Generalized Linear Models 5

Linear models are well suited for regression analyses when the response variable is continuous and at least approximately normal. In some cases, an appropriate transformation is needed to ensure approximate normality of the response. In addition, the expectation of the response is assumed to be a linear combination of covariates. Again, these covariates may be transformed before being included in the linear predictor. However, in many applications the response is not a continuous variable, but rather binary, categorical, or a count variable as in the following examples:

- Patent opposition (yes/no), see Sect. 2.3 (p. 33).
- Creditworthiness of a client (yes/no).
- Benign or malignant tumor.
- Person is unemployed, part-time employed, or fully employed.
- Tree is very damaged, averagely or lightly damaged, or not damaged at all.
- Number of cases of illness, insurance claims, or problematic credits within a certain time frame.

Moreover, we are not always able to perform a satisfactory regression analysis for certain types of continuous response variables using a linear model. This is the case when dealing with a variable whose distribution is considerably skewed, as, for example, a life span, the amount of damages, or income. Even though data with a skewed distribution can sometimes be transformed into one with an approximately symmetric distribution, it is often advantageous to apply, for example, a gamma regression model to the original response variable.

Within a broad framework, generalized linear models (GLMs) unify many regression approaches with response variables that do not necessarily follow a normal distribution, including, for example, the logit model for binary response variables (Sect. 2.3) as well as the classical linear model with normally distributed errors. GLMs still rely on the assumption that the effect of covariates can be modeled through a linear predictor, similar as in logit and linear models. We start our description of GLMs with regression models for binary responses in Sect. 5.1. Next, Sect. 5.2 describes regression models for count data, especially Poisson regression. Section 5.3 is dedicated to models for nonnegative, continuous

L. Fahrmeir et al., *Regression*, DOI 10.1007/978-3-642-34333-9_5,

responses. Along with the introduction of suitable models, we discuss statistical inference relying on the likelihood principle. Section 5.4 offers a general unified discussion of GLMs and likelihood inference, while Sect. 5.5 outlines quasi-likelihood inference. Section 5.6 considers Bayesian GLMs. Finally, Sect. 5.7 transfers the boosting idea outlined for linear models in Sect. 4.3 to GLMs.

5.1 Binary Regression

5.1.1 Binary Regression Models

As in the previous chapters, we assume that (ungrouped) data on n objects or individuals are given in the form $(y_i, x_{i1}, \ldots, x_{ik})$, $i = 1, \ldots, n$, with the binary response y coded by 0 and 1 and covariates denoted by $x_1, \ldots, x_k$. Similar to linear and logit models in Example 2.8, $x_1, \ldots, x_k$ may have been derived from an appropriate transformation or coding of the original covariates. The main goal of a binary regression analysis is then to model and estimate the effects of the covariates on the (conditional) probability

$$\pi_i = \mathrm{P}(y_i = 1) = \mathrm{E}(y_i),$$

for the outcome $y_i = 1$ and given values of the covariates $x_{i1}, \ldots, x_{ik}$. In this specification, the response variables are assumed to be (conditionally) independent.

We already discussed the disadvantages of the *linear probability model*

$$\pi_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik}$$

for binary response variables in Sect. 2.3. In particular, the *linear predictor*

$$\eta_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} = \boldsymbol{x}_i' \boldsymbol{\beta},$$

with $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_k)'$ and $\boldsymbol{x}_i = (1, x_{i1}, \ldots, x_{ik})'$ must lie in the interval $[0, 1]$ for all vectors $\boldsymbol{x}$. This requires restrictions on the parameters $\boldsymbol{\beta}$ that are difficult to handle in the estimation process. Thus, all popular binary regression models combine the probability π_i with the linear predictor η_i through a relation of the form

$$\pi_i = h(\eta_i) = h(\beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik}), \qquad (5.1)$$

where h is a strictly monotonically increasing cumulative distribution function on the real line. This ensures $h(\eta) \in [0, 1]$ and Eq. (5.1) can always be expressed in the form

$$\eta_i = g(\pi_i),$$

with the inverse function $g = h^{-1}$. Within the framework of GLMs, h is called the *response function* and $g = h^{-1}$ is known as the *link function*. Logit and probit models are the most widely used binary regression models.

