this confirms the high collinearity between `Density` and `Area` which implies that we do not need both variables in the model. We remark that the AIC values in Listing 5.9 are not on our scale, as stated in Remark 5.22. ■

5.3.6 *Zero-Inflated Poisson Model*

In many applications it is the case that the Poisson distribution does not fully fit the claim counts data because there are too many policies with zero claims, i.e., policies with $Y = 0$, compared to a Poisson assumption. This topic has attracted some attention in the recent actuarial literature, see, e.g., Boucher et al. [43–45], Frees et al. [137], Calderín-Ojeda et al. [62] and Lee [239]. An obvious solution to this problem is to 'artificially' increase the probability of a zero claim compared to a Poisson model, this is the proposal introduced by Lambert [232]. Y has a zero-inflated Poisson (ZIP) distribution if the probability weights of Y are given by (set $v = 1$)

$$f_{\mathrm{ZIP}}(y; \theta, \pi_0) = \begin{cases} \pi_0 + (1 - \pi_0)e^{-\mu} & \text{for } y = 0, \\ (1 - \pi_0)e^{-\mu}\frac{\mu^y}{y!} & \text{for } y \in \mathbb{N}, \end{cases}$$

for $\pi_0 \in (0, 1)$, $\mu = e^{\theta} > 0$, and for the Poisson probability weights we refer to (2.4). For $\pi_0 > 0$ the weight of a zero claim $Y = 0$ is increased (inflated) compared to the original Poisson distribution.

Remarks 5.24

- The ZIP distribution has different interpretations. It can be interpreted as a hierarchical model where we have a latent variable Z which indicates with probability π_0 that we have an excess zero, and with probability $1 - \pi_0$ we have an ordinary Poisson distribution, i.e. for $y \in \mathbb{N}_0$

$$\mathbb{P}_{\theta}\left[Y = y \middle| Z = z\right] = \begin{cases} \mathbb{1}_{\{y=0\}} & \text{for } z = 0, \\ e^{-\mu}\frac{\mu^y}{y!} & \text{for } z = 1, \end{cases} \tag{5.41}$$

  with $\mathbb{P}[Z = 0] = 1 - \mathbb{P}[Z = 1] = \pi_0$.

  The latter shows that we can also understand it as a mixture of two distributions, namely, of the Poisson distribution and of a single point measure in $y = 0$ with mixing probability π_0. Mixture distributions are going to be discussed in Sect. 6.3.1, below. In this sense, we can also interpret the model as a mixed Poisson model with mixing distribution $\pi(\lambda)$ being a Bernoulli distribution taking values 0 and μ with probability π_0 and $1 - \pi_0$, respectively, see (5.36), and the former parameter $\lambda = 0$ leads to a degenerate Poisson model.
- We have introduced the ZIP model, but this approach is neither limited to the Poisson model nor the zeros. For instance, we could also consider an inflated negative-binomial model where both the zeros and the ones are inflated with probabilities $\pi_0, \pi_1 > 0$ such that $\pi_0 + \pi_1 < 1$.
- Hurdle models are an alternative way to model excess zeros. Hurdle models have been introduced by Cragg [83], and they also allow for too little zeros. A hurdle (Poisson) model mixes a lower-truncated (Poisson) count distribution with a point mass in zero

$$f_{\text{hurdle Poisson}}(y; \theta, \pi_0) = \begin{cases} \pi_0 & \text{for } y = 0, \\ (1 - \pi_0)\frac{e^{-\mu}\frac{\mu^y}{y!}}{1-e^{-\mu}} & \text{for } y \in \mathbb{N}, \end{cases} \tag{5.42}$$

for $\pi_0 \in (0,1)$ and $\mu > 0$. For $\pi_0 > e^{-\mu}$ the weight of a zero claim is increased and for $\pi_0 < e^{-\mu}$ it is decreased. This distribution is called a hurdle distribution, because we first need to overcome the hurdle at zero to come to the Poisson model. Lower-truncated distributions are studied in Sect. 6.4, below, and mixture distributions are discussed in Sect. 6.3.1. In general, fitting lower-truncated distributions is challenging because the density and the distribution function should both have tractable forms to perform MLE for truncated distributions. The Expectation-Maximization (EM) algorithm is a useful tool to perform model fitting under truncation. We come back to the hurdle Poisson model in Example 6.19, below, and it is also closely related to the zero-truncated Poisson (ZTP) model discussed in Remarks 6.20.

The first two moments of a ZIP random variable $Y \sim f_{\mathrm{ZIP}}(\cdot;\theta,\pi_0)$ are given by

$$\begin{aligned}\mathbb{E}_{\theta,\pi_0}[Y] &= (1-\pi_0)\mu,\\ \mathrm{Var}_{\theta,\pi_0}(Y) &= (1-\pi_0)\mu + (\pi_0-\pi_0^2)\mu^2 \;=\; \mathbb{E}_{\theta,\pi_0}[Y]\,(1+\pi_0\mu)\,,\end{aligned}$$

these calculations easily follow with the latent variable Z interpretation from above. As a consequence, we receive an over-dispersed model with over-dispersion $\pi_0\mu$ (the latter also follows from the fact that we consider a mixed Poisson distribution with a Bernoulli mixing distribution having weights π_0 in 0 and $1-\pi_0$ in $\mu > 0$, see (5.37)).

Unfortunately, MLE does not allow for explicit solutions in this model. The score equations of $Y_i \overset{\text{i.i.d.}}{\sim} f_{\mathrm{ZIP}}(\cdot;\theta,\pi_0)$ are given by

$$\begin{aligned}\nabla_{(\pi_0,\mu)}\ell_{\boldsymbol{Y}}(\pi_0,\mu) &= \nabla_{(\pi_0,\mu)}\sum_{i=1}^n \log\left(\pi_0+(1-\pi_0)e^{-\mu}\right)\mathbb{1}_{\{Y_i=0\}}\\ &\qquad + \nabla_{(\pi_0,\mu)}\sum_{i=1}^n \log\left((1-\pi_0)e^{-\mu}\frac{\mu^y}{y!}\right)\mathbb{1}_{\{Y_i>0\}} \;=\; 0.\end{aligned}$$

The R package `pscl` [401] has a function called `zeroinfl` which uses the general purpose optimizer `optim` to find the MLEs in the ZIP model. Alternatively, we could explore the EM algorithm for mixture distributions presented in Sect. 6.3, below.

In insurance applications, the ZIP application can be problematic if we have different exposures $v_i > 0$ for different insurance policies i. In the Poisson GLM case with canonical link choice we typically integrate the different exposures into the offset, see (5.27). However, it is not clear whether and how we should integrate the different exposures into the zero-inflation probability π_0. It seems natural to believe that shorter exposures should increase π_0, but the explicit functional form of this increase can be debated, some options are discussed in Section 5 of Lee [239].

Listing 5.10 Implementation of model ZIP GLM3

```
1  d.ZIP <- zeroinfl(ClaimNb ~ VehPowerGLM + VehAgeGLM
2                  + log(DrivAge) + I(DrivAge^3) + I(DrivAge^4)
3                  + BonusMalusGLM*DrivAge + BonusMalusGLM*I(DrivAge^2)
4                  + VehBrand + VehGas + DensityGLM + Region
5                  + AreaGLM | 1,
6                  data=learn, offset=log(Exposure), dist=’poisson’, link=’logit’,
7                  start=list(count=glm3$coefficients, zero=c(-0.4153)) )
```

Table 5.10 Run times, number of parameters, AICs, in-sample and out-of-sample deviance losses (units are in 10^{-2}) and in-sample average frequency of the null models (Poisson, negative-binomial and ZIP) and the Poisson, negative-binomial and ZIP GLMs. The optimal model is highlighted in boldface

	Run time	# Param.	AIC	In-sample loss on $\mathcal{L}$	Out-of-sample loss on $\mathcal{T}$	Aver. freq.
Poisson null	–	1	199’506	25.213	25.445	7.36%
Poisson GLM3	15 s	50	192’716	24.084	24.102	7.36%
NB null $\widehat{\alpha}^{\mathrm{MLE}}_{\mathrm{null}} = 1.059$	–	2	198’466	20.357	20.489	7.36%
NB null $\widehat{\alpha}^{\mathrm{MLE}}_{\mathrm{NB}} = 1.810$	–	1	198’564	21.796	21.948	7.36%
NB GLM3 $\widehat{\alpha}^{\mathrm{MLE}}_{\mathrm{NB}} = 1.810$	85 s	51	**192’113**	20.722	20.674	7.38%
ZIP null	20 s	2	198’638	–	–	7.43%
ZIP GLM3 (null π_0)	270 s	51	192’393	–	–	7.37%

In the following application, we simply choose π_0 independent of the exposures, but certainly this is not the best modeling choice.

Example 5.25 (ZIP Model for Claim Counts) We revisit the MTPL claim frequency example of Sect. 5.3.4, but this time we fit a ZIP model. For the Poisson part we use exactly the same GLM regression function as in model Poisson GLM3 and, in particular, we use for the different exposures v_i of the insurance policies the offset term $o_i = \log v_i$, see line 6 of Listing 5.10. This offset only acts on the Poisson part of the ZIP GLM. The zero-inflating probability π_0 is modeled with a logistic Bernoulli model, see Sect. 2.1.2. For computational reasons, we choose the null model for the Bernoulli part modeling the zero-inflation π_0. This is indicated by the “1” on line 5 of Listing 5.10. This 1 should be expanded if we also want to consider a regression model for the zero-inflating probability π_0 and, in particular, if we want to integrate an offset term for the exposure. We can set this term to `offset(f)`, where `f` is a suitable transformation of the exposure. Furthermore, successful calibration requires meaningful starting values, otherwise `zeroinfl` will not find the MLEs. We start the algorithm in the parameters of model Poisson GLM3, see line 7 of Listing 5.10. The results are presented in Table 5.10.

Firstly, we see that the run times are not fully competitive in this implementation, even if we choose the null model for the zero-inflating probability π_0, i.e., only

Table 5.11 Out-of-sample deviance losses: forecast dominance. The optimal model is highlighted in boldface

Model	Poisson deviance	NB deviance $\widehat{\alpha}^{\mathrm{MLE}}_{\mathrm{null}} = 1.059$	NB deviance $\widehat{\alpha}^{\mathrm{MLE}}_{\mathrm{NB}} = 1.810$
Null model	25.445	20.489	21.948
Poisson GLM3	24.102	19.266	20.678
NB GLM3 $\widehat{\alpha}^{\mathrm{MLE}}_{\mathrm{NB}} = 1.810$	**24.100**	**19.262**	**20.674**
ZIP null model	25.446	20.490	21.949
ZIP GLM3	24.103	19.267	20.679

Table 5.12 Contingency table of observed numbers of policies against predicted numbers of policies with given claim counts `ClaimNb`

	Numbers of claims `ClaimNb`					
	0	1	2	3	4	5
Observed number of policies	587’772	21’198	1’174	57	4	1
Poisson predicted number of policies	587’325	22’064	779	34	3	0.3
NB predicted number of policies	587’902	20’982	1’200	100	15	4
ZIP predicted number of policies	587’829	21’094	1’191	79	9	4

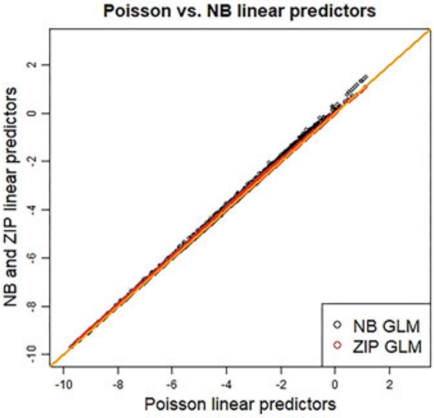
one intercept parameter is involved for determining π_0. Secondly, in this model we cannot calculate deviance losses because the saturated model has two parameters for each observation. Thirdly, the model does not satisfy the balance property though we work with the canonical links for the Poisson part and the Bernoulli part, however, this property gets lost under the combination of these two model parts.

Most interesting are the AIC values. We observe that the ZIP GLM improves the Poisson GLM, but it has a bigger AIC value than the negative-binomial GLM. From this we conclude that we give preference to the negative-binomial model in our case.

Considering forecast dominance according to Definition 4.20, but restricted to the three deviance losses studied in Example 5.23, we receive Table 5.11. Also this table gives preference to the negative-binomial GLM. However, if we consider the table of the observed numbers of policies against the predicted numbers of claims, see Table 5.12, we give preference to the ZIP GLM because it has the lowest χ^2-value, i.e., it reflects best (in-sample) our observations.

Figure 5.8 compares the resulting predictors on the log-scale. From this plot we conclude that in our example the predictors of the ZIP GLM are closer to the Poisson ones than the NB GLM predictors. In a next step, one could refine the zero-inflating probability π_0 modeling by integrating the exposure and further feature information. This would lead to a further model improvement. We refrain here from doing so and close this example; in Example 6.19, below, we study the hurdle Poisson model. ■

Fig. 5.8 Comparison linear predictors of the NB and ZIP GLMs against the ones of the Poisson GLM



5.3.7 Lab: Gamma GLM for Claim Sizes

As a second example we consider claim size modeling within GLMs. For this example we do not use the French MTPL claims data because the empirical density plot in Fig. 13.15 indicates that a GLM will not fit to that data. The French MTPL data seems to have three distinct modes, which suggests to use a mixture distribution. Moreover, the log-log plot indicates a regularly varying tail, which cannot be captured by the EDF on the original observation scale; we are going to study this data in Example 6.14, below. Here, we use the Swedish motorcycle data, previously used in the textbook of Ohlsson–Johansson [290] and described in Chap. 13.2. From Fig. 5.9 we see that the empirical density has one mode, and the log-log plot supports light tails, i.e., the gamma model might be a suitable choice for this data. Therefore, we choose a gamma GLM with log-link g. As described above, the log-link is not the canonical link for the gamma EDF distribution but it ensures the right sign w.r.t. the linear predictor $\eta_i = \langle \boldsymbol{\beta}, \boldsymbol{x}_i \rangle$. Working with the log-link in the gamma model will imply that the balance property is not fulfilled.

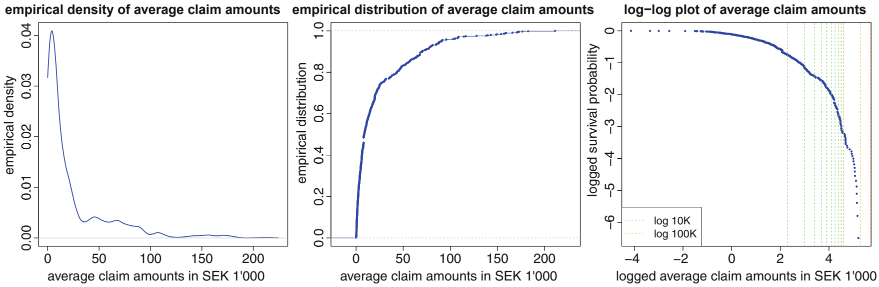


Fig. 5.9 (lhs) Empirical density, (middle) empirical distribution and (rhs) log-log plot of claim amounts of the Swedish motorcycle data presented in Chap. 13.2

Feature Engineering

We have 4 continuous feature components `OwnerAge`, `RiskClass`, `VehAge` and `BonusClass`, one binary feature component `Gender` and a categorical component `Area`, see Listing 13.4. We have decided for a minimal feature engineering; we refer to Figs. 13.19 (rhs) and 13.20 (rhs) for descriptive plots. We use the continuous variables directly in a log-linear fashion, we add quadratic terms for `OwnerAge` and `VehAge`, we merge `RiskClass` 6 and 7, and we censor `VehAge` at 20. `Area` is categorical, but we may interpret the `Zone` levels as ordinal categorical, and mapping them to integers allows us to use them in a continuous fashion; Fig. 13.19 (middle row, rhs) shows that this is a reasonable choice. Moreover, we merge `Zone` 5, 6 and 7 due to small volumes and their similar behavior.

Gamma Generalized Linear Model

The Swedish motorcycle claim amount data poses the special difficulty that we do not have individual claim observations $Z_{i,j}$, but we only know the total claim amounts $S_i = \sum_{j=1}^{N_i} Z_{i,j}$ and the number of claims N_i on each insurance policy; Fig. 5.9 shows average claims S_i/N_i of insurance policies i with $N_i > 0$. In general, this poses a problem in statistical modeling, but in the gamma model this problem can be handled because the gamma distribution is closed under aggregation of i.i.d. gamma claims $Z_{i,j}$. In all what follows in this section, we only study insurance policies with $N_i > 0$, and we label these insurance policies i accordingly.