Logit Model

The logit model presented in Sect. 2.3 results from the choice of the logistic response function

$$\pi = h(\eta) = \frac{\exp(\eta)}{1 + \exp(\eta)} \tag{5.2}$$

or (equivalently) the logit link function

$$g(\pi) = \log\left(\frac{\pi}{1-\pi}\right) = \eta = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k. \tag{5.3}$$

This yields a linear model for the logarithmic odds (log-odds) $\log(\pi/(1-\pi))$. Transformation with the exponential function gives

$$\frac{\pi}{1-\pi} = \exp(\beta_0)\exp(\beta_1 x_1) \cdot \ldots \cdot \exp(\beta_k x_k), \tag{5.4}$$

implying that the effects of the covariates affect the odds $\pi/(1-\pi)$ in an exponential-multiplicative form; see Sect. 2.3 for this interpretation. Another interpretation—which is also available for the two models introduced in the following—results from the connection to latent linear models; see p. 274 for details.

Probit Model

For h, we use the standard normal cumulative distribution function Φ, i.e.,

$$\pi = \Phi(\eta) = \Phi(\boldsymbol{x}'\boldsymbol{\beta}). \tag{5.5}$$

A (minor) disadvantage is the required numerical evaluation of Φ in the maximum likelihood estimation of the parameter $\boldsymbol{\beta}$.

Complementary Log–Log Model

The complementary log–log model uses the extreme minimum-value cumulative distribution function

$$h(\eta) = 1 - \exp(-\exp(\eta)) \tag{5.6}$$

as response function, with the inverse

$$g(\pi) = \log(-\log(1-\pi))$$

as link function. In comparison to logit and probit models, this model is useful in more specific applications, for example, when modeling discrete duration times; see, e.g., Fahrmeir and Tutz (2001) for an introduction to discrete time duration models.

Figure 5.1 (left) shows the response functions of the three binary regression models, i.e., the logistic distribution function (5.2), the standard normal distribution function (5.5), and the extreme-value distribution function (5.6).

At first glance, the three models seem very different from each other: Even though the response function of logit and probit models are both symmetric around

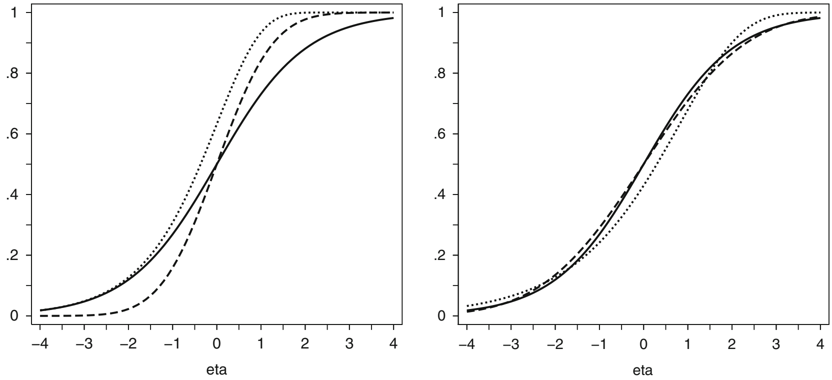

Fig. 5.1 Response functions (*left*) and adjusted response functions (*right*) in binary regression models: logit model (—), probit model (- - -), complementary log–log model (···)

zero, the logistic distribution function approaches 0 or 1 much slower for $\eta \to -\infty$ or $\eta \to +\infty$, respectively. In contrast, the response function of the complementary log–log model is asymmetric, following a similar pattern as the logit response function for small η, but showing a faster approach towards 1 as $\eta \to +\infty$. Thus, statistical analyses involving the three models might be expected to lead to very different results. However, for an adequate comparison of the models, we have to keep in mind that we could have used the more general cumulative distribution function h of a $\mathrm{N}(0, \sigma^2)$ distribution with any choice of variance $\sigma^2 \neq 1$, rather than the standard normal cumulative distribution function of the N(0, 1) distribution that defines the probit model. Standardizing h yields the relation

$$\pi(\eta) = h(\boldsymbol{x}'\boldsymbol{\beta}) = \Phi(\boldsymbol{x}'\boldsymbol{\beta}/\sigma) = \Phi(\boldsymbol{x}'\tilde{\boldsymbol{\beta}}),$$

where $\tilde{\boldsymbol{\beta}} = \boldsymbol{\beta}/\sigma$. Hence, even though the two response functions Φ (with $\sigma^2 = 1$) and h (with $\sigma^2 \neq 1$, e.g. $\sigma^2 = 4$) differ from each other, the resulting model for the probability $\pi(\eta)$ based on $h(\eta)$ with $\eta = \boldsymbol{x}'\boldsymbol{\beta}$ is equivalent to a probit model with the rescaled parameters $\tilde{\boldsymbol{\beta}} = \boldsymbol{\beta}/\sigma$. In this sense, the requirement of $\sigma^2 = 1$ in the probit model is arbitrary and we might just as well have assumed $\sigma^2 = 4$. We also obtain the same equivalence when deriving binary regression models from latent linear models; see p. 274.

For a fair comparison of logit and probit models, we need to put each on equal footing. Since the logistic distribution function has variance $\pi^2/3$ with the circular constant $\pi = 3.141593\ldots$, we need to compare it to a rescaled normal distribution function whose variance is adjusted to $\sigma^2 = \pi^2/3$. Figure 5.1 (right) shows the similarity of the logit and the adjusted probit response function.

Statistical analyses with logit and probit models therefore lead to similar estimated probabilities. The scaling $\tilde{\boldsymbol{\beta}} = \boldsymbol{\beta}/\sigma$ or $\boldsymbol{\beta} = \tilde{\boldsymbol{\beta}}\sigma$ will automatically be taken into account in the estimation process. Thus, the estimated coefficients

5.1 Binary Regression Models

Data

The binary response variables y_i are coded 0/1 and are (conditionally) independent given the covariates $x_{i1}, \ldots, x_{ik}$.

Models

The probability $\pi_i = \mathrm{P}(y_i = 1) = \mathrm{E}(y_i)$ and the *linear predictor*

$$\eta_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} = \boldsymbol{x}_i' \boldsymbol{\beta}$$

are connected by the *response function* $h(\eta) \in [0, 1]$ via

$$\pi_i = h(\eta_i).$$

Logit model

$$\pi = \frac{\exp(\eta)}{1 + \exp(\eta)} \quad \Longleftrightarrow \quad \log \frac{\pi}{1 - \pi} = \eta.$$

Probit model

$$\pi = \Phi(\eta) \quad \Longleftrightarrow \quad \Phi^{-1}(\pi) = \eta.$$

Complementary log–log model

$$\pi = 1 - \exp(-\exp(\eta)) \quad \Longleftrightarrow \quad \log(-\log(1 - \pi)) = \eta.$$

$\tilde{\beta}_1, \tilde{\beta}_2, \ldots$ of a logit model differ from the corresponding values $\beta_1, \beta_2, \ldots$ of a probit model (with $\sigma^2 = 1$) approximately by the factor $\sigma = \pi/\sqrt{3} \approx 1.814$, yet the estimated probabilities $\pi(\eta)$ are very similar. Since the ratios $\tilde{\beta}_1/\tilde{\beta}_2 = \beta_1/\beta_2$ etc. are independent of σ, therefore we should not interpret the absolute (estimated) coefficients, but rather the ratios β_1/β_2 etc., as illustrated in Example 5.1 (p. 275).

Similar considerations apply to the comparison with the complementary log–log model. Since the extreme-value distribution has variance $\sigma^2 = \pi^2/6$ and expectation -0.5772, the response function has to be adjusted to the variance $\sigma^2 = \pi^2/3$ and expectation 0 for a comparison with the logistic distribution function. This adjustment does have additional impact on the (estimated) intercept β_0. Figure 5.1 (right) shows the corresponding adjusted response function, which follows a similar form as those of the logit and probit function for small η, but also shows clear differences for larger η. Accordingly, the results of statistical analyses obtained with the complementary log–log model differ more substantially from those obtained by logit or probit models.

Binary Models and Latent Linear Models

Binary regression models can be derived by considering a *latent (unobserved) continuous response variable*, which is connected with the observed binary response via a threshold mechanism. Suppose we are investigating the decision of some individuals $i = 1, \dots, n$ when choosing between two alternatives $y = 0$ and $y = 1$. Typical examples include decision problems, e.g., related to buying a certain product or not. We further assume that individuals assign utilities u_{i0} and u_{i1} to each of the two alternatives. The alternative that maximizes the utility is chosen, i.e., $y_i = 1$ if $u_{i1} > u_{i0}$ and $y_i = 0$ if $u_{i1} \leq u_{i0}$.