Assume that $Z_{i,j}$ are i.i.d. gamma distributed with shape parameter α_i and scale parameter c_i, we refer to (2.6). The mean, the variance and the moment generating function of $Z_{i,j}$ are given by

$$\mathbb{E}[Z_{i,j}] = \frac{\alpha_i}{c_i}, \qquad \mathrm{Var}(Z_{i,j}) = \frac{\alpha_i}{c_i^2} \qquad \text{and} \qquad M_{Z_{i,j}}(r) = \left(\frac{c_i}{c_i - r}\right)^{\alpha_i}, \tag{5.43}$$

where the moment generating function requires $r < c_i$ to be finite. Assuming that the number of claims N_i is a known positive integer $n_i \in \mathbb{N}$, we see from the moment generating function that $S_i = \sum_{j=1}^{n_i} Z_{i,j}$ is again gamma distributed with shape parameter $n_i\alpha_i$ and scale parameter c_i. We change the notation from N_i to n_i to emphasize that the number of claims is treated as a known constant (and also to avoid using the notation of conditional probabilities, here). Finally, we scale $Y_i = S_i/(n_i\alpha_i) \sim \Gamma(n_i\alpha_i, n_i\alpha_i c_i)$. This random variable Y_i has a single-parameter EDF gamma distribution with weight $v_i = n_i$, dispersion $\varphi_i = 1/\alpha_i$ and cumulant function $\kappa(\theta_i) = -\log(-\theta_i)$, for $\theta_i \in \boldsymbol{\Theta} = (-\infty, 0)$,

$$\begin{aligned} Y_i \sim f(y; \theta_i, v_i/\varphi_i) &= \exp\left\{\frac{y\theta_i - \kappa(\theta_i)}{\varphi_i/v_i} + a(y; v_i/\varphi_i)\right\} \qquad (5.44) \\ &= \frac{(-\theta_i\alpha_i v_i)^{v_i\alpha_i}}{\Gamma(v_i\alpha_i)} y^{v_i\alpha_i - 1} \exp\left\{-(-\theta_i\alpha_i v_i)y\right\}, \end{aligned}$$

and the canonical parameter is $\theta_i = -c_i$. For our GLM analysis we treat the shape parameter $\alpha_i \equiv \alpha > 0$ as a nuisance parameter that does not depend on the specific policy i, i.e., we set constant dispersion $\varphi = 1/\alpha$, and only the scale parameter c_i is chosen policy dependent through $\theta_i = -c_i$.

Random variable $Y_i = S_i/(n_i\alpha) \sim \Gamma(n_i\alpha, n_i\alpha c_i)$ gives the reproductive form of the gamma EDF, see Remarks 2.13. In applications, this form is not directly useful because under unknown shape parameter α, we cannot calculate observations $Y_i = S_i/(n_i\alpha)$. For this reason, we parametrize the model differently, here. We consider instead

$$Y_i = S_i/n_i \sim \Gamma(n_i\alpha, n_i c_i). \tag{5.45}$$

This (new) random variable has the same gamma EDF (5.44), we only need to reinterpret the canonical parameter as $\theta_i = -c_i/\alpha$. Then, we choose the log-link for g which implies

$$\mu_i = \mathbb{E}_{\theta_i}[Y_i] = \kappa'(\theta_i) = -\frac{1}{\theta_i} = \exp\{\eta_i\} = \exp\langle \boldsymbol{\beta}, \boldsymbol{x}_i\rangle,$$

if $\boldsymbol{x}_i \in \mathcal{X} \subset \mathbb{R}^{q+1}$ describes the pre-processed features of policy i. The gamma GLM is now fully specified and can be fitted to the data; from Example 5.5 we know that we have a concave maximization problem. We call this model Gamma GLM1 (with the feature pre-processing as described above). Note that the (constant) dispersion parameter φ cancels in the score equations, thus, we do not need to explicitly specify the nuisance parameter α to estimate regression parameter $\boldsymbol{\beta} \in \mathbb{R}^{q+1}$.

Maximum Likelihood Estimation and Model Selection

Because we have only few claims data in this Swedish motorcycle example (only $m = 656$ insurance policies suffer claims), we do not perform a generalization analysis with learning and test samples. In this situation we need all data for model fitting, and model performance is analyzed with AIC and with tenfold cross-validation.

The in-sample deviance loss in the gamma GLM is given by

$$\mathfrak{D}(\mathcal{L}, \widehat{\mu}(\cdot)) = \frac{2}{m}\sum_{i=1}^{m} \frac{n_i}{\varphi}\left(\frac{Y_i - \widehat{\mu}(\boldsymbol{x}_i)}{\widehat{\mu}(\boldsymbol{x}_i)} - \log\left(\frac{Y_i}{\widehat{\mu}(\boldsymbol{x}_i)}\right)\right), \tag{5.46}$$

where i runs over the policies $i = 1, \ldots, m$ with positive claims $Y_i = S_i/n_i > 0$, and $\widehat{\mu}(\boldsymbol{x}_i) = \exp\langle \widehat{\boldsymbol{\beta}}^{\text{MLE}}, \boldsymbol{x}_i\rangle$ is the MLE estimated regression function. Similar to the Poisson case (5.29), McCullagh–Nelder [265] derive the following behavior

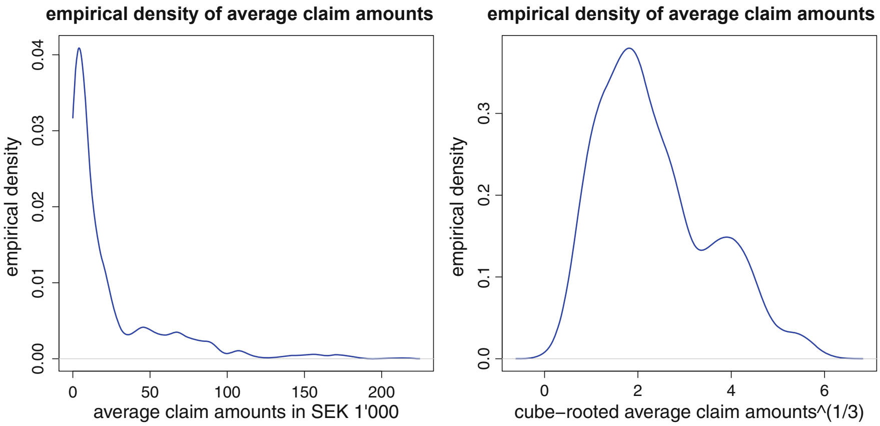


Fig. 5.10 (lhs) Empirical density of Y_i and (rhs) empirical density of $Y_i^{1/3}$

for the gamma unit deviance around its mode, see Section 7.2 and Figure 7.2 in McCullagh–Nelder [265],

$$\mathfrak{d}\left(Y_i, \mu_i\right) \approx 9Y_i^{2/3}\left(Y_i^{-1/3} - \mu_i^{-1/3}\right)^2, \tag{5.47}$$

this uses that the log-likelihood is symmetric around its mode for scale $\mu_i^{-1/3}$, see Fig. 5.5 (middle). This shows that the gamma deviance scales differently around Y_i compared to the square loss function. From this we receive an approximation to the deviance residuals (for $v/\varphi = 1$)

$$r_i^{\mathrm{D}} = \operatorname{sign}(Y_i - \mu_i)\sqrt{\mathfrak{d}\left(Y_i, \mu_i\right)} \approx 3\left(\left(\frac{Y_i}{\mu_i}\right)^{1/3} - 1\right) = 3\,\frac{Y_i^{1/3} - \mu_i^{1/3}}{\mu_i^{1/3}}. \tag{5.48}$$

This is the cube-root transformation derived by Wilson–Hilferty [383]. This suggests that if the empirical distribution of $Y_i^{1/3}$ looks roughly Gaussian we can use a gamma distribution. Figure 5.10 gives the empirical densities of Y_i on the left-hand side and of $Y_i^{1/3}$ on the right-hand side. The latter looks roughly Gaussian (except of the second mode close to 4), this supports the use of a gamma model.

Listing 5.11 provides the summary statistics of the fitted model Gamma GLM1; note that we integrate the number of claims n_i through scaling into the `weights`. We have $q + 1 = 9$ regression parameters, and from this summary statistics we observe that not all variables should be kept in the model. If we perform backward elimination using `drop1` in each step, see Sect. 5.3.3, we first drop `BonusClass` and then `Gender`, resulting in a reduced model with 7 parameters. We call this reduced model Gamma GLM2.

Listing 5.11 Results in model Gamma GLM1 using the R command `glm`

```
Call:
glm(formula = ClaimAmount/ClaimNb ~ OwnerAge + I(OwnerAge^2) +
    AreaGLM + RiskClass + VehAge + I(VehAge^2) + Gender + BonusClass,
    family = Gamma(link = "log"), data = mcdata0, weights = ClaimNb)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.3683  -1.4585  -0.5979   0.4354   3.4763

Coefficients:
                Estimate Std. Error t value Pr(>!t!)
(Intercept)    8.9737854  0.5532821  16.219  < 2e-16 ***
OwnerAge       0.1072781  0.0280862   3.820 0.000147 ***
I(OwnerAge^2) -0.0014508  0.0003489  -4.158 3.65e-05 ***
AreaGLM       -0.0768512  0.0368284  -2.087 0.037303 *
RiskClass      0.0615575  0.0327553   1.879 0.060651 .
VehAge        -0.2051148  0.0296184  -6.925 1.05e-11 ***
I(VehAge^2)    0.0062649  0.0015946   3.929 9.45e-05 ***
GenderMale     0.1085538  0.1673443   0.649 0.516772
BonusClass     0.0089004  0.0225371   0.395 0.693029
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Gamma family taken to be 1.536577)

    Null deviance: 1368.0  on 655  degrees of freedom
Residual deviance: 1126.5  on 647  degrees of freedom
AIC: 14922

Number of Fisher Scoring iterations: 11
```

Table 5.13 Run times, number of parameters, AICs, Pearson's dispersion estimate, in-sample losses, tenfold cross-validation losses and the in-sample average claim amounts of the null model (gamma intercept model) and the gamma GLMs

	Run time	# Param.	AIC	Dispersion est. $\widehat{\varphi}^{\mathrm{P}}$	In-sample loss on $\mathcal{L}$	Tenfold CV loss $\widehat{\mathfrak{D}}^{\mathrm{CV}}$	Average amount
Gamma null	–	$1+1$	14'416	2.057	2.085	2.091	24'641
Gamma GLM1	1s	$9+1$	14'277	1.537	1.717	1.752	25'105
Gamma GLM2	1s	$7+1$	14'274	1.544	1.719	1.747	25'130

The results of models Gamma GLM1 and Gamma GLM2 are presented in Table 5.13. We show AICs, Pearson's dispersion estimate, the in-sample deviance losses on all available data, the corresponding tenfold cross-validation losses, and the average claim amounts.

Firstly, we observe that the GLMs do not meet the balance property. This is implied by the fact that we do not use the canonical link to avoid any sort of difficulty of dealing with the one-sided bounded effective domain $\boldsymbol{\Theta} = (-\infty, 0)$. For pricing, the intercept parameter $\widehat{\beta}_0^{\mathrm{MLE}}$ should be shifted to eliminate this bias, i.e, we need to shift this parameter under the log-link by $-\log(25'130/24'641)$ for model Gamma GLM2.

Secondly, the in-sample and tenfold cross-validation losses are not directly comparable to AIC. Observe that we need to know the dispersion parameter φ in order to calculate both of these statistics. For the in-sample and cross-validation losses we have set $\varphi = 1$, thus, all these figures are directly comparable. For AIC we have estimated the dispersion parameter φ with MLE. This is the reason for increasing the number of parameters in Table 5.13 by +1. Moreover, the resulting AICs differ from the ones received from the R command `glm`, see, for instance, Listing 5.11. The AIC value in Listing 5.11 does not consider all terms appropriately due to the inclusion of `weights`, this is similar to Remark 5.22, it uses the deviance dispersion estimate $\widehat{\varphi}^{\mathrm{D}}$, i.e., not the MLE and (still) increases the number of parameters by 1 because the dispersion is estimated. For these reasons, we have implemented our own code for calculating AIC. Both AIC and the tenfold cross-validation losses say that we should give preference to model Gamma GLM2.

The dispersion estimate in Listing 5.11 corresponds to Pearson's estimate

$$\widehat{\varphi}^{\mathrm{P}} = \frac{1}{m-(q+1)} \sum_{i=1}^{m} n_i \frac{(Y_i - \widehat{\mu}_i)^2}{\widehat{\mu}_i^2}. \tag{5.49}$$

We observe that the dispersion estimate is roughly 1.5 which gives an estimate of the shape parameter $\alpha = 1/\varphi$ of $2/3$. A shape parameter less than 1 implies that the density of the gamma distribution is strictly decreasing, see Fig. 2.1. Often this is a sign that the model does not fully fit the data, and if we use this model for simulation we may receive too many observations close to zero compared to the true data. A shape parameter less than 1 may be implied by more heterogeneity in the data compared to what the chosen gamma GLM allows for or by large claims that cannot be explained by the present gamma density structure. Thus, there is some sign here that the data is more heavy-tailed than our model choice suggests. Alternatively, there might be some need to also model the shape parameter with a regression model; this could be done using the vector-valued parameter EF representation of the gamma model, see Sect. 2.1.3. In view of Fig. 5.10 (rhs) it may also be that the feature information is not sufficient to describe the second mode in 4, thus, we probably need more explanatory information to reduce dispersion.

In Fig. 5.11 we give the Tukey–Anscombe plot and a QQ plot. Note that the observations for $n_i = 1$ follow a gamma distribution with shape parameter α and scale parameter $c_i = \alpha/\mu_i = -\alpha\theta_i$. Thus, if we scale Y_i/μ_i, we receive i.i.d. gamma random variables with shape and scale parameters equal to α. This then allows us for $n_i = 1$ to plot the empirical distribution of $Y_i/\widehat{\mu}_i$ against $\Gamma(\alpha, \alpha)$ in a QQ plot where we estimate $1/\alpha$ by Pearson's dispersion estimate. The Tukey–Anscombe plot looks reasonable, but the QQ plot shows that the gamma model does not entirely fit the data. From this plot we cannot conclude whether the gamma distribution is causing the problem or whether it is a missing term in the regression structure. We only see that the data is over-dispersed, resulting in more heavy-tailed observations than the theoretical gamma model can explain, and a compensation by too many small observations (which is induced by over-dispersion, i.e., a shape parameter smaller than one). In the network chapter we will refine the regression function, keeping the gamma assumption, to understand which modeling part is causing the difficulty.

Remark 5.26 For the calculation of AIC in Table 5.13 we have used the MLE of the dispersion parameter φ. This is obtained by solving the score equation (5.11) for the

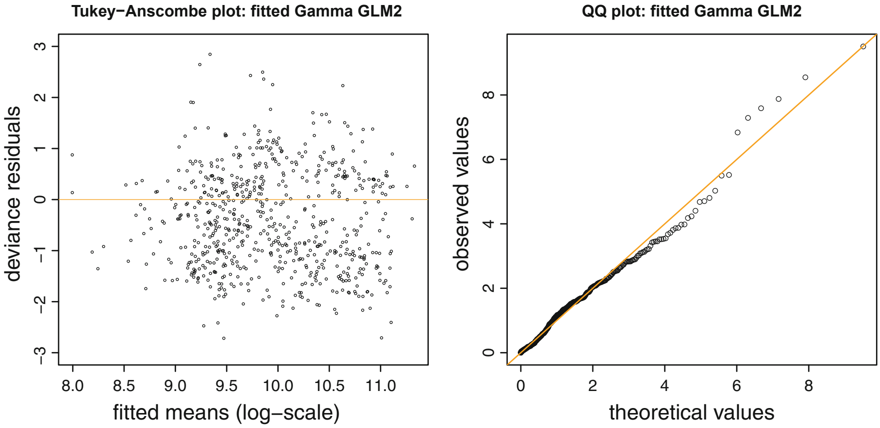


Fig. 5.11 (lhs) Tukey–Anscombe plot of the fitted model Gamma GLM2, and (rhs) QQ plot of the fitted model Gamma GLM2

gamma case. It is given by, we set $\alpha = 1/\varphi$ and we calculate the MLE of α instead,

$$\frac{\partial}{\partial\alpha}\ell_{\boldsymbol{Y}}(\boldsymbol{\beta},\alpha) = \sum_{i=1}^{n} v_i\Big[Y_i h(\mu(\boldsymbol{x}_i)) - \kappa\left(h(\mu(\boldsymbol{x}_i))\right) + \log Y_i + \log(\alpha v_i) + 1 - \Psi(\alpha v_i)\Big] = 0,$$

where $\Psi(\alpha) = \Gamma'(\alpha)/\Gamma(\alpha)$ is the digamma function. We calculate the second derivative w.r.t. α, see also (2.30),

$$\frac{\partial^2}{\partial\alpha^2}\ell_{\boldsymbol{Y}}(\boldsymbol{\beta},\alpha) = \sum_{i=1}^{n} v_i\left[\frac{1}{\alpha} - v_i\Psi'(\alpha v_i)\right] = \sum_{i=1}^{n} v_i^2\left[\frac{1}{\alpha v_i} - \Psi'(\alpha v_i)\right] < 0 \qquad \text{for } \alpha > 0,$$

the negativity follows from Theorem 1 in Alzner [9]. In fact, the function $\log\alpha - \Psi(\alpha)$ is strictly completely monotonic for $\alpha > 0$. This says that the log-likelihood $\ell_{\boldsymbol{Y}}(\boldsymbol{\beta},\alpha)$ is a concave function in $\alpha > 0$ and the solution to the score equation is unique, giving the MLE of α and φ, respectively.