Now suppose a researcher investigates the choice problem. However, one is not able to observe the latent utilities behind the decision, but rather observes the binary decisions y_i together with a number of explanatory variables $x_{i1}, \dots, x_{ik}$, which may influence the choice between the two alternatives. Assuming that the unobserved utilities can be additively decomposed and follow a linear model, we obtain

$$u_{i1} = \boldsymbol{x}_i'\tilde{\boldsymbol{\beta}}_1 + \tilde{\varepsilon}_{i1},$$

$$u_{i0} = \boldsymbol{x}_i'\tilde{\boldsymbol{\beta}}_0 + \tilde{\varepsilon}_{i0},$$

with $\boldsymbol{x}_i = (1, x_{i1}, \dots, x_{ik})'$. The unknown coefficient vectors $\tilde{\boldsymbol{\beta}}_1$ and $\tilde{\boldsymbol{\beta}}_0$ determine the effect of the explanatory variables on the utilities. The "errors" $\tilde{\varepsilon}_{i1}$ and $\tilde{\varepsilon}_{i0}$ include the effects of unobserved explanatory variables. Equivalently, we may choose to investigate utility differences, then obtaining

$$\tilde{y}_i = u_{i1} - u_{i0} = \boldsymbol{x}_i'(\tilde{\boldsymbol{\beta}}_1 - \tilde{\boldsymbol{\beta}}_0) + \tilde{\varepsilon}_{i1} - \tilde{\varepsilon}_{i0} = \boldsymbol{x}_i'\boldsymbol{\beta} + \varepsilon_i,$$

with $\boldsymbol{\beta} = \tilde{\boldsymbol{\beta}}_1 - \tilde{\boldsymbol{\beta}}_0$ and $\varepsilon_i = \tilde{\varepsilon}_{i1} - \tilde{\varepsilon}_{i0}$. The connection to the observable binary variables y_i is now given by $y_i = 1$ if $\tilde{y}_i = u_{i1} - u_{i0} > 0$ and $y_i = 0$ if $\tilde{y}_i = u_{i1} - u_{i0} \leq 0$.

Based on this framework, the binary responses y_i follow a Bernoulli distribution, i.e., $y_i \sim \mathrm{B}(1, \pi_i)$ with

$$\pi_i = \mathrm{P}(y_i = 1) = \mathrm{P}(\tilde{y}_i > 0) = \mathrm{P}(\boldsymbol{x}_i'\boldsymbol{\beta} + \varepsilon_i > 0) = \int I(\boldsymbol{x}_i'\boldsymbol{\beta} + \varepsilon_i > 0) f(\varepsilon_i)\, d\varepsilon_i,$$

where $I(\cdot)$ is the indicator function and f is the probability density of ε_i. We obtain different models depending on the choice of f. Specifically, when ε_i follows a logistic distribution, we obtain the logit model, while for standard normal errors $\varepsilon_i \sim \mathrm{N}(0, 1)$ we have the probit model $\pi_i = \Phi(\boldsymbol{x}_i'\boldsymbol{\beta})$. For $\varepsilon_i \sim \mathrm{N}(0, \sigma^2)$, we have

$$\pi_i = \Phi(\boldsymbol{x}_i'\boldsymbol{\beta}/\sigma) = \Phi(\boldsymbol{x}_i'\tilde{\boldsymbol{\beta}}),$$

through standardization with $\tilde{\boldsymbol{\beta}} = \boldsymbol{\beta}/\sigma$. This implies that regression coefficients $\boldsymbol{\beta}$ of a latent linear regression model can only be identified up to a factor $1/\sigma$. However, the ratio of two coefficients, for example, β_1 and β_2, is identifiable, since $\beta_1/\beta_2 = \tilde{\beta}_1/\tilde{\beta}_2$.

5.2 Interpretation of the Logit Model

Based on the linear predictor

$$\eta_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} = \boldsymbol{x}_i' \boldsymbol{\beta},$$

the odds

$$\frac{\pi_i}{1 - \pi_i} = \frac{\mathrm{P}(y_i = 1 \,|\, \boldsymbol{x}_i)}{\mathrm{P}(y_i = 0 \,|\, \boldsymbol{x}_i)}$$

follow the multiplicative model

$$\frac{\mathrm{P}(y_i = 1 \,|\, \boldsymbol{x}_i)}{\mathrm{P}(y_i = 0 \,|\, \boldsymbol{x}_i)} = \exp(\beta_0) \cdot \exp(x_{i1}\beta_1) \cdot \ldots \cdot \exp(x_{ik}\beta_k).$$

If, for example, x_{i1} increases by 1 unit to $x_{i1} + 1$, the following changes apply to the relationship of the odds:

$$\frac{\mathrm{P}(y_i = 1 \,|\, x_{i1}, \ldots)}{\mathrm{P}(y_i = 0 \,|\, x_{i1}, \ldots)} \Big/ \frac{\mathrm{P}(y_i = 1 \,|\, x_{i1} + 1, \ldots)}{\mathrm{P}(y_i = 0 \,|\, x_{i1} + 1, \ldots)} = \exp(\beta_1).$$

$$\begin{aligned}
\beta_1 > 0 &: \ \mathrm{P}(y_i = 1)/\mathrm{P}(y_i = 0) \text{ increases,}\\
\beta_1 < 0 &: \ \mathrm{P}(y_i = 1)/\mathrm{P}(y_i = 0) \text{ decreases,}\\
\beta_1 = 0 &: \ \mathrm{P}(y_i = 1)/\mathrm{P}(y_i = 0) \text{ remains unchanged.}
\end{aligned}$$

Interpretation of Parameters

One of the main reasons for the popularity of the logit model is its interpretation as a linear model for log-odds, as well as a multiplicative model for the odds $\pi/(1-\pi)$, as outlined in Sect. 2.3 and formulae (5.3) and (5.4). The latent linear model is useful to interpret effects in the probit model, since the covariate effects can be interpreted in the usual way with this model formulation (up to a common multiplicative factor). In general, interpretation best proceeds in two steps: For the linear predictor, we interpret the effects in the same way as in the linear model. Then we transform the linear effect for $\eta = \boldsymbol{x}'\boldsymbol{\beta}$ into a nonlinear effect for $\pi = h(\eta)$ with the response function h.

Example 5.1 Patent Opposition—Binary Regression

In Example 2.8 (p. 35), we analyzed the probability of patent opposition using a logit model with linear predictor

Table 5.1 Patent opposition: estimated regression coefficients for the logit, probit, and complementary log–log model. Adjusted coefficients for the probit and complementary log–log model are also included

Variable	Logit	Probit	Probit (adj.)	Log–Log	Log–Log (adj.)
intercept	201.740	119.204	216.212	164.519	211.744
year	−0.102	−0.060	−0.109	−0.083	−0.106
ncit	0.113	0.068	0.123	0.088	0.113
nclaim	0.026	0.016	0.029	0.021	0.027
ustwin	−0.406	−0.243	−0.441	−0.310	−0.398
patus	−0.526	−0.309	−0.560	−0.439	−0.563
patgsgr	0.196	0.121	0.219	0.154	0.198
ncountry	0.097	0.058	0.105	0.080	0.103

$$\begin{aligned}\eta_i &= \beta_0 + \beta_1 year_i + \beta_2 ncit_i + \beta_3 nclaims_i + \beta_4 ustwin_i \\ &\quad + \beta_5 patus_i + \beta_6 patgsgr_i + \beta_7 ncountry_i .\end{aligned}$$

For an interpretation of the estimated parameters in the logit model compare Example 2.8. For comparison, we now choose a probit model and a complementary log–log model using the same linear predictor and reanalyze the data. Table 5.1 contains parameter estimates for all three models. In order to compare the probit and logit fits, we have to multiply the estimated coefficients of the probit model with the factor $\pi/\sqrt{3} \approx 1.814$, following our previous considerations. For example, we obtain the estimated effect $-0.060 \cdot 1.814 \approx -0.109$ for the covariate *year* compared to -0.102 in the logit model. For the other coefficients, somewhat higher discrepancies occur at some places (see the fourth column in Table 5.1); however, the discrepancies are much smaller than the standard deviations of the estimates. Since, according to the interpretation of binary models, coefficients can only be interpreted up to a factor of $1/\sigma$, the probit and the logit models provide essentially the same results. After rescaling with the factor $\pi/\sqrt{6} \approx 1.283$, we also obtain comparable coefficients for the complementary log–log model, which are close to those of the logit model; see column 6 in Table 5.1.