5.3.8 Lab: Inverse Gaussian GLM for Claim Sizes

We present the inverse Gaussian GLM in this section as a competing model to the gamma GLM studied in the previous section.

Infinite Divisibility

In the gamma model above we have used that the total claim amount $S = \sum_{j=1}^{n} Z_j$ has a gamma distribution for given claim counts $N = n > 0$ and i.i.d. gamma claim sizes Z_j. This property is closely related to divisibility. A random variable S is called divisible by $n \in \mathbb{N}$ if there exist i.i.d. random variables $Z_1, \ldots, Z_n$ such that

$$S \overset{(d)}{=} \sum_{j=1}^{n} Z_j,$$

and S is called *infinitely divisible* if S is divisible by n for all $n \in \mathbb{N}$. The EDF is based on parameters $(\theta, \omega) \in \mathbf{\Theta} \times \mathcal{W}$. Jørgensen [203] gives the following interesting result.

Theorem 5.27 (Theorem 3.7 in Jørgensen [203], Without Proof) *Choose a member of the EDF with parameter set $\mathbf{\Theta} \times \mathcal{W}$. Then*

- *the index set $\mathcal{W}$ is an additive semi-group and $\mathbb{N} \subseteq \mathcal{W} \subseteq \mathbb{R}_+$, and*
- *the members of the chosen EDF are infinitely divisible if and only if $\mathcal{W} = \mathbb{R}_+$.*

This theorem tells us how to aggregate and disaggregate within EDFs, e.g., the Poisson, gamma and inverse Gaussian models are infinitely divisible, and the binomial distribution is divisible by n with the disaggregated random variables belonging to the same EDF and the same canonical parameter, see Sect. 2.2.2. In particular, we also refer to Corollary 2.15 on the convolution property.

Inverse Gaussian Generalized Linear Model

Alternatively to the gamma GLM one often explores an inverse Gaussian GLM which has a cubic variance function $V(\mu) = \mu^3$. We bring this inverse Gaussian model into the same form as the gamma model of Sect. 5.3.7, so that we can aggregate claims within insurance policies. The mean, the variance and the moment generating function of an inverse Gaussian random variable $Z_{i,j}$ with parameters $\alpha_i, c_i > 0$ are given by

$$\mathbb{E}[Z_{i,j}] = \frac{\alpha_i}{c_i}, \quad \mathrm{Var}(Z_{i,j}) = \frac{\alpha_i}{c_i^3} \quad \text{and} \quad M_{Z_{i,j}}(r) = \exp\left\{\alpha_i \left[c_i - \sqrt{c_i^2 - 2r}\right]\right\},$$

where the moment generating function requires $r < c_i^2/2$ to be finite. From the moment generating function we see that $S_i = \sum_{j=1}^{n_i} Z_{i,j}$ is inverse Gaussian distributed with parameters $n_i\alpha_i$ and c_i. Finally, we scale $Y_i = S_i/(n_i\alpha_i)$ which provides us with an inverse Gaussian distribution with parameters $n_i^{1/2}\alpha_i^{1/2}$ and $n_i^{1/2}\alpha_i^{1/2}c_i$. This random variable Y_i has a single-parameter EDF inverse Gaussian distribution in its reproductive form, namely,

$$\begin{aligned} Y_i \sim f(y; \theta_i, v_i/\varphi_i) &= \exp\left\{\frac{y\theta_i - \kappa(\theta_i)}{\varphi_i/v_i} + a(y; v_i/\varphi_i)\right\} \qquad (5.50) \\ &= \frac{\alpha_i^{1/2}}{\sqrt{\frac{2\pi}{v_i} y^3}} \exp\left\{-\frac{\alpha_i}{2y/v_i}\left(1 - \sqrt{-2\theta_i y}\right)^2\right\}, \end{aligned}$$

with cumulant function $\kappa(\theta) = -\sqrt{-2\theta}$ for $\theta \in \boldsymbol{\Theta} = (-\infty, 0]$, weight $v_i = n_i$, dispersion parameter $\varphi_i = 1/\alpha_i$ and canonical parameter $\theta_i = -c_i^2/2$.

Similarly to the gamma case, this representation is not directly useful if the parameter α_i is not known. Therefore, we parametrize this model differently. Namely, we consider

$$Y_i = S_i/n_i \sim \text{InvGauss}\left(n_i^{1/2}\alpha_i, n_i^{1/2}c_i\right). \tag{5.51}$$

This re-scaled random variable has that same inverse Gaussian EDF (5.50), but we need to re-interpret the parameters. We have dispersion parameter $\varphi_i = 1/\alpha_i^2$ and canonical parameter $\theta_i = -c_i^2/(2\alpha_i^2)$. For our GLM analysis we will treat the parameter $\alpha_i \equiv \alpha > 0$ as a nuisance parameter that does not depend on the specific policy i. Thus, we have constant dispersion $\varphi = 1/\alpha^2$ and only the scale parameter c_i is assumed to be policy dependent through the canonical parameter $\theta_i = -c_i^2/(2\alpha^2)$.

We are now in the same situation as in the gamma case in Sect. 5.3.7. We choose the log-link for g which implies

$$\mu_i = \mathbb{E}_{\theta_i}[Y_i] = \kappa'(\theta_i) = \frac{1}{\sqrt{-2\theta_i}} = \exp\{\eta_i\} = \exp\langle\boldsymbol{\beta}, \boldsymbol{x}_i\rangle,$$

for $\boldsymbol{x}_i \in \mathcal{X} \subset \mathbb{R}^{q+1}$ describing the pre-processed features of policy i. We use the same feature pre-processing as in model Gamma GLM2, and we call this resulting model IG GLM2. Again the constant dispersion parameter $\varphi = 1/\alpha^2$ cancels in the score equations, thus, we do not need to explicitly specify the nuisance parameter α to estimate the regression parameter $\boldsymbol{\beta} \in \mathbb{R}^{q+1}$. However, there is an important difference to the gamma GLM, namely, as stated in Example 5.6, we do not have a concave maximization problem and Fisher's scoring method needs a suitable initial value. We start the fitting algorithm in the parameters of model Gamma GLM2.

The in-sample deviance loss in the inverse Gaussian GLM is given by

$$\mathfrak{D}(\mathcal{L}, \widehat{\mu}(\cdot)) = \frac{1}{m}\sum_{i=1}^{m} \frac{n_i}{\varphi} \frac{(Y_i - \widehat{\mu}(\boldsymbol{x}_i))^2}{\widehat{\mu}(\boldsymbol{x}_i)^2\, Y_i}, \tag{5.52}$$

where i runs over the policies $i = 1, \ldots, m$ with positive claims $Y_i = S_i/n_i > 0$, and $\widehat{\mu}(\boldsymbol{x}_i) = \exp\langle\widehat{\boldsymbol{\beta}}^{\text{MLE}}, \boldsymbol{x}_i\rangle$ is the MLE estimated regression function. The unit deviances behave as

$$\mathfrak{d}\,(Y_i, \mu_i) = Y_i \left(Y_i^{-1} - \mu_i^{-1}\right)^2, \tag{5.53}$$

Table 5.14 Run times, number of parameters, AICs, in-sample losses, tenfold cross-validation losses and the in-sample average claim amounts of the null gamma model, model Gamma GLM2, the null inverse Gaussian model, and model inverse Gaussian GLM2; the deviance losses use unit dispersion $\varphi = 1$

	Run time	# Param.	AIC	In-sample loss on $\mathcal{L}$	Tenfold CV loss $\widehat{\mathfrak{D}}^{\text{CV}}$	Average amount
Gamma null	–	$1 + 1$	14'416	2.085	2.091	24'641
Gamma GLM2	1 s	$7 + 1$	14'274	1.719	1.747	25'130
IG null	–	$1 + 1$	14'715	$5.012 \cdot 10^{-4}$	$5.016 \cdot 10^{-4}$	24'641
IG GLM2	1 s	$7 + 1$	14'686	$4.793 \cdot 10^{-4}$	$4.820 \cdot 10^{-4}$	32'268

note that the log-likelihood is symmetric around its mode for scale μ_i^{-1}, see Fig. 5.5 (rhs). From this we receive deviance residuals (for $v/\varphi = 1$)

$$r_i^{\text{D}} = \text{sign}(Y_i - \mu_i)\sqrt{\mathfrak{d}\,(Y_i, \mu_i)} = Y_i^{1/2}\left(\mu_i^{-1} - Y_i^{-1}\right).$$

Thus, these residuals behave as $Y_i^{1/2}$ for $Y_i \to \infty$ (and fixed μ_i^{-1}), which is more heavy-tailed than the cube-root behavior $Y_i^{1/3}$ in the gamma case, see (5.48). Another difference to the gamma case is that the deviance loss (5.52) is not scale-invariant, see also (11.4), below.

We revisit the example of Table 5.13, but we replace the gamma distribution by the inverse Gaussian distribution. The results in Table 5.14 show that the inverse Gaussian model is not fully competitive on this data set. In view of (5.43) we observe that the coefficient of variation (standard deviation divided by mean) is in the gamma model given by $1/\sqrt{\alpha}$, thus, in the gamma model this coefficient of variation is independent of the expected claim size μ_i and only depends on the shape parameter α. In the inverse Gaussian model the coefficient of variation is given by

$$\text{Vco}(Z_{i,j}) = \frac{\sqrt{\text{Var}(Z_{i,j})}}{\mathbb{E}[Z_{i,j}]} = \frac{\sqrt{\mu_i}}{\alpha},$$

thus, it monotonically increases in the expected claim size μ_i. It seems that this structure is not fully suitable for this data set, i.e., there is no indication that the coefficient of variation increases in the expected claim size. We come back to a comparison of the gamma and the inverse Gaussian model in Sect. 11.1, below.

5.3.9 Log-Normal Model for Claim Sizes: A Short Discussion

Another way to improve the gamma model of Sect. 5.3.7 could be to use a log-normal distribution instead. In the above situation this does not work because the observations are not in the right format. If the claim observations $Z_{i,j}$ are log-normally distributed, then $\log(Z_{i,j})$ are normally distributed. Unfortunately, in our Swedish motorcycle data set we do not have individual claim observations $Z_{i,j}$, but the provided information is aggregated over all claims per insurance policy, i.e., $S_i = \sum_{j=1}^{N_i} Z_{i,j}$. Therefore, there is no possibility here to challenge the gamma framework of Sect. 5.3.7 with a corresponding log-normal framework, because the log-normal framework is not closed under summation of i.i.d. log-normally distributed random variables.

We would like to give some remarks that concern calculations on the log-scale (or any other strictly increasing and concave transformation of the original data). For the log-normal distribution, as well as in similar cases like the log-gamma distribution, one works with logged observations $Y_i = \log(Z_i)$. This is a strictly monotone transformation and the MLEs in the log-normal model based on observations Z_i and in the normal model based on observations $Y_i = \log(Z_i)$ coincide. This can be seen from the following calculation. We start from the log-normal density on $\mathbb{R}_+$, and we do a transformation of variable $z > 0 \mapsto y = \log(z) \in \mathbb{R}$ with $dy = dz/z$

$$\begin{aligned} f_{\mathrm{LN}}(z;\mu,\sigma^2)dz &= \frac{1}{\sqrt{2\pi\sigma^2}}\frac{1}{z}\exp\left\{-\frac{1}{2\sigma^2}\left(\log(z)-\mu\right)^2\right\}dz \\ &= \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{1}{2\sigma^2}\left(y-\mu\right)^2\right\}dy \;=\; f_\Phi(y;\mu,\sigma^2)dy. \end{aligned}$$

From this we see that the MLEs will coincide.

In many situations, one assumes that $\sigma^2 > 0$ is a given nuisance parameter, and one models $\boldsymbol{x} \mapsto \mu(\boldsymbol{x})$ with a GLM within the single-parameter EDF. In the log-normal/Gaussian case one typically chooses the canonical link on the log-scale which is the identity function. This then allows one to perform a classical linear regression for $\mu(\boldsymbol{x}) = \langle\boldsymbol{\beta},\boldsymbol{x}\rangle$ using the logged observations $\boldsymbol{Y} = (Y_1,\ldots,Y_n)^\top = (\log(Z_1),\ldots,\log(Z_n))^\top$, and the corresponding MLE is given by

$$\widehat{\boldsymbol{\beta}}^{\mathrm{MLE}} = (\mathfrak{X}^\top\mathfrak{X})^{-1}\mathfrak{X}^\top\boldsymbol{Y}, \tag{5.54}$$

for full rank $q+1 \le n$ design matrix $\mathfrak{X}$. Note that in this case we have a closed-form solution for the MLE of $\boldsymbol{\beta}$. This is called the homoskedastic case because all observations Y_i are assumed to have the same variance σ^2, otherwise, in the heteroskedastic case, we would still have to include the covariance matrix.

Since we work with the canonical link on the log-scale we have the balance property on the log-scale, see Corollary 5.7. Thus, we receive unbiasedness

$$\sum_{i=1}^n \mathbb{E}_{\boldsymbol{\beta}}\left[\mathbb{E}_{\widehat{\boldsymbol{\beta}}^{\mathrm{MLE}}}[Y_i]\right] = \sum_{i=1}^n \mathbb{E}_{\boldsymbol{\beta}}\left[\langle\widehat{\boldsymbol{\beta}}^{\mathrm{MLE}},\boldsymbol{x}_i\rangle\right] = \sum_{i=1}^n \mathbb{E}_{\boldsymbol{\beta}}\left[Y_i\right] = \sum_{i=1}^n \mu(\boldsymbol{x}_i). \tag{5.55}$$

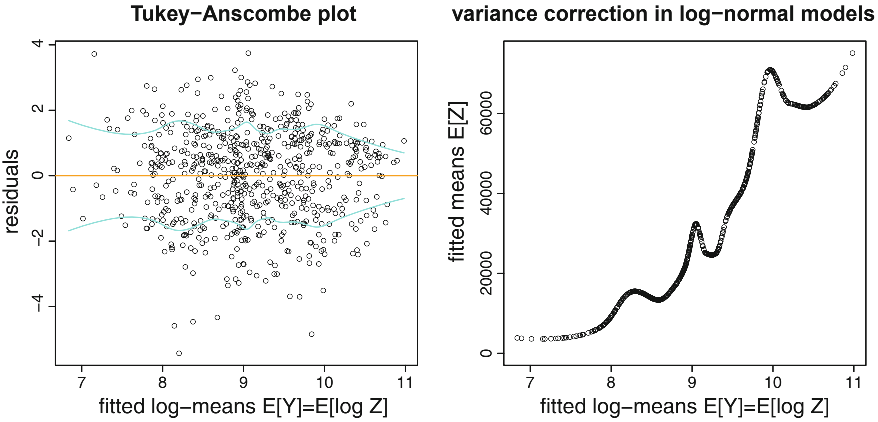


Fig. 5.12 (lhs) Tukey–Anscombe plot of the fitted Gaussian model $\widehat{\mu}(\boldsymbol{x}_i)$ on the logged claim sizes $Y_i = \log(Z_i)$, and (rhs) estimated means $\widehat{\mu}_{Z_i}$ as a function of $\widehat{\mu}(\boldsymbol{x}_i)$ considering heteroskedasticity $\widehat{\sigma}(\boldsymbol{x}_i)$

If we move back to the original scale of the observations Z_i we receive from the log-normal assumption

$$\mathbb{E}_{(\widehat{\boldsymbol{\beta}}^{\text{MLE}},\sigma^2)}[Z_i] = \exp\left\{\langle \widehat{\boldsymbol{\beta}}^{\text{MLE}}, \boldsymbol{x}_i\rangle + \sigma^2/2\right\}.$$

Therefore, we need to adjust with the nuisance parameter σ^2 for the back-transformation to the original observation scale. At this point, typically, the difficulties start. Often, a good back-transformation involves a feature dependent variance parameter $\sigma^2(\boldsymbol{x}_i)$, thus, in many practical applications the homoskedasticity assumption is not fulfilled, and a constant variance parameter choice leads to a poor model on the original observation scale.

A suitable estimation of $\sigma^2(\boldsymbol{x}_i)$ may turn out to be rather difficult. This is illustrated in Fig. 5.12. The left-hand side of this figure shows the Tukey–Anscombe plot of the homoskedastic case providing unscaled ($\sigma^2 \equiv 1$) (Pearson's) residuals on the log-scale

$$r_i^{\text{P}} = \log(Z_i) - \widehat{\mu}(\boldsymbol{x}_i) = Y_i - \widehat{\mu}(\boldsymbol{x}_i).$$

The light-blue color shows an insurance policy dependent standard deviation estimate $\widehat{\sigma}(\boldsymbol{x}_i)$. In our case this estimate is non-monotone in $\widehat{\mu}(\boldsymbol{x}_i)$ (which is quite common on real data). Using this estimate we can estimate the means of the log-normal random variables by

$$\widehat{\mu}_{Z_i} = \widehat{\mathbb{E}}[Z_i] = \exp\left\{\widehat{\mu}(\boldsymbol{x}_i) + \widehat{\sigma}(\boldsymbol{x}_i)^2/2\right\}.$$

The right-hand side of Fig. 5.12 plots these estimated means $\widehat{\mu}_{Z_i}$ against the estimated means $\widehat{\mu}(\boldsymbol{x}_i)$ on the log-scale. We observe a graph that is non-monotone, implied by the non-monotonicity of the standard deviation estimate $\widehat{\sigma}(\boldsymbol{x}_i)$ as a function of $\widehat{\mu}(\boldsymbol{x}_i)$. This non-monotonicity is not bad per se, as we still have a proper statistical model, however, it might be rather counter-intuitive and difficult to explain. For this reason it is advisable to directly model the expected value by one single function, and not to decompose it into different regression functions.

Another important point to be considered is that for model selection using AIC we have to work on the same scale for all models. Thus, if we use a gamma model to model Z_i, then for an AIC selection we need to evaluate also the log-normal model on that scale. This can be seen from the justification in Sect. 4.2.3.

Finally, we focus on unbiasedness. Note that on the log-scale we have unbiasedness (5.55) through the balance property. Unfortunately, this does not carry over to the original scale. We give a small example, where we assume that there is neither any uncertainty about the distributional model nor about the nuisance parameter. That is, we assume that Z_i are i.i.d. log-normally distributed with parameters μ and σ^2, where only μ is unknown. The MLE of μ is given by

$$\widehat{\mu}^{\mathrm{MLE}} = \frac{1}{n}\sum_{i=1}^{n} \log(Z_i) \;\sim\; \mathcal{N}(\mu, \sigma^2/n).$$

In this case we have

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{(\mu,\sigma^2)}\left[\mathbb{E}_{(\widehat{\mu}^{\mathrm{MLE}},\sigma^2)}[Z_i]\right] &= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{(\mu,\sigma^2)}\left[\exp\{\widehat{\mu}^{\mathrm{MLE}}\}\right]\exp\{\sigma^2/2\}\\ &= \exp\left\{\mu + (1+n^{-1})\sigma^2/2\right\}\\ &> \exp\left\{\mu + \sigma^2/2\right\} \;=\; \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{(\mu,\sigma^2)}\left[Z_i\right].\end{aligned}$$

Volatility in parameter estimation $\widehat{\mu}^{\mathrm{MLE}}$ leads to a positive bias in this case. Note that we have assumed full knowledge of the distributional model (i.i.d. log-normal) and the nuisance parameter σ^2 in this calculation. If, for instance, we do not know the true nuisance parameter and we work with (deterministic) $\widetilde{\sigma}^2 \ll \sigma^2$ and $n > 1$, we can get a negative bias

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{(\mu,\sigma^2)}\left[\mathbb{E}_{(\widehat{\mu}^{\mathrm{MLE}},\widetilde{\sigma}^2)}[Z_i]\right] &= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{(\mu,\sigma^2)}\left[\exp\{\widehat{\mu}^{\mathrm{MLE}}\}\right]\exp\{\widetilde{\sigma}^2/2\}\\ &= \exp\left\{\mu + \sigma^2/(2n) + \widetilde{\sigma}^2/2\right\}\\ &< \exp\left\{\mu + \sigma^2/2\right\} \;=\; \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{(\mu,\sigma^2)}\left[Z_i\right].\end{aligned}$$

This shows that working on the log-scale is rather difficult because the back-transformation is far from being trivial, and for unknown nuisance parameter not even the sign of the bias is clear. Similar considerations apply to the frequently used Box–Cox transformation [48] for $\chi \neq 1$

$$Z_i \;\mapsto\; Y_i = \frac{Z_i^{\chi} - 1}{\chi}.$$

For this reason, if unbiasedness is a central requirement (like in insurance pricing) non-linear transformations should only be used with great care (and only if necessary).

5.4 Quasi-Likelihoods

Above we have been mentioning the notion of over-dispersed Poisson models. This naturally leads to so-called quasi-Poisson models and quasi-likelihoods. The framework of quasi-likelihoods has been introduced by Wedderburn [376]. In this section we give the main idea behind quasi-likelihoods, and for a more detailed treatment and mathematical results we refer to Chapter 8 of McCullagh–Nelder [265].

In Sect. 5.1.4 we have discussed the estimation of GLMs. This has been based on the explicit knowledge of the full log-likelihood function $\ell_{\boldsymbol{Y}}(\boldsymbol{\beta})$ for given data $\boldsymbol{Y}$. This has allowed us to calculate the score equations $s(\boldsymbol{\beta}, \boldsymbol{Y}) = \nabla_{\boldsymbol{\beta}} \ell_{\boldsymbol{Y}}(\boldsymbol{\beta}) = 0$ whose solutions (Z-estimators) contain the MLE for $\boldsymbol{\beta}$. The solutions of the score equations themselves, using Fisher's scoring method, no longer need the explicit functional form of the log-likelihood, but they are only based on the first and second moments, see (5.9) and Remarks 5.4. Thus, all models where these first two moments coincide will provide the same MLE for the regression parameter $\boldsymbol{\beta}$; this is also the explanation behind the IRLS algorithm. Moreover, the first two moments are sufficient for prediction and uncertainty quantification based on mean squared errors, and they are also sufficient to quantify asymptotic normality. This is exactly what motivates the quasi-likelihood considerations, and these considerations are also related to the quasi-generalized pseudo maximum likelihood estimator (QPMLE) that we are going to discuss in Theorem 11.8, below.

Assume that $\boldsymbol{Y}$ is a random vector having first moment $\boldsymbol{\mu} \in \mathbb{R}^n$, positive definite variance function $V(\boldsymbol{\mu}) \in \mathbb{R}^{n \times n}$ and dispersion parameter φ. The quasi-(log-)likelihood function $\ell_{\boldsymbol{Y}}(\boldsymbol{\mu})$ assumes that its gradient is given by

$$\nabla_{\boldsymbol{\mu}} \ell_{\boldsymbol{Y}}(\boldsymbol{\mu}) = \frac{1}{\varphi} V(\boldsymbol{\mu})^{-1} \left(\boldsymbol{Y} - \boldsymbol{\mu}\right).$$

In case of a diagonal variance function $V(\boldsymbol{\mu})$ this relates to the score (5.9). The remaining step is to model the mean parameter $\boldsymbol{\mu} = \boldsymbol{\mu}(\boldsymbol{\beta}) \in \mathbb{R}^n$ as a function of a lower dimensional regression parameter $\boldsymbol{\beta} \in \mathbb{R}^{q+1}$, we also refer to Fig. 5.2. For this last step we assume that the Jacobian $B \in \mathbb{R}^{n\times(q+1)}$ of $d\boldsymbol{\mu}/d\boldsymbol{\beta}$ has full rank $q+1$. The score equations for $\boldsymbol{\beta}$ and given observations $\boldsymbol{Y}$ then read as

$$\frac{1}{\varphi} B^\top V(\boldsymbol{\mu}(\boldsymbol{\beta}))^{-1} \left(\boldsymbol{Y} - \boldsymbol{\mu}(\boldsymbol{\beta})\right) = 0.$$

This is of exactly the same structure as the score equations in Proposition 5.1, and the roots are found by using the IRLS algorithm for $t \geq 0$, see (5.12),

$$\widehat{\boldsymbol{\beta}}^{(t)} \mapsto \widehat{\boldsymbol{\beta}}^{(t+1)} = \left(B^\top V(\widehat{\boldsymbol{\mu}}^{(t)})^{-1} B\right)^{-1} B^\top V(\widehat{\boldsymbol{\mu}}^{(t)})^{-1} \left(B\widehat{\boldsymbol{\beta}}^{(t)} + \boldsymbol{Y} - \widehat{\boldsymbol{\mu}}^{(t)}\right),$$

where $\widehat{\boldsymbol{\mu}}^{(t)} = \boldsymbol{\mu}(\widehat{\boldsymbol{\beta}}^{(t)})$.

We conclude with the following points about quasi-likelihoods:

- For regression parameter estimation within the quasi-likelihood framework it is sufficient to know the structure of the first two moments $\boldsymbol{\mu}(\boldsymbol{\beta}) \in \mathbb{R}^n$ and $V(\boldsymbol{\mu}) \in \mathbb{R}^{n\times n}$ as well as the score equations. Thus, we do not need to explicitly specify a distributional family for the observations $\boldsymbol{Y}$. This structure of the first two moments is then sufficient for their estimation using the IRLS algorithm, i.e., we receive the predictors within this framework.
- Since we do not specify the full distribution of $\boldsymbol{Y}$ we can neither simulate from this model nor can we calculate quantities where the full log-likelihood of the model needs to be known. For example, we cannot calculate AIC in a quasi-likelihood model.
- The quasi-likelihood model is characterized by the functional forms of $\boldsymbol{\mu}(\boldsymbol{\beta})$ and $V(\boldsymbol{\mu})$. The former plays the role of the link function and the linear predictor in the GLM, and the latter plays the role of the variance function within the EDF which is characterized through the cumulant function κ. For instance, if we assume to have a diagonal matrix

  $$V(\boldsymbol{\mu}) = \mathrm{diag}(V(\mu_1), \ldots, V(\mu_n)),$$

  then, the choice of the variance function $\mu \mapsto V(\mu)$ describes the explicit selection of the quasi-likelihood model. If we choose the power variance function $V(\mu) = \mu^p$, $p \notin (0,1)$, we have a quasi-Tweedie's model.
- For prediction uncertainty evaluation we also need an estimate of the dispersion parameter $\varphi > 0$. Since we do not know the full likelihood in this approach, Pearson's estimate $\widehat{\varphi}^{\mathrm{P}}$ is the only option we have to estimate φ.
- For asymptotic normality results and hypothesis testing within the quasi-likelihood framework we refer to Section 8.4 of McCullagh–Nelder [265].

5.5 Double Generalized Linear Model

In the derivations above we have treated the dispersion parameter φ in the GLM as a nuisance parameter. In the case of a homogeneous dispersion parameter it can be canceled in the score equations for MLE, see (5.9). Therefore, it does not influence MLE, and in a subsequent step this nuisance parameter can still be estimated using, e.g., Pearson's or deviance residuals, see Sect. 5.3.1 and Remark 5.26. In some examples we may have systematic effects in the dispersion parameter, too. In this case the above approach will not work because a heterogeneous dispersion parameter no longer cancels in the score equations. This has been considered in Smyth [341] and Smyth–Verbyla [343]. The heterogeneous dispersion situation is of general interest for GLMs, and it is of particular interest for Tweedie's CP GLM if we interpret Tweedie's distribution [358] as a CP model with i.i.d. gamma claim sizes, see Proposition 2.17; we also refer to Jørgensen–de Souza [204], Smyth–Jørgensen [342] and Delong et al. [94].

5.5.1 *The Dispersion Submodel*

We extend model assumption (5.1) by assuming that also the dispersion parameter φ_i is policy i dependent. Assume that all random variables Y_i are independent and have densities w.r.t. a σ-finite measure ν on $\mathbb{R}$ given by

$$Y_i \sim f(y_i; \theta_i, v_i/\varphi_i) = \exp\left\{ \frac{y_i\theta_i - \kappa(\theta_i)}{\varphi_i/v_i} + a(y_i; v_i/\varphi_i) \right\},$$

for $1 \leq i \leq n$, with canonical parameters $\theta_i \in \mathring{\boldsymbol{\Theta}}$, exposures $v_i > 0$ and dispersion parameters $\varphi_i > 0$. As in (5.5) we assume that every policy i is equipped with feature information $\boldsymbol{x}_i \in \mathcal{X}$ such that for a given link function $g : \mathcal{M} \to \mathbb{R}$ we can model its mean as

$$\boldsymbol{x}_i \mapsto g(\mu_i) = g(\mu(\boldsymbol{x}_i)) = g\left(\mathbb{E}_{\theta(\boldsymbol{x}_i)}[Y_i]\right) = \eta_i = \eta(\boldsymbol{x}_i) = \langle \boldsymbol{\beta}, \boldsymbol{x}_i \rangle. \tag{5.56}$$

This provides us with log-likelihood function for observation $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^\top$

$$\boldsymbol{\beta} \mapsto \ell_{\boldsymbol{Y}}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \frac{v_i}{\varphi_i}\Big[Y_i h(\mu(\boldsymbol{x}_i)) - \kappa\left(h(\mu(\boldsymbol{x}_i))\right) \Big] + a(Y_i; v_i/\varphi_i),$$

with canonical link $h = (\kappa')^{-1}$. The difference to (5.7) is that the dispersion parameter φ_i now depends on the insurance policy which requires additional modeling. We choose a second strictly monotone and smooth link function g_φ :

$\mathbb{R}_+ \to \mathbb{R}$, and we express the dispersion of policy $1 \le i \le n$ by

$$g_\varphi(\varphi_i) = g_\varphi(\varphi(\boldsymbol{z}_i)) = \langle \boldsymbol{\gamma}, \boldsymbol{z}_i \rangle, \tag{5.57}$$

where $\boldsymbol{z}_i$ is the feature of policy i, which may potentially differ from $\boldsymbol{x}_i$. The rationale behind this different feature is that different information might be relevant for modeling the dispersion parameter, or feature information might be differently pre-processed compared to the response Y_i. We now need to estimate two regression parameters $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ in this approach on possibly differently pre-processed feature information $\boldsymbol{x}_i$ and $\boldsymbol{z}_i$ of policy i. In general, this is not easily doable because the term $a(Y_i; v_i/\varphi_i)$ of the log-likelihood of Y_i may have a complicated structure (or may not be available in closed form like in Tweedie's CP model).

5.5.2 Saddlepoint Approximation

We reformulate the EDF density using the unit deviance $\mathfrak{d}(Y, \mu)$ defined in (2.25); we drop the lower index i for the moment. Set $\theta = h(\mu) \in \mathring{\boldsymbol{\Theta}}$ for the canonical link h, then

$$\begin{aligned} f(y; \theta, v/\varphi) &= \exp\left\{ \frac{v}{\varphi} \left[y h(\mu) - \kappa(h(\mu)) \right] + a(y; v/\varphi) \right\} \\ &= \exp\left\{ \frac{v}{\varphi} \left[y h(y) - \kappa(h(y)) \right] + a(y; v/\varphi) \right\} \exp\left\{ -\frac{1}{2\varphi/v} \mathfrak{d}(y, \mu) \right\} \\ &\overset{\text{def.}}{=} a^*(y; \omega) \exp\left\{ -\frac{\omega}{2}\, \mathfrak{d}(y, \mu) \right\}, \end{aligned} \tag{5.58}$$

with $\omega = v/\varphi \in \mathcal{W}$. This corresponds to (2.27), and it brings the EDF density into a Gaussian-looking form. A general difficulty is that the term $a^*(y; \omega)$ may have a complicated structure or may not be given in closed form. Therefore, we consider its saddlepoint approximation; this is based on Section 3.5 of Jørgensen [203].

Suppose that we are in the absolutely continuous EDF case and that κ is steep. In that case $Y \in \mathcal{M}$, a.s., and the variance function $y \mapsto V(y)$ is well-defined for all observations $Y = y$, a.s. Based on Daniels [87], Barndorff-Nielsen–Cox [24] proved the following statement, see Theorem 3.10 in Jørgensen [203]: assume there exists $\omega_0 \in \mathcal{W}$ such that for all $\omega > \omega_0$ the density (5.58) is bounded. Then, the following saddlepoint approximation is uniform on compact subsets of the support $\mathfrak{T}$ of Y

$$f(y; \theta, v/\varphi) = \left(\frac{2\pi\varphi}{v} V(y) \right)^{-1/2} \exp\left\{ -\frac{1}{2\varphi/v}\, \mathfrak{d}(y, \mu) \right\} \left(1 + O(\varphi/v)\right), \tag{5.59}$$

as $\varphi/v \to 0$. What makes this saddlepoint approximation attractive is that we can get rid of a complicated function $a^*(y;\omega)$ by a neat approximation $(\frac{2\pi\varphi}{v}V(y))^{-1/2}$ for sufficiently large volumes v, and at the same time, this does not affect the unit deviance $\mathfrak{d}(y,\mu)$, preserving the estimation properties of μ. The discrete counterpart is given in Theorem 3.11 of Jørgensen [203].

Using saddlepoint approximation (5.59) we receive an approximate log-likelihood function

$$\ell_Y(\mu,\varphi) \;\approx\; \frac{1}{2}\left[-\varphi^{-1}v\mathfrak{d}(Y,\mu) - \log(\varphi)\right] - \frac{1}{2}\log\left(\frac{2\pi}{v}V(Y)\right).$$

This approximation has an attractive form for dispersion estimation because it gives an approximate EDF for observation $\mathfrak{d} \overset{\text{def.}}{=} v\mathfrak{d}(Y,\mu)$, for given μ. Namely, for canonical parameter $\phi = -\varphi^{-1} < 0$ we have approximation

$$\ell_Y(\mu,\phi) \;\approx\; \frac{\mathfrak{d}\phi - (-\log(-\phi))}{2} - \frac{1}{2}\log\left(\frac{2\pi}{v}V(Y)\right). \tag{5.60}$$

The right-hand side has the structure of a gamma EDF for observation $\mathfrak{d}$ with canonical parameter $\phi < 0$, cumulant function $\kappa_\varphi(\phi) = -\log(-\phi)$ and dispersion parameter 2. Thus, we have the structure of an approximate gamma model on the right-hand side of (5.60) with, for given μ,

$$\mathbb{E}_\phi[\mathfrak{d}|\mu] \approx \kappa'_\varphi(\phi) \;=\; -\frac{1}{\phi} \;=\; \varphi, \tag{5.61}$$

$$\mathrm{Var}_\phi(\mathfrak{d}|\mu) \approx 2\kappa''_\varphi(\phi) \;=\; 2\frac{1}{\phi^2} \;=\; 2\varphi^2. \tag{5.62}$$

These statements say that for given μ and assuming that the saddlepoint approximation is sufficiently accurate, $\mathfrak{d}$ is approximately gamma distributed with shape parameter 1/2 and canonical parameter ϕ (which relates to the dispersion φ in the mean parametrization). Thus, we can estimate ϕ and φ, respectively, with a (second) GLM from (5.60), for given mean parameter μ.

Remarks 5.28

- The accuracy of the saddlepoint approximation is discussed in Section 3.2 of Smyth–Verbyla [343]. The saddlepoint approximation is exact in the Gaussian and the inverse Gaussian case. In the Gaussian case, we have log-likelihood

$$\ell_Y(\mu,\phi) = \frac{\mathfrak{d}\phi - (-\log(-\phi))}{2} - \frac{1}{2}\log\left(\frac{2\pi}{v}\right),$$

with variance function $V(Y) = 1$. In the inverse Gaussian case, we have log-likelihood

$$\ell_Y(\mu, \phi) = \frac{\mathfrak{d}\phi - (-\log(-\phi))}{2} - \frac{1}{2}\log\left(\frac{2\pi}{v}Y^3\right),$$

with variance function $V(Y) = Y^3$. Thus, in the Gaussian case and in the inverse Gaussian case we have a gamma model for $\mathfrak{d}$ with mean φ and shape parameter $1/2$, for given μ; for a related result we also refer to Theorem 3 of Blæsild–Jensen [38]. For Tweedie's models with $p \geq 1$, one can show that the relative error of the saddlepoint approximation is a non-increasing function of the squared coefficient of variation $\tau = \frac{\varphi}{v}V(y)/y^2 = \frac{\varphi}{v}y^{p-2}$, leading to small approximation errors if φ/v is sufficiently small; typically one requires $\tau < 1/3$, see Section 3.2 of Smyth–Verbyla [343].

- The saddlepoint approximation itself does not provide a density because in general the term $O(\varphi/v)$ in (5.59) is non-zero. Nelder–Pregibon [282] renormalized the saddlepoint approximation to a proper density and studied its properties.
- In the gamma EDF case, the saddlepoint approximation would not be necessary because this case can still be solved in closed form. In fact, in the gamma EDF case we have log-likelihood, set $\phi = -v/\varphi < 0$,

$$\ell_Y(\mu, \phi) = \frac{\phi\mathfrak{d}(Y, \mu) - \chi(\phi)}{2} - \log Y, \tag{5.63}$$

with $\chi(\phi) = 2(\log\Gamma(-\phi) + \phi\log(-\phi) - \phi)$. For given μ, this is an EDF for $\mathfrak{d}(Y, \mu)$ with cumulant function χ on the effective domain $(-\infty, 0)$. This provides us with expected value and variance

$$\begin{aligned}
\mathbb{E}_\phi[\mathfrak{d}(Y, \mu)|\mu] = \chi'(\phi) \;&=\; 2\left(-\Psi(-\phi) + \log(-\phi)\right) \;\approx\; -\frac{1}{\phi},\\
\mathrm{Var}_\phi(\mathfrak{d}(Y, \mu)|\mu) = 2\chi''(\phi) \;&=\; 4\left(\Psi'(-\phi) - \frac{1}{-\phi}\right),
\end{aligned}$$

with digamma function Ψ and the approximation exactly refers to the saddlepoint approximation; for the variance statement we also refer to Fisher's information (2.30). For receiving more accurate mean approximations one can consider higher order terms, e.g., the second order approximation is $\chi'(\phi) \approx -1/\phi + 1/(6\phi^2)$. In fact, from the saddlepoint approximation (5.60) and from the exact formula (5.63) we receive in the gamma case Stirling's formula

$$\Gamma(\gamma) \;\approx\; \sqrt{2\pi}\gamma^{\gamma-1/2}e^{-\gamma}.$$

In the subsequent examples we will just use the saddlepoint approximation also in the gamma EDF case.

5.5.3 Residual Maximum Likelihood Estimation

The saddlepoint approximation (5.60) proposes to alternate MLE of $\boldsymbol{\beta}$ for the mean model (5.56) and of $\boldsymbol{\gamma}$ for the dispersion model (5.57). Fisher's information matrix of the saddlepoint approximation (5.60) w.r.t. the canonical parameters θ and ϕ is given by

$$\mathcal{I}(\theta, \phi) = -\mathbb{E}_{\theta,\phi}\begin{pmatrix} \phi v \kappa''(\theta) & -v\left(Y - \kappa'(\theta)\right) \\ -v\left(Y - \kappa'(\theta)\right) & -\frac{1}{2}\frac{1}{\phi^2} \end{pmatrix} = \begin{pmatrix} \frac{v}{\varphi(\phi)} V(\mu(\theta)) & 0 \\ 0 & \frac{1}{2} V_\varphi(\varphi(\phi)) \end{pmatrix},$$

with variance function $V_\varphi(\varphi) = \varphi^2$, and emphasizing that we work in the canonical parametrization (θ, ϕ). This is a positive definite diagonal matrix which suggests that the algorithm alternating the $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ estimations will have a fast convergence. For fixed estimate $\widehat{\boldsymbol{\gamma}}$ we calculate estimated dispersion parameters $\widehat{\varphi}_i = g_\varphi^{-1}\langle \widehat{\boldsymbol{\gamma}}, \boldsymbol{z}_i \rangle$ of policies $1 \le i \le n$, see (5.57). These then allow us to calculate diagonal working weight matrix

$$W(\boldsymbol{\beta}) = \operatorname{diag}\left(\left(\frac{\partial g(\mu_i)}{\partial \mu_i}\right)^{-2} \frac{v_i}{\widehat{\varphi}_i} \frac{1}{V(\mu_i)}\right)_{1 \le i \le n} \in \mathbb{R}^{n \times n},$$

which is used in Fisher's scoring method/IRLS algorithm (5.12) to receive MLE $\widehat{\boldsymbol{\beta}}$, given the estimates $(\widehat{\varphi}_i)_i$. These MLEs allow us to estimate the mean parameters $\widehat{\mu}_i = g^{-1}\langle \widehat{\boldsymbol{\beta}}, \boldsymbol{x}_i \rangle$, and to calculate the deviances

$$\mathfrak{d}_i = v_i \mathfrak{d}\left(Y_i, \widehat{\mu}_i\right) = 2 v_i \Big(Y_i h\left(Y_i\right) - \kappa\left(h\left(Y_i\right)\right) - Y_i h\left(\widehat{\mu}_i\right) + \kappa\left(h\left(\widehat{\mu}_i\right)\right) \Big) \ \ge\ 0.$$

Using (5.60) we know that these deviances can be approximated by gamma distributions $\Gamma(1/2, 1/(2\varphi_i))$. This is a single-parameter EDF with dispersion parameter 2 (as nuisance parameter) and mean parameter φ_i. This motivates the definition of the working weight matrix (based on the gamma EDF model)

$$W_\varphi(\boldsymbol{\gamma}) = \operatorname{diag}\left(\left(\frac{\partial g_\varphi(\varphi_i)}{\partial \varphi_i}\right)^{-2} \frac{1}{2} \frac{1}{V_\varphi(\varphi_i)}\right)_{1 \le i \le n} \in \mathbb{R}^{n \times n},$$

and the working residuals

$$\boldsymbol{R}_\varphi(\mathfrak{d}, \boldsymbol{\gamma}) = \left(\frac{\partial g_\varphi(\varphi_i)}{\partial \varphi_i}(\mathfrak{d}_i - \varphi_i)\right)_{1 \le i \le n}^{\top} \in \mathbb{R}^{n}.$$

Fisher's scoring method (5.12) iterates for $s \geq 0$ the following recursion to receive $\widehat{\boldsymbol{\gamma}}$

$$\widehat{\boldsymbol{\gamma}}^{(s)} \mapsto \widehat{\boldsymbol{\gamma}}^{(s+1)} = \left(\mathfrak{Z}^\top W_\varphi(\widehat{\boldsymbol{\gamma}}^{(s)})\mathfrak{Z}\right)^{-1} \mathfrak{Z}^\top W_\varphi(\widehat{\boldsymbol{\gamma}}^{(s)}) \left(\mathfrak{Z}\widehat{\boldsymbol{\gamma}}^{(s)} + \boldsymbol{R}_\varphi(\mathfrak{d}, \widehat{\boldsymbol{\gamma}}^{(s)})\right), \tag{5.64}$$

where $\mathfrak{Z} = (\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n)^\top$ is the design matrix used to estimate the dispersion parameters.

5.5.4 Lab: Double GLM Algorithm for Gamma Claim Sizes

We revisit the Swedish motorcycle claim size data studied in Sect. 5.3.7. We expand the gamma claim size GLM to a double GLM also modeling the systematic effects in the dispersion parameter. In a first step we need to change the parametrization of the gamma model of Sect. 5.3.7. In the former section we have modeled the average claim size $S_i/n_i \sim \Gamma(n_i\alpha_i, n_ic_i)$, but for applying the saddlepoint approximation we should use the reproductive form (5.44) of the gamma model. We therefore set

$$Y_i = S_i/(n_i\alpha_i) \sim \Gamma(n_i\alpha_i, n_i\alpha_ic_i). \tag{5.65}$$

The reason for the different parametrization in Sect. 5.3.7 has been that (5.65) is not directly useful if α_i is unknown because in that case the observations Y_i cannot be calculated. In this section we estimate $\varphi_i = 1/\alpha_i$ which allows us to model (5.65); a different treatment within Tweedie's family is presented in Sect. 11.1.3. The only difficulty is to initialize the double GLM algorithm. We proceed as follows.

(0) In an initial step we assume constant dispersion $\varphi_i = 1/\alpha_i \equiv 1/\alpha = 1$. This gives us exactly the mean estimates of Sect. 5.3.7 for $S_i/n_i \sim \Gamma(n_i\alpha, n_ic_i)$; note that for constant shape parameter α the mean of S_i/n_i can be estimated without explicit knowledge of α (because it cancels in the score equations). Using these mean estimates we calculate the MLE $\widehat{\alpha}^{(0)}$ of the (constant) shape parameter α, see Remark 5.26. This then allows us to determine the (scaled) observations $Y_i^{(1)} = S_i/(n_i\widehat{\alpha}^{(0)})$ and we initialize $\widehat{\varphi}_i^{(0)} = 1/\widehat{\alpha}^{(0)}$.

(1) Iterate for $t \geq 1$:

- estimate the mean μ_i of Y_i using the mean GLM (5.56) based on the observations $Y_i^{(t)}$ and the dispersion estimates $\widehat{\varphi}_i^{(t-1)}$. This provides us with $\widehat{\mu}_i^{(t)}$;
- based on the deviances $\mathfrak{d}_i^{(t)} = v_i\mathfrak{d}(Y_i^{(t)}, \widehat{\mu}_i^{(t)})$, calculate the updated dispersion estimates $\widehat{\varphi}_i^{(t)}$ using the dispersion GLM (5.57) and the residual MLE iteration (5.64) with the saddlepoint approximation. Set for the updated observations $Y_i^{(t+1)} = S_i\widehat{\varphi}_i^{(t)}/n_i$.

Table 5.15 Number of parameters, AICs, Pearson's dispersion estimate, in-sample losses, tenfold cross-validation losses and the in-sample average claim amounts of the null model (gamma intercept model) and the (double) gamma GLM

	# Param.	AIC	Dispersion est. $\widehat{\varphi}^{\text{P}}$	In-sample loss on $\mathcal{L}$	Tenfold CV loss $\widehat{\mathfrak{D}}^{\text{CV}}$	Average amount
Gamma null	$1+1$	14'416	2.057	2.085	2.091	24'641
Gamma GLM2	$7+1$	14'274	1.544	1.719	1.747	25'130
Double gamma GLM	$7+6$	14'258	–	(1.721)	–	26'413

In an initial double GLM analysis we use the feature information $z_i = \boldsymbol{x}_i$ for the dispersion φ_i modeling (5.57). We choose for both GLMs the log-link which leads to concave maximization problems, see Example 5.5. Running the above double GLM algorithm converges in 4 iterations, and analyzing the resulting model we observe that we should drop the variable `RiskClass` from the feature z_i. We then run the same double GLM algorithm with the feature information $\boldsymbol{x}_i$ and the new z_i again, and the results are presented in Table 5.15.

The considered double GLM has parameter dimensions $\boldsymbol{\beta} \in \mathbb{R}^7$ and $\boldsymbol{\gamma} \in \mathbb{R}^6$. To have comparability with AIC of Sect. 5.3.7, we evaluate AIC of the double GLM in the observations S_i/n_i (and not in Y_i; i.e., similar to the gamma GLM). We observe that it has an improved AIC value compared to model Gamma GLM2. Thus, indeed, dispersion modeling seems necessary in this example (under the GLM2 regression structure). We do not calculate in-sample and cross-validation losses in the double GLM because in the other two models of Table 5.15 we have set $\varphi = 1$ in these statistics. However, the in-sample loss of model Gamma GLM2 with $\varphi = 1$ corresponds to the (homogeneous) deviance dispersion estimate (up to scaling $n/(n-(q+1))$), and this in-sample loss of 1.719 can directly be compared to the average estimated dispersion $m^{-1}\sum_{i=1}^{m}\widehat{\varphi}_i = 1.721$ (in round brackets in Table 5.15). On the downside, the double GLM has a bigger bias which needs an adjustment.

In Fig. 5.13 (lhs) we give the normal plots of model Gamma GLM2 and the double gamma GLM model. This plot is received by transforming the observations to normal quantiles using the corresponding estimated gamma models. We see quite some similarity between the two estimated gamma models. Both models seem to have similar deficiencies, i.e., dispersion modeling improves explanation of observations, however, either the regression function or the gamma distributional assumption does not fully fit the data, especially for small claims. Finally, in Fig. 5.13 (rhs) we plot the estimated dispersion parameters $\widehat{\varphi}_i$ against the logged estimated means $\log(\widehat{\mu}_i)$ (linear predictors). We observe that the estimated dispersion has a (weak) U-shape as a function of the expected claim sizes which indicates that the tails cannot fully be captured by our model. This closes this example.

Remark 5.29 For the dispersion estimation $\widehat{\varphi}_i$ we use as observations the deviances $\mathfrak{d}_i = v_i \mathfrak{d}(Y_i, \widehat{\mu}_i)$, $1 \le i \le n$. On a finite sample, these deviances are typically biased due to the use of the estimated means $\widehat{\mu}_i$. Smyth–Verbyla [343] propose the

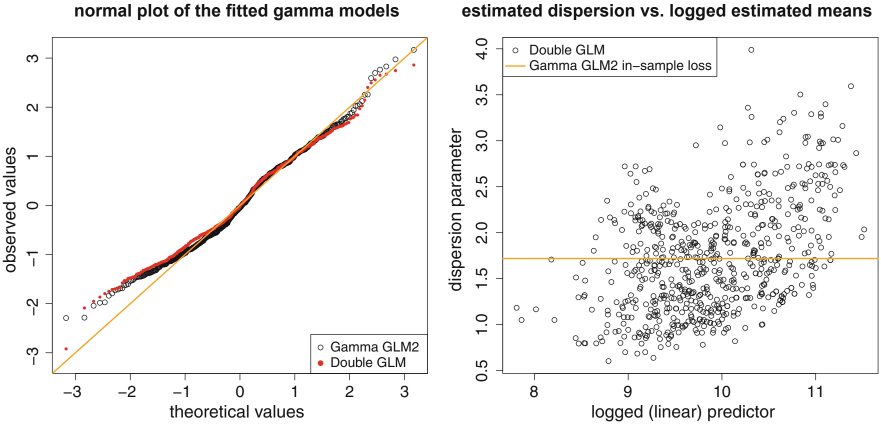


Fig. 5.13 (lhs) Normal plot of the fitted models Gamma GLM2 and double GLM, (rhs) estimated dispersion parameters $\widehat{\varphi}_i$ against the logged estimated means $\log(\widehat{\mu}_i)$ (the orange line gives the in-sample loss in model Gamma GLM2)

following bias correction. Consider the estimated hat matrix defined by

$$H = W(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\gamma}})^{1/2}\mathfrak{X} \left(\mathfrak{X}^\top W(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\gamma}})\,\mathfrak{X}\right)^{-1} \mathfrak{X}^\top W(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\gamma}})^{1/2},$$

with the diagonal work weight matrix $W(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\gamma}})$ depending on the estimated regression parameters $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\gamma}}$ through μ and φ. Denote the diagonal entries of the hat matrix by $(h_{i,i})_{1\le i\le n}$. A bias corrected version of the deviances is received by considering observations $(1-h_{i,i})^{-1}\mathfrak{d}_i = (1-h_{i,i})^{-1}v_i\mathfrak{d}\,(Y_i, \widehat{\mu}_i)$, $1 \le i \le n$. We will come back to the hat matrix H in Sect. 5.6.1, below.

5.5.5 Tweedie's Compound Poisson GLM

A popular situation for applying the double GLM framework is Tweedie's CP model introduced in Sect. 2.2.3, in particular, we refer to Proposition 2.17 for the corresponding parametrization. Having claim frequency and claim sizes involved, such a model can hardly be calibrated with one single regression function and a constant dispersion parameter. An obvious choice is a double GLM, this is the proposal presented in Smyth–Jørgensen [342]. In most of the cases one chooses for both link functions g and g_φ the log-links because positivity needs to be guaranteed.

This implies for the two working weight matrices of the double GLM

$$W(\boldsymbol{\beta}) = \mathrm{diag}\left(\mu_i^2 \frac{v_i}{\varphi_i}\,\frac{1}{V(\mu_i)}\right)_{1\le i\le n} = \mathrm{diag}\left(\mu_i^{2-p}\frac{v_i}{\varphi_i}\right)_{1\le i\le n},$$

$$W_\varphi(\boldsymbol{\gamma}) = \mathrm{diag}\left(\varphi_i^2\,\frac{1}{2}\,\frac{1}{V_\varphi(\varphi_i)}\right)_{1\le i\le n} = \mathrm{diag}(1/2, \ldots, 1/2).$$

The deviances in Tweedie's CP model are given by, see (4.18),

$$\mathfrak{d}_i = v_i\mathfrak{d}\left(Y_i, \widehat{\mu}_i\right) = 2v_i\left(Y_i\frac{Y_i^{1-p} - \widehat{\mu}_i^{\,1-p}}{1-p} - \frac{Y_i^{2-p} - \widehat{\mu}_i^{\,2-p}}{2-p}\right) \ge 0,$$

and these deviances could still be de-biased, see Remark 5.29. The working responses for the two GLMs are

$$\boldsymbol{R} = \left(Y_i/\mu_i - 1\right)^\top_{1\le i\le n} \qquad \text{and} \qquad \boldsymbol{R}_\varphi = \left(\mathfrak{d}_i/\varphi_i - 1\right)^\top_{1\le i\le n}.$$

The drawback of this approach is that it only considers the (scaled) total claim amounts $Y_i = S_i\varphi_i/v_i$ as observations, see Proposition 2.17. These total claim amounts consist of the number of claims N_i and i.i.d. individual claim sizes $Z_{i,j} \sim \Gamma(\alpha, c_i)$, supposed $N_i \ge 1$. Having observations of both claim amounts S_i and claim counts N_i allows one to build a Poisson GLM for claim counts and a gamma GLM for claim sizes which can be estimated separately. This has also been the reason of Smyth–Jørgensen [342] to enhance Tweedie's model estimation for known claim counts in their Section 4. Moreover, in Theorem 4 of Delong et al. [94] it is proved that the two GLM approaches can be identified under log-link choices.

5.6 Diagnostic Tools

In our examples we have studied several figures like AIC, cross-validation losses, etc., for model and parameter selection. Moreover, we have plotted the results, for instance, using the Tukey–Anscombe plot or the QQ plot. Of course, there are numerous other plots and tools that can help us to analyze the results and to improve the resulting models. We present some of these in this section.

5.6.1 The Hat Matrix

The MLE $\widehat{\boldsymbol{\beta}}^{\mathrm{MLE}}$ satisfies at convergence of the IRLS algorithm, see (5.12),

$$\widehat{\boldsymbol{\beta}}^{\mathrm{MLE}} = \left(\mathfrak{X}^\top W(\widehat{\boldsymbol{\beta}}^{\mathrm{MLE}})\mathfrak{X}\right)^{-1}\mathfrak{X}^\top W(\widehat{\boldsymbol{\beta}}^{\mathrm{MLE}})\left(\mathfrak{X}\widehat{\boldsymbol{\beta}}^{\mathrm{MLE}} + \boldsymbol{R}(\boldsymbol{Y}, \widehat{\boldsymbol{\beta}}^{\mathrm{MLE}})\right),$$

with working residuals for $\boldsymbol{\beta} \in \mathbb{R}^{q+1}$

$$\boldsymbol{R}(\boldsymbol{Y}, \boldsymbol{\beta}) = \left(\left. \frac{\partial g(\mu_i)}{\partial \mu_i} \right|_{\mu_i = \mu_i(\boldsymbol{\beta})} (Y_i - \mu_i(\boldsymbol{\beta})) \right)^{\top}_{1 \le i \le n} \in \mathbb{R}^n.$$

Following Section 4.2.2 of Fahrmeir–Tutz [123], this allows us to define the so-called *hat matrix*, see also Remark 5.29,

$$H = H(\widehat{\boldsymbol{\beta}}^{\text{MLE}}) = W(\widehat{\boldsymbol{\beta}}^{\text{MLE}})^{1/2} \mathfrak{X} \left(\mathfrak{X}^{\top} W(\widehat{\boldsymbol{\beta}}^{\text{MLE}}) \mathfrak{X} \right)^{-1} \mathfrak{X}^{\top} W(\widehat{\boldsymbol{\beta}}^{\text{MLE}})^{1/2} \in \mathbb{R}^{n \times n}, \tag{5.66}$$

recall that the working weight matrix $W(\boldsymbol{\beta})$ is diagonal. The hat matrix H is symmetric and idempotent, i.e. $H^2 = H$, with $\text{trace}(H) = \text{rank}(H) = q + 1$. Therefore, H acts as a projection, mapping the observations $\widetilde{\boldsymbol{Y}}$ to the fitted values

$$\widetilde{\boldsymbol{Y}} \stackrel{\text{def.}}{=} W(\widehat{\boldsymbol{\beta}}^{\text{MLE}})^{1/2} \left(\mathfrak{X} \widehat{\boldsymbol{\beta}}^{\text{MLE}} + \boldsymbol{R}(\boldsymbol{Y}, \widehat{\boldsymbol{\beta}}^{\text{MLE}}) \right) \mapsto H\widetilde{\boldsymbol{Y}} = W(\widehat{\boldsymbol{\beta}}^{\text{MLE}})^{1/2} \mathfrak{X} \widehat{\boldsymbol{\beta}}^{\text{MLE}}$$
$$= W(\widehat{\boldsymbol{\beta}}^{\text{MLE}})^{1/2} \widehat{\boldsymbol{\eta}},$$

the latter being the fitted linear predictors. The diagonal elements $h_{i,i}$ of this hat matrix H satisfy $0 \le h_{i,i} \le 1$, and values close to 1 correspond to extreme data points i, in particular, for $h_{i,i} = 1$ only observation $\widetilde{Y}_i$ influences $\widehat{\eta}_i$, whereas for $h_{i,i} = 0$ observation $\widetilde{Y}_i$ has no influence on $\widehat{\eta}_i$.

Figure 5.14 gives the resulting hat matrices of the double gamma GLM of Sect. 5.5.4. On the left-hand side we show the diagonal entries $h_{i,i}$ of the claim

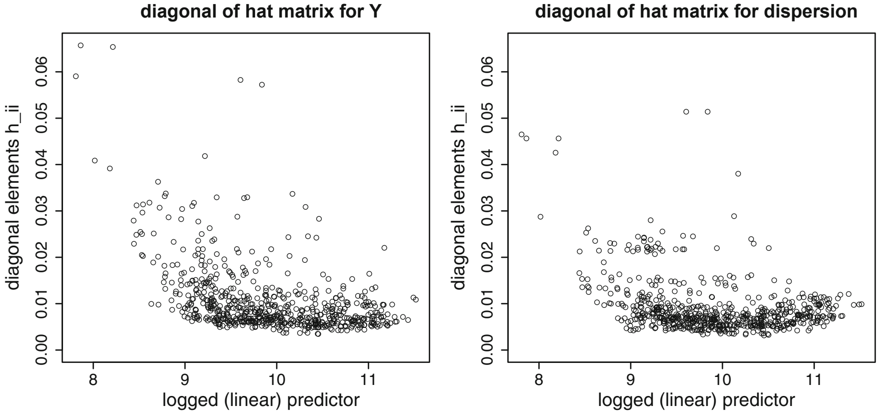


Fig. 5.14 Diagonal entries $h_{i,i}$ of the two hat matrices of the example in Sect. 5.5.4: (lhs) for means $\widehat{\mu}_i$ and responses Y_i, and (rhs) for dispersions $\widehat{\varphi}_i$ and responses $\mathfrak{d}_i$

amount responses Y_i (for the estimation of μ_i), and on the right-hand side the corresponding plots for the deviance responses $\mathfrak{d}_i$ (for the estimation of φ_i). These diagonal elements $h_{i,i}$ are ordered on the x-axis w.r.t. the linear predictors $\widehat{\eta}_i$. From this figure we conclude that the diagonal entries of the hat matrices are bigger for very small responses in our example, and the dispersion plot has a couple of more special observations that may require further analysis.

5.6.2 Case Deletion and Generalized Cross-Validation

As a continuation of the previous subsection we can analyze the influence of an individual observation Y_i on the estimation of regression parameter $\boldsymbol{\beta}$. This influence is naturally measured by fitting the regression parameter based on the full data $\mathcal{D}$ and based only on the observations $\mathcal{L}_{(-i)} = \mathcal{D} \setminus \{Y_i\}$, we also refer to leave-one-out cross-validation in Sect. 4.2.2. The influence of observation Y_i is then obtained by comparing $\widehat{\boldsymbol{\beta}}^{\mathrm{MLE}}$ and $\widehat{\boldsymbol{\beta}}^{\mathrm{MLE}}_{(-i)}$. Since fitting n different models by individually leaving out each observation Y_i is too costly, one only explores a one-step Fisher's scoring update starting from $\widehat{\boldsymbol{\beta}}^{\mathrm{MLE}}$ that provides an approximation to $\widehat{\boldsymbol{\beta}}^{\mathrm{MLE}}_{(-i)}$, that is,

$$\begin{aligned}\widehat{\boldsymbol{\beta}}^{(1)}_{(-i)} &= \left(\mathfrak{X}^\top_{(-i)} W_{(-i)}(\widehat{\boldsymbol{\beta}}^{\mathrm{MLE}})\mathfrak{X}_{(-i)}\right)^{-1}\mathfrak{X}^\top_{(-i)} W_{(-i)}(\widehat{\boldsymbol{\beta}}^{\mathrm{MLE}})\left(\mathfrak{X}\widehat{\boldsymbol{\beta}}^{\mathrm{MLE}} + \boldsymbol{R}(\boldsymbol{Y},\widehat{\boldsymbol{\beta}}^{\mathrm{MLE}})\right)_{(-i)}\\ &= \left(\mathfrak{X}^\top_{(-i)} W_{(-i)}(\widehat{\boldsymbol{\beta}}^{\mathrm{MLE}})\mathfrak{X}_{(-i)}\right)^{-1}\mathfrak{X}^\top_{(-i)} W_{(-i)}(\widehat{\boldsymbol{\beta}}^{\mathrm{MLE}})^{1/2}\,\widetilde{\boldsymbol{Y}}_{(-i)},\end{aligned}$$

where all lower indices ${}_{(-i)}$ indicate that we drop the corresponding row or/and column from the matrices and vectors, and where $\widetilde{\boldsymbol{Y}}$ has been defined in the previous subsection. This allows us to compare $\widehat{\boldsymbol{\beta}}^{\mathrm{MLE}}$ and $\widehat{\boldsymbol{\beta}}^{(1)}_{(-i)}$ to analyze the influence of observation Y_i.

To reformulate this approximation, we come back to the hat matrix $H = H(\widehat{\boldsymbol{\beta}}^{\mathrm{MLE}}) = (h_{i,j})_{1\le i,j\le n}$ defined in (5.66). It fulfills

$$W(\widehat{\boldsymbol{\beta}}^{\mathrm{MLE}})^{1/2}\mathfrak{X}\widehat{\boldsymbol{\beta}}^{\mathrm{MLE}} = H\widetilde{\boldsymbol{Y}} = \left(\sum_{j=1}^n h_{1,j}\widetilde{Y}_j, \ldots, \sum_{j=1}^n h_{n,j}\widetilde{Y}_j\right)^\top \in \mathbb{R}^n.$$

Thus, for predicting Y_i we can consider the linear predictor (for the chosen link g)

$$\widehat{\eta}_i = g(\widehat{\mu}_i) = \langle \widehat{\boldsymbol{\beta}}^{\mathrm{MLE}}, \boldsymbol{x}_i\rangle = (\mathfrak{X}\widehat{\boldsymbol{\beta}}^{\mathrm{MLE}})_i = W_{i,i}(\widehat{\boldsymbol{\beta}}^{\mathrm{MLE}})^{-1/2}\sum_{j=1}^n h_{i,j}\widetilde{Y}_j.$$

A computation of the linear predictor of Y_i using the leave-one-out approximation $\widehat{\boldsymbol{\beta}}^{(1)}_{(-i)}$ gives

$$\widehat{\eta}_i^{(-i,1)} = \langle \widehat{\boldsymbol{\beta}}^{(1)}_{(-i)}, \boldsymbol{x}_i \rangle = \frac{1}{1-h_{i,i}} \widehat{\eta}_i - W_{i,i}(\widehat{\boldsymbol{\beta}}^{\mathrm{MLE}})^{-1/2} \frac{h_{i,i}}{1-h_{i,i}} \widetilde{Y}_i.$$

This allows one to efficiently calculate a leave-one-out prediction using the hat matrix H. This also motivates to study the *generalized cross-validation* (GCV) loss which is an approximation to leave-one-out cross-validation, see Sect. 4.2.2,

$$\widehat{\mathfrak{D}}^{\mathrm{GCV}} = \frac{1}{n} \sum_{i=1}^{n} \frac{v_i}{\varphi}\, \mathfrak{d}\left(Y_i, g^{-1}(\widehat{\eta}_i^{(-i,1)})\right) \tag{5.67}$$

$$= \frac{2}{n} \sum_{i=1}^{n} \frac{v_i}{\varphi} \Big[Y_i h\left(Y_i\right) - \kappa\left(h\left(Y_i\right)\right) - Y_i h\left(g^{-1}(\widehat{\eta}_i^{(-i,1)})\right) + \kappa\left(h\left(g^{-1}(\widehat{\eta}_i^{(-i,1)})\right)\right) \Big].$$

Example 5.30 (Generalized Cross-Validation Loss in the Gaussian Case) We study the generalized cross-validation loss $\widehat{\mathfrak{D}}^{\mathrm{GCV}}$ in the homoskedastic Gaussian case $v_i/\varphi \equiv 1/\sigma^2$ with cumulant function $\kappa(\theta) = \theta^2/2$ and canonical link $g(\mu) = h(\mu) = \mu$. The generalized cross-validation loss in the Gaussian case is given by

$$\widehat{\mathfrak{D}}^{\mathrm{GCV}} = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{\sigma^2} \left(Y_i - \widehat{\eta}_i^{(-i,1)} \right)^2,$$

with (linear) leave-one-out predictor

$$\widehat{\eta}_i^{(-i,1)} = \langle \widehat{\boldsymbol{\beta}}^{(1)}_{(-i)}, \boldsymbol{x}_i \rangle = \sum_{j=1, j\neq i}^{n} \frac{h_{i,j}}{1-h_{i,i}} Y_j = \frac{1}{1-h_{i,i}} \widehat{\eta}_i - \frac{h_{i,i}}{1-h_{i,i}} Y_i.$$

This gives us generalized cross-validation loss in the Gaussian case

$$\widehat{\mathfrak{D}}^{\mathrm{GCV}} = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{\sigma^2} \left(\frac{Y_i - \widehat{\eta}_i}{1-h_{i,i}} \right)^2,$$

with $\boldsymbol{\beta}$ independent hat matrix

$$H = \mathfrak{X} \left(\mathfrak{X}^\top \mathfrak{X} \right)^{-1} \mathfrak{X}^\top.$$

The generalized cross-validation loss is used, for instance, for generalized additive model (GAM) fitting where an efficient and fast cross-validation method is required to select regularization parameters. Generalized cross-validation has been introduced by Craven–Wahba [84] but these authors replaced $h_{i,i}$ by $\sum_{j=1}^{n} h_{j,j}/n$. It holds that $\sum_{j=1}^{n} h_{j,j} = \text{trace}(H) = q + 1$, thus, using this approximation we receive

$$\widehat{\mathfrak{D}}^{\text{GCV}} \approx \frac{1}{n}\sum_{i=1}^{n}\frac{1}{\sigma^2}\left(\frac{Y_i - \widehat{\eta}_i}{1-\sum_{j=1}^{n} h_{j,j}/n}\right)^2 = \frac{n}{(n-(q+1))^2}\sum_{i=1}^{n}\frac{(Y_i-\widehat{\eta}_i)^2}{\sigma^2}$$
$$= \frac{n}{n-(q+1)}\frac{\widehat{\varphi}^{\text{P}}}{\sigma^2},$$

with $\widehat{\varphi}^{\text{P}}$ being Pearson's dispersion estimate in the Gaussian model, see (5.30). ■

We give a numerical example based on the gamma GLM for the claim sizes studied in Sect. 5.3.7.

Example 5.31 (Leave-One-Out Cross-Validation) The aim of this example is to compare the generalized cross-validation loss $\widehat{\mathfrak{D}}^{\text{GCV}}$ to the leave-one-out cross-validation loss $\widehat{\mathfrak{D}}^{\text{loo}}$, see (4.34), the former being an approximation to the latter. We do this for the gamma claim size model studied in Sect. 5.3.7. In this example it is feasible to exactly calculate the leave-one-out cross-validation loss because we have only 656 claims.

The results are presented in Table 5.16. Firstly, the different cross-validation losses confirm that the model slightly (in-sample) over-fits to the data, which is not a surprise when estimating 7 regression parameters based on 656 observations. Secondly, the cross-validation losses provide similar numbers with leave-one-out being slightly bigger than tenfold cross-validation, here. Thirdly, the generalized cross-validation loss $\widehat{\mathfrak{D}}^{\text{GCV}}$ manages to approximate the leave-one-out cross-validation loss $\widehat{\mathfrak{D}}^{\text{loo}}$ very well in this example.

Table 5.17 gives the corresponding results for model Poisson GLM1 of Sect. 5.2.4. Firstly, in this example with 610'206 observations it is not feasible to calculate the leave-one-out cross-validation loss (for computational reasons). Therefore, we rely on the generalized cross-validation loss as an approximation. From the results of Table 5.17 it seems that this approximation (rather) underestimates the loss (compared to tenfold cross-validation). Indeed, this is an observation that we have made also in other examples. ■

Table 5.16 Comparison of different cross-validation losses for model Gamma GLM2

	Gamma GLM2
In-sample loss $\mathfrak{D}(\mathcal{L}, \widehat{\mu}_{\mathcal{L}}^{\text{MLE}})$	1.719
Tenfold CV loss $\widehat{\mathfrak{D}}^{\text{CV}}$	1.747
Leave-one-out CV loss $\widehat{\mathfrak{D}}^{\text{loo}}$	1.756
Generalized CV loss $\widehat{\mathfrak{D}}^{\text{GCV}}$	1.758

Table 5.17 Comparison of different cross-validation losses for model Poisson GLM1

	Poisson GLM1
In-sample loss $\mathfrak{D}(\mathcal{L}, \widehat{\mu}_{\mathcal{L}}^{\mathrm{MLE}})$	24.101
Tenfold CV loss $\widehat{\mathfrak{D}}^{\mathrm{CV}}$	24.121
Leave-one-out CV loss $\widehat{\mathfrak{D}}^{\mathrm{loo}}$	N/A
Generalized CV loss $\widehat{\mathfrak{D}}^{\mathrm{GCV}}$	24.105

5.7 Generalized Linear Models with Categorical Responses

The reader will have noticed that the discussion of GLMs in this chapter has been focusing on the single-parameter linear EDF case (5.1). In many actuarial applications we also want to study examples of the vector-valued parameter EF (2.2). We briefly discuss the categorical case since this case is frequently used.

5.7.1 Logistic Categorical Generalized Linear Model

We recall the EF representation of the categorical distribution studied in Sect. 2.1.4. We choose as ν the counting measure on the finite set $\mathcal{Y} = \{1, \ldots, k+1\}$. A random variable Y taking values in $\mathcal{Y}$ is called categorical, and the levels $y \in \mathcal{Y}$ can either be ordinal or nominal. This motivates dummy coding of the categorical random variable Y providing

$$T(Y) = (\mathbb{1}_{\{Y=1\}}, \ldots, \mathbb{1}_{\{Y=k\}})^\top \in \{0, 1\}^k, \tag{5.68}$$

thus, $k + 1$ has been chosen as reference level. For the canonical parameter $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)^\top \in \boldsymbol{\Theta} = \mathbb{R}^k$ we have cumulant function and mean functional, respectively,

$$\kappa(\boldsymbol{\theta}) = \log\left(1 + \sum_{j=1}^{k} e^{\theta_j}\right), \qquad \boldsymbol{p} = \mathbb{E}_{\boldsymbol{\theta}}[T(Y)] = \nabla_{\boldsymbol{\theta}}\kappa(\boldsymbol{\theta}) = \frac{e^{\boldsymbol{\theta}}}{1 + \sum_{j=1}^{k} e^{\theta_j}}.$$

With these choices we receive the EF representation of the categorical distribution (set $\theta_{k+1} = 0$)

$$dF(y; \boldsymbol{\theta}) = \exp\left\{\boldsymbol{\theta}^\top T(y) - \log\left(1 + \sum_{j=1}^{k} e^{\theta_j}\right)\right\} d\nu(y) = \prod_{l=1}^{k+1}\left(\frac{e^{\theta_l}}{\sum_{j=1}^{k+1} e^{\theta_j}}\right)^{\mathbb{1}_{\{y=l\}}} d\nu(y).$$

The covariance matrix of $T(Y)$ is given by

$$\Sigma(\boldsymbol{\theta}) = \mathrm{Var}_{\boldsymbol{\theta}}(T(Y)) = \nabla_{\boldsymbol{\theta}}^2\kappa(\boldsymbol{\theta}) = \mathrm{diag}\,(\boldsymbol{p}) - \boldsymbol{p}\boldsymbol{p}^\top \in \mathbb{R}^{k\times k}.$$

Assume that we have feature information $\boldsymbol{x} \in \mathcal{X} \subset \{1\} \times \mathbb{R}^q$ for response variable Y. This allows us to lift this categorical model to a GLM. The *logistic GLM* assumes for $\boldsymbol{p} = (p_1, \ldots, p_k)^\top \in (0,1)^k$ a regression function, $1 \le l \le k$,

$$\boldsymbol{x} \mapsto p_l = p_l(\boldsymbol{x}) = \mathbb{P}_{\boldsymbol{\beta}}[Y = l] = \frac{\exp\langle \boldsymbol{\beta}_l, \boldsymbol{x}\rangle}{1 + \sum_{j=1}^{k} \exp\langle \boldsymbol{\beta}_j, \boldsymbol{x}\rangle}, \tag{5.69}$$

for regression parameter $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^\top, \ldots, \boldsymbol{\beta}_k^\top)^\top \in \mathbb{R}^{k(q+1)}$. Equivalently, we can rewrite these regression probabilities relative to the reference level, that is, we consider linear predictors for $1 \le l \le k$

$$\eta_l(\boldsymbol{x}) = \log\left(\frac{\mathbb{P}_{\boldsymbol{\beta}}[Y = l]}{\mathbb{P}_{\boldsymbol{\beta}}[Y = k+1]}\right) = \langle \boldsymbol{\beta}_l, \boldsymbol{x}\rangle. \tag{5.70}$$

Note that this naturally gives us the canonical link h which we have already derived in Sect. 2.1.4. Define the matrix for feature $\boldsymbol{x} \in \mathcal{X} \subset \{1\} \times \mathbb{R}^q$

$$\boldsymbol{X} = \begin{pmatrix} \boldsymbol{x}^\top & 0 & 0 & \cdots & 0 \\ 0 & \boldsymbol{x}^\top & 0 & \cdots & 0 \\ 0 & 0 & \boldsymbol{x}^\top & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \boldsymbol{x}^\top \end{pmatrix} \in \mathbb{R}^{k \times k(q+1)}. \tag{5.71}$$

This gives linear predictor and canonical parameter, respectively, under the canonical link h

$$\boldsymbol{\theta} = h(\boldsymbol{p}(\boldsymbol{x})) = \boldsymbol{\eta}(\boldsymbol{x}) = \boldsymbol{X}\boldsymbol{\beta} = \big(\langle \boldsymbol{\beta}_1, \boldsymbol{x}\rangle, \ldots, \langle \boldsymbol{\beta}_k, \boldsymbol{x}\rangle\big)^\top \in \boldsymbol{\Theta} = \mathbb{R}^k. \tag{5.72}$$

5.7.2 Maximum Likelihood Estimation in Categorical Models

Assume we have n independent observations Y_i following the logistic categorical GLM (5.69) with features $\boldsymbol{x}_i \in \mathbb{R}^{q+1}$ and $\boldsymbol{X}_i \in \mathbb{R}^{k \times k(q+1)}$, respectively, for $1 \le i \le n$. The joint log-likelihood function is given by, we use (5.72),

$$\boldsymbol{\beta} \mapsto \ell_{\boldsymbol{Y}}(\boldsymbol{\beta}) = \sum_{i=1}^{n} (\boldsymbol{X}_i\boldsymbol{\beta})^\top T(Y_i) - \kappa(\boldsymbol{X}_i\boldsymbol{\beta}).$$

This provides us with score equations

$$s(\boldsymbol{\beta}, \boldsymbol{Y}) = \nabla_{\boldsymbol{\beta}} \ell_{\boldsymbol{Y}}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \boldsymbol{X}_i^\top \left[T(Y_i) - \nabla_{\boldsymbol{\theta}}\kappa(\boldsymbol{X}_i\boldsymbol{\beta})\right] = \sum_{i=1}^{n} \boldsymbol{X}_i^\top \left[T(Y_i) - \boldsymbol{p}(\boldsymbol{x}_i)\right] = 0,$$

with logistic regression function (5.69) for $\boldsymbol{p}(\boldsymbol{x})$. For the score equations with canonical link we also refer to the second case in Proposition 5.1. Next, we calculate Fisher's information matrix, we also refer to (3.16),

$$\mathcal{I}_n(\boldsymbol{\beta}) = -\mathbb{E}_{\boldsymbol{\beta}}\left[\nabla^2_{\boldsymbol{\beta}}\ell_{\boldsymbol{Y}}(\boldsymbol{\beta})\right] = \sum_{i=1}^{n} \boldsymbol{X}_i^\top \Sigma_i(\boldsymbol{\beta})\boldsymbol{X}_i,$$

with covariance matrix of $T(Y_i)$

$$\Sigma_i(\boldsymbol{\beta}) = \nabla^2_{\boldsymbol{\theta}}\kappa(\boldsymbol{X}_i\boldsymbol{\beta}) = \operatorname{diag}\left(\boldsymbol{p}(\boldsymbol{x}_i)\right) - \boldsymbol{p}(\boldsymbol{x}_i)\boldsymbol{p}(\boldsymbol{x}_i)^\top.$$

We rewrite the score in a similar way as in Sect. 5.1.4. This requires for general link $g(\boldsymbol{p}) = \boldsymbol{\eta}$ and inverse link $\boldsymbol{p} = g^{-1}(\boldsymbol{\eta})$, respectively, the following block diagonal matrix

$$\begin{aligned} W(\boldsymbol{\beta}) &= \operatorname{diag}\left(\left(\nabla_{\boldsymbol{\eta}} g^{-1}(\boldsymbol{\eta})\Big|_{\boldsymbol{\eta}=\boldsymbol{X}_i\boldsymbol{\beta}}\right)\Sigma_i(\boldsymbol{\beta})^{-1}\left(\nabla_{\boldsymbol{\eta}} g^{-1}(\boldsymbol{\eta})\Big|_{\boldsymbol{\eta}=\boldsymbol{X}_i\boldsymbol{\beta}}\right)^\top\right)_{1\le i\le n} \\ &= \operatorname{diag}\left(\left(\nabla_{\boldsymbol{p}} g(\boldsymbol{p})\big|_{\boldsymbol{p}=g^{-1}(\boldsymbol{X}_i\boldsymbol{\beta})}\right)^\top \Sigma_i(\boldsymbol{\beta})\left(\nabla_{\boldsymbol{p}} g(\boldsymbol{p})\big|_{\boldsymbol{p}=g^{-1}(\boldsymbol{X}_i\boldsymbol{\beta})}\right)\right)^{-1}_{1\le i\le n}, \qquad (5.73) \end{aligned}$$

and the working residuals

$$\boldsymbol{R}(\boldsymbol{Y},\boldsymbol{\beta}) = \left(\left(\nabla_{\boldsymbol{p}} g(\boldsymbol{p})\big|_{\boldsymbol{p}=g^{-1}(\boldsymbol{X}_i\boldsymbol{\beta})}\right)^\top \left(T(Y_i) - \boldsymbol{p}(\boldsymbol{x}_i)\right)\right)_{1\le i\le n}. \qquad (5.74)$$

Because we work with the canonical link $g = h$ and $g^{-1} = \nabla_{\boldsymbol{\theta}}\kappa$, we can use the simplified block diagonal matrix

$$W(\boldsymbol{\beta}) = \operatorname{diag}\left(\Sigma_1(\boldsymbol{\beta}), \ldots, \Sigma_n(\boldsymbol{\beta})\right) \in \mathbb{R}^{kn\times kn},$$

and the working residuals

$$\boldsymbol{R}(\boldsymbol{Y},\boldsymbol{\beta}) = \left(\Sigma_i(\boldsymbol{\beta})^{-1}\left(T(Y_i) - \boldsymbol{p}(\boldsymbol{x}_i)\right)\right)_{1\le i\le n} \in \mathbb{R}^{kn}.$$

Finally, we define the design matrix

$$\mathfrak{X} = \begin{pmatrix} \boldsymbol{X}_1 \\ \boldsymbol{X}_2 \\ \vdots \\ \boldsymbol{X}_n \end{pmatrix} \in \mathbb{R}^{kn\times k(q+1)}.$$

Putting everything together we receive the score equations

$$s(\boldsymbol{\beta}, \boldsymbol{Y}) = \nabla_{\boldsymbol{\beta}} \ell_{\boldsymbol{Y}}(\boldsymbol{\beta}) = \mathfrak{X}^\top W(\boldsymbol{\beta}) \boldsymbol{R}(\boldsymbol{Y}, \boldsymbol{\beta}) = 0. \tag{5.75}$$

This is now exactly in the same form as in Proposition 5.1. Fisher's scoring method/IRLS algorithm then allows us to recursively calculate the MLE of $\boldsymbol{\beta} \in \mathbb{R}^{k(q+1)}$ by

$$\widehat{\boldsymbol{\beta}}^{(t)} \mapsto \widehat{\boldsymbol{\beta}}^{(t+1)} = \left(\mathfrak{X}^\top W(\widehat{\boldsymbol{\beta}}^{(t)})\mathfrak{X}\right)^{-1} \mathfrak{X}^\top W(\widehat{\boldsymbol{\beta}}^{(t)}) \left(\mathfrak{X}\widehat{\boldsymbol{\beta}}^{(t)} + \boldsymbol{R}(\boldsymbol{Y}, \widehat{\boldsymbol{\beta}}^{(t)})\right).$$

We have asymptotic normality of the MLE (under suitable regularity conditions)

$$\widehat{\boldsymbol{\beta}}_n^{\mathrm{MLE}} \overset{(\mathrm{d})}{\approx} \mathcal{N}(\boldsymbol{\beta}, \mathcal{I}_n(\boldsymbol{\beta})^{-1}),$$

for large sample sizes n. This allows us to apply the Wald test (5.32) for backward parameter elimination. Moreover, in-sample and out-of-sample losses can be analyzed with unit deviances coming from the categorical cross-entropy loss function (4.19).

Remarks 5.32 The above derivations have been done for the categorical distribution under the canonical link choice. However, these considerations hold true for more general links g within the vector-valued parameter EF. That is, the block diagonal matrix $W(\boldsymbol{\beta})$ in (5.73) and the working residuals $\boldsymbol{R}(\boldsymbol{Y}, \boldsymbol{\beta})$ in (5.74) provide score equations (5.75) for general vector-valued parameter EF examples, and where we replace the categorical probability $\boldsymbol{p}$ by the mean $\mu = \mathbb{E}_{\boldsymbol{\beta}}[T(Y)]$.

5.8 Further Topics of Regression Modeling

There are several special topics and tools in regression modeling that we have not discussed, yet. Some of them will be considered in selected chapters below, and some points are mentioned here, without going into detail.

5.8.1 Longitudinal Data and Random Effects

The GLMs studied above have been considering cross-sectional data, meaning that we have fixed one time period t and studied this time period in an isolated fashion. Time-dependent extensions are called longitudinal or panel data. Consider a time series of data $(Y_{i,t}, \boldsymbol{x}_{i,t})$ for policies $1 \le i \le n$ and time points $t \ge 1$. For the prediction of response variable $Y_{i,t}$ we may then regress on the individual past history of policy i, given by the data

$$\mathcal{D}_{i,t} = \left\{Y_{i,1}, \ldots, Y_{i,t-1}, \boldsymbol{x}_{i,1}, \ldots, \boldsymbol{x}_{i,t}\right\}.$$

In particular, we may explore the distribution of $Y_{i,t}$, conditionally given $\mathcal{D}_{i,t}$,

$$Y_{i,t}|_{\mathcal{D}_{i,t}} \sim F(\cdot|\mathcal{D}_{i,t}; \theta),$$

for canonical parameter $\theta \in \boldsymbol{\Theta}$ and $F(\cdot|\mathcal{D}_{i,t}; \theta)$ being a member of the EDF. For a GLM we choose a link function g and make the assumption

$$g\left(\mathbb{E}_{\boldsymbol{\beta}}[Y_{i,t}|\mathcal{D}_{i,t}]\right) = \langle \boldsymbol{\beta}, \boldsymbol{z}_{i,t}\rangle, \tag{5.76}$$

where $\boldsymbol{z}_{i,t} \in \mathbb{R}^{q+1}$ is a $(q+1)$-dimensional and $\sigma(\mathcal{D}_{i,t})$-measurable feature vector, and regression parameter $\boldsymbol{\beta} \in \mathbb{R}^{q+1}$ describes the common systematic effects across all policies $1 \le i \le n$. This gives a generalized auto-regressive model, and if we have the Markov property

$$F(\cdot|\mathcal{D}_{i,t}; \theta) \overset{(\mathrm{d})}{=} F(\cdot|Y_{i,t-1}, \boldsymbol{x}_{i,t}; \theta) \qquad \text{for all } t \ge 2 \text{ and } \theta \in \boldsymbol{\Theta},$$

we obtain a generalized auto-regressive model of order 1. These longitudinal models allow one to model experience rating, for instance, in car insurance where the past claims history directly influences the future insurance prices, we refer to Remark 5.15 on bonus-malus systems (BMS).

The next level of complexity is obtained by extending regression structure (5.76) by policy i specific random effects $\boldsymbol{B}_i$ such that we may postulate

$$g\left(\mathbb{E}_{\boldsymbol{\beta}}[Y_{i,t}|\mathcal{D}_{i,t}, \boldsymbol{B}_i]\right) = \langle \boldsymbol{\beta}, \boldsymbol{z}_{i,t}\rangle + \langle \boldsymbol{B}_i, \boldsymbol{w}_{i,t}\rangle, \tag{5.77}$$

with $\sigma(\mathcal{D}_{i,t})$-measurable feature vector $\boldsymbol{w}_{i,t}$. Regression parameter $\boldsymbol{\beta}$ then describes the fixed systematic effects that are common over the entire portfolio $1 \le i \le n$ and $\boldsymbol{B}_i$ describes the policy dependent random effects (assumed to be normalized $\mathbb{E}[\boldsymbol{B}_i] = 0$). Typically one assumes that $\boldsymbol{B}_1, \ldots, \boldsymbol{B}_n$ are centered and i.i.d. Such effects are called static random effects because they are not time-dependent, and they may also be interpreted in a Bayesian sense.

Finally, extending these static random effects to dynamic random effects $\boldsymbol{B}_{i,t}$, $t \ge 1$, leads to so-called state-space models, the linear state-space model being the most popular example and being fitted using the Kalman filter [207].

5.8.2 Regression Models Beyond the GLM Framework

There are several ways in which the GLM framework can be modified.

Siblings of Generalized Linear Regression Functions

The most common modification of GLMs concerns the regression structure, namely, that the scalar product in the linear predictor

$$\boldsymbol{x} \;\mapsto\; g(\mu) = \eta = \langle \boldsymbol{\beta}, \boldsymbol{x} \rangle,$$

is replaced by another regression function. A popular alternative is the framework of generalized additive models (GAMs). GAMs go back to Hastie–Tibshirani [181, 182] and the standard reference is Wood [384]. GAMs consider the regression functions

$$\boldsymbol{x} \;\mapsto\; g(\mu) = \eta = \beta_0 + \sum_j \beta_j s_j(x_j), \tag{5.78}$$

where $s_j : \mathbb{R} \to \mathbb{R}$ are natural cubic splines. Natural cubic splines s_j are obtained by concatenating cubic functions in so-called nodes. A GAM can have as many nodes in each cubic spline s_j as there are different levels $x_{i,j}$ in the data $1 \le i \le n$. In general, this leads to very flexible regression models, and to control in-sample over-fitting regularization is applied, for regularization we also refer to Sect. 6.2. Regularization requires setting a tuning parameter, and an efficient determination of this tuning parameter uses generalized cross-validation, see Sect. 5.6. Nevertheless, fitting GAMs can be very computational, already for portfolios with 1 million policies and involving 20 feature components the calibration can be very slow. Moreover, regression function (5.78) does not (directly) allow for a data driven method of finding interactions between feature components. For these reasons, we do not further study GAMs in this monograph.

A modification in the regression function that is able to consider interactions between feature components is the framework of classification and regression trees (CARTs). CARTs have been introduced by Breiman et al. [54] in 1984, and they are still used in its original form today. Regression trees aim to partition the feature space $\mathcal{X}$ into a finite number of disjoint subsets $\mathcal{X}_t$, $1 \le t \le T$, such that all policies $(Y_i, \boldsymbol{x}_i)$ in the same subset $\boldsymbol{x}_i \in \mathcal{X}_t$ satisfy a certain homogeneity property w.r.t. the regression task (and the chosen loss function). The CART regression function is then defined by

$$\boldsymbol{x} \;\mapsto\; \mu(\boldsymbol{x}) = \sum_{t=1}^{T} \widehat{\mu}_t \; \mathbb{1}_{\{\boldsymbol{x} \in \mathcal{X}_t\}},$$

where $\widehat{\mu}_t$ is the homogeneous mean estimator on $\mathcal{X}_t$. These CARTs are popular building blocks for ensemble methods where different regression functions are combined, we mention random forests and boosting algorithms that mainly rely on CARTs. Random forests have been introduced by Breiman [52], and boosting has been popularized by Valiant [362], Kearns–Valiant [209, 210], Schapire [328], Freund [139] and Freund–Schapire [140]. Today boosting belongs to the most powerful predictive regression methods, we mention the XGBoost algorithm of Chen–Guestrin [71] that has won many competitions. We will not further study CARTs and boosting in these notes because these methods also have some drawbacks. For instance, resulting regression functions are not continuous nor do they easily allow to extrapolate data beyond the (observed) feature space, e.g., if we have a time component. Moreover, they are more difficult in the use of unstructured data such as text data. For more on CARTs and boosting in actuarial science we refer to Denuit et al. [100] and Ferrario–Hämmerli [125].

Other Distributional Models

The theory above has been relying on the EDF, but, of course, we could also study any other family of distribution functions. A clear drawback of the EDF is that it only considers light-tailed distribution functions, i.e., distribution functions for which the moment generating function exists around the origin. If the data is more heavy-tailed, one may need to transform this data and then use the EDF on the transformed data (with the drawback that one loses the balance property) or one chooses another family of distribution functions. Transformations have already been discussed in Remarks 2.11 and Sect. 5.3.9. Another two families of distributions that have been studied in the actuarial literature are the generalized beta of the second kind (GB2) distribution, see Venter [369], Frees et al. [137] and Chan et al. [66], and inhomogeneous phase type (IHP) distributions, see Albrecher et al. [8] and Bladt [37]. The GB2 family is a 4-parameter family, and it nests several examples such as the gamma, the Weibull, the Pareto and the Lomax distributions, see Table B1 in Chan et al. [66]. The density of the GB2 distribution is for $y > 0$ given by

$$\begin{aligned} f(y; a, b, \alpha_1, \alpha_2) &= \frac{\frac{|a|}{b}\left(\frac{y}{b}\right)^{a\alpha_1-1}}{B(\alpha_1,\alpha_2)\left(1+\left(\frac{y}{b}\right)^a\right)^{\alpha_1+\alpha_2}} \qquad (5.79)\\ &= \frac{\frac{|a|}{y}}{B(\alpha_1,\alpha_2)}\left(\frac{\left(\frac{y}{b}\right)^a}{1+\left(\frac{y}{b}\right)^a}\right)^{\alpha_1}\left(\frac{1}{1+\left(\frac{y}{b}\right)^a}\right)^{\alpha_2}, \end{aligned}$$

with scale parameter $b > 0$, shape parameters $a \in \mathbb{R}$ and $\alpha_1, \alpha_2 > 0$, and beta function

$$B(\alpha_1, \alpha_2) = \frac{\Gamma(\alpha_1)\Gamma(\alpha_2)}{\Gamma(\alpha_1+\alpha_2)}.$$

Consider a modified logistic transformation of variable $y \mapsto z = (y/b)^a/(1 + (y/b)^a) \in (0, 1)$. This gives us the beta density

$$f(z; \alpha_1, \alpha_2) = \frac{z^{\alpha_1-1}(1-z)^{\alpha_2-1}}{B(\alpha_1,\alpha_2)}.$$

Thus, the GB2 distribution can be obtained by a transformation of the beta distribution. The latter provides that a GB2 distributed random variable Y can be simulated from $Y \overset{(d)}{=} b(Z/(1-Z))^{1/a}$ with $Z \sim \text{Beta}(\alpha_1, \alpha_2)$.

A GB2 distributed random variable Y has first moment

$$\mathbb{E}_{a,b,\alpha_1,\alpha_2}[Y] \;=\; \frac{B(\alpha_1 + 1/a, \alpha_2 - 1/a)}{B(\alpha_1, \alpha_2)}\, b,$$

for $-\alpha_1 a < 1 < \alpha_2 a$. Observe that for $a > 0$ we have that the survival function of Y is regularly varying with tail index $\alpha_2 a > 0$. Thus, we can model Pareto-like tails with the GB2 family; for regular variation we refer to (1.3).

As proposed in Frees et al. [137], one can introduce a regression structure for $b > 0$ by choosing a log-link and setting

$$\log\left(\mathbb{E}_{a,b,\alpha_1,\alpha_2}[Y]\right) = \log\left(\frac{B(\alpha_1 + 1/a, \alpha_2 - 1/a)}{B(\alpha_1, \alpha_2)}\right) + \langle \boldsymbol{\beta}, \boldsymbol{x} \rangle.$$

MLE of $\boldsymbol{\beta}$ may pose some challenge because it depends on nuisance parameters a, α_1, α_2. In a recent paper Li et al. [251], there is a proposal to extend this GB2 regression to a composite regression model; composite models are discussed in Sect. 6.4.4, below. This closes this short section, and for more examples we refer to the literature.

5.8.3 Quantile Regression

Pinball Loss Function

The GLMs introduced above aim at estimating the means $\mu(\boldsymbol{x}) = \mathbb{E}_{\theta(\boldsymbol{x})}[Y]$ of random variables Y being explained by features $\boldsymbol{x}$. Since mean estimation can be rather sensitive in situations where we have large claims, the more robust quantile regression has attracted some attention, recently. Quantile regression has been introduced by Koenker–Bassett [220]. The idea is that instead of estimating the mean μ of a random variable Y, we rather try to estimate its τ-quantile for given $\tau \in (0, 1)$. The τ-quantile is given by the generalized inverse $F^{-1}(\tau)$ of the distribution function F of Y, that is,

$$F^{-1}(\tau) = \inf\{y \in \mathbb{R};\; F(y) \geq \tau\}. \tag{5.80}$$

Consider the *pinball loss function* for $y \in \mathfrak{C}$ (convex closure of the support of Y) and actions $a \in \mathbb{A} = \mathbb{R}$

$$(y, a) \;\mapsto\; L_\tau(y, a) = (y - a)\left(\tau - \mathbb{1}_{\{y-a<0\}}\right) \;\geq\; 0. \tag{5.81}$$

This provides us with the expected loss for $Y \sim F$ and action $a \in \mathbb{A}$

$$\begin{aligned}
\mathbb{E}_F\left[L_\tau(Y,a)\right] &= \mathbb{E}_F\left[(Y-a)\left(\tau - \mathbb{1}_{\{Y<a\}}\right)\right] \\
&= (\tau-1)\mathbb{E}_F\left[(Y-a)\mathbb{1}_{\{Y<a\}}\right] + \tau\mathbb{E}_F\left[(Y-a)\mathbb{1}_{\{Y\ge a\}}\right] \\
&= (\tau-1)\int_{-\infty}^{a}(y-a)dF(y) + \tau\int_{a}^{\infty}(y-a)dF(y).
\end{aligned}$$

The aim is to find an optimal action $\widehat{a}(F)$ that minimizes this expected loss, see (4.24),

$$\widehat{a}(F) \;\in\; \mathfrak{A}(F) \;=\; \underset{a\in\mathbb{A}}{\arg\min}\, \mathbb{E}_F\left[L_\tau(Y,a)\right].$$

Note that for the time being we do not know whether the solution to this minimization problem is a singleton. For this reason, we state the solution (subject to existence) as a set-valued functional $\mathfrak{A}$, see (4.25).

We calculate the score equation of the expected loss using the Leibniz rule

$$\begin{aligned}
\frac{\partial}{\partial a}\mathbb{E}_F\left[L_\tau(Y,a)\right] &= -(\tau-1)\int_{-\infty}^{a} dF(y) - \tau\int_{a}^{\infty} dF(y) \\
&= -(\tau-1)F(a) - \tau\left(1-F(a)\right) \;=\; F(a)-\tau \;\overset{!}{=}\; 0.
\end{aligned}$$

Assume the distribution F is continuous. This implies $F(F^{-1}(\tau)) = \tau$, and we have

$$F^{-1}(\tau) \;\in\; \mathfrak{A}(F) \;=\; \underset{a\in\mathbb{A}}{\arg\min}\, \mathbb{E}_F\left[L_\tau(Y,a)\right].$$

In fact, using the pinball loss, we have just seen that the τ-quantile is elicitable within the class of continuous distributions, see Definition 4.18.

For a more general result we need a more general definition of a (set-valued) τ-quantile

$$Q_\tau(F) = \left\{ y \in \mathbb{R};\; \lim_{z\uparrow y} F(z) \le \tau \le F(y) \right\}. \tag{5.82}$$

This defines a closed interval and its lower endpoint corresponds to the generalized inverse $F^{-1}(\tau)$ given in (5.80). In complete analogy to Theorem 4.19 on the elicitability of the mean functional, we have the following statement for the τ-quantile; this result goes back to Thomson [351] and Saerens [326].

Theorem 5.33 (Gneiting [162, Theorem 9], Without Proof) *Let $\mathcal{F}$ be the class of distribution functions on an interval $\mathfrak{C} \subseteq \mathbb{R}$ and choose quantile level $\tau \in (0,1)$.*

- *The τ-quantile* (5.82) *is elicitable relative to $\mathcal{F}$.*

- *Assume the loss function $L : \mathfrak{C} \times \mathbb{A} \to \mathbb{R}_+$ satisfies (L0)-(L2) on page 92 for interval $\mathfrak{C} = \mathbb{A} \subseteq \mathbb{R}$. L is consistent for the τ-quantile* (5.82) *relative to the class $\mathcal{F}$ of compactly supported distributions on $\mathfrak{C}$ if and only if L is of the form*

$$L(y, a) = (G(y) - G(a)) \left(\tau - \mathbb{1}_{\{y-a<0\}}\right),$$

  for a non-decreasing function G on $\mathfrak{C}$.
- *If G is strictly increasing on $\mathfrak{C}$ and if $\mathbb{E}_F[G(Y)]$ exists and is finite for all $F \in \mathcal{F}$, then the above loss function L is strictly consistent for the τ-quantile* (5.82) *relative to the class $\mathcal{F}$.*

Theorem 5.33 characterizes the strictly consistent loss functions for quantile estimation, the pinball loss being the special case $G(y) = y$.

Quantile Regression

The idea behind quantile regression is that we build a regression model for the τ-quantile. Assume we have a datum $(Y, \boldsymbol{x})$ whose conditional τ-quantile, given $\boldsymbol{x} \in \{1\} \times \mathbb{R}^q$, can be described by the regression function

$$\boldsymbol{x} \mapsto g\left(F^{-1}_{Y|\boldsymbol{x}}(\tau)\right) = \langle \boldsymbol{\beta}_\tau, \boldsymbol{x} \rangle,$$

for a strictly monotone and smooth link function $g : \mathfrak{C} \to \mathbb{R}$, and for a regression parameter $\boldsymbol{\beta}_\tau \in \mathbb{R}^{q+1}$. The aim now is to estimate this regression parameter from independent data $(Y_i, \boldsymbol{x}_i)$, $1 \leq i \leq n$. The pinball loss L_τ, given in (5.81), provides us with the following optimization problem

$$\widehat{\boldsymbol{\beta}}_\tau = \underset{\boldsymbol{\beta} \in \mathbb{R}^{q+1}}{\arg\min} \sum_{i=1}^{n} L_\tau \left(Y_i, g^{-1}\langle \boldsymbol{\beta}, \boldsymbol{x}_i \rangle\right).$$

This then allows us to estimate the corresponding τ-quantile as a function of the feature information $\boldsymbol{x}$. For $\tau = 1/2$ we estimate the median by

$$\widehat{F}^{-1}_{Y|\boldsymbol{x}}(1/2) = g^{-1}\left\langle \widehat{\boldsymbol{\beta}}_{1/2}, \boldsymbol{x} \right\rangle.$$

We conclude from this short section that we can regress any quantity $a(F)$ that is elicitable, i.e., for which a loss function exists that is strictly consistent for $a(F)$ on $F \in \mathcal{F}$. For more on quantile regression we refer to the monograph of Uribe–Guillén [361], and an interesting paper is Dimitriades et al. [106]. We will study quantile regression within deep networks in Chap. 11.2, below.