△

Grouped Data

Thus far, we have assumed *individual data* or *ungrouped data*, which means that one observation $(y_i, \boldsymbol{x}_i)$ is given for each individual or object i in a sample of size n. Every binary, 0/1 coded value y_i of the response variable and every covariate vector $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ik})$ then belongs to exactly one unit $i = 1, \ldots, n$.

If some covariate vectors (i.e., rows of the design matrix) are identical, the data can be *grouped* as in Sect. 4.1.2 (p. 181). Specifically, after sorting and summarizing the data, the design matrix only contains rows with unique covariate vectors $\boldsymbol{x}_i$. In addition, the number n_i of replications of $\boldsymbol{x}_i$ in the original sample of the individual data and the relative frequencies $\bar{y}_i$ of the corresponding individual binary values of the response variables are given:

$$\begin{array}{c} \text{Group } 1 \\ \vdots \\ \text{Group } i \\ \vdots \\ \text{Group } G \end{array} \begin{bmatrix} n_1 \\ \vdots \\ n_i \\ \vdots \\ n_G \end{bmatrix} \begin{bmatrix} \bar{y}_1 \\ \vdots \\ \bar{y}_i \\ \vdots \\ \bar{y}_G \end{bmatrix} \begin{bmatrix} 1 & x_{11} & \cdots & x_{1k} \\ & \vdots & & \vdots \\ 1 & x_{i1} & & x_{ik} \\ & \vdots & & \vdots \\ 1 & x_{G1} & \cdots & x_{Gk} \end{bmatrix}$$

The number of unique covariate vectors in the sample G is often much smaller than the sample size n, especially when covariates are binary or categorial. Rather than *relative* frequencies $\bar{y}_i$, we can also provide the *absolute* frequencies $n_i \bar{y}_i$. Grouped data are then often presented in condensed form in a contingency table, as in the following Example 5.2.

The grouping of individual data decreases computing time, as well as memory requirements, and is also done to ensure data identification protection. Moreover, some inferential methods are only applicable for grouped data, especially when testing the goodness of fit for the model or for model diagnostics; see Sect. 5.1.4 (p. 287).

Individual data y_i are Bernoulli distributed with $\mathrm{P}(y_i = 1) = \pi_i$, i.e. $y_i \sim \mathrm{B}(1, \pi_i)$. If the response variables y_i are (conditionally) independent, the absolute frequencies $n_i \bar{y}_i$ of grouped data are binomially distributed, i.e.,

$$n_i \bar{y}_i \sim \mathrm{B}(n_i, \pi_i),$$

with $\mathrm{E}(n_i \bar{y}_i) = n_i \pi_i$, $\mathrm{Var}(n_i \bar{y}_i) = n_i \pi_i (1 - \pi_i)$. The relative frequencies then follow a “scaled” binomial distribution

$$\bar{y}_i \sim \mathrm{B}(n_i; \pi_i)/n_i,$$

i.e., the range of values of the probability function for relative frequencies is $\{0, 1/n_i, 2/n_i, \ldots, 1\}$, instead of $\{0, 1, 2, \ldots, n_i\}$. The probability function is

$$\mathrm{P}(\bar{y}_i = j/n_i) = \binom{n_i}{j} \pi_i^j (1 - \pi_i)^{n_i - j} \qquad j = 0, \ldots, n_i.$$

The mean and the variance are given by

$$\mathrm{E}(\bar{y}_i) = \pi_i, \quad \mathrm{Var}(\bar{y}_i) = \frac{\pi_i (1 - \pi_i)}{n_i}.$$

For modeling the probability π_i, we can use the same binary regression models as in case of individual data.

Table 5.2 Grouped infection data

	C-section			
	Planned		Not planned	
	Infection		Infection	
	Yes	No	Yes	No
Antibiotics				
Risk factor	1	17	11	87
No risk factor	0	2	0	0
No antibiotics				
Risk factor	28	30	23	3
No risk factor	8	32	0	9

Example 5.2 Caesarean Delivery—Grouped Data

Table 5.2 contains grouped data on infections of mothers after a C-section collected at the clinical center of the University of Munich. The response variable y "infection" is binary with

$$y = \begin{cases} 1 & \text{infection,} \\ 0 & \text{no infection.} \end{cases}$$

After each childbirth the following three binary covariates were collected:

$$\mathit{NPLAN} = \begin{cases} 1 & \text{C-section was not planned,} \\ 0 & \text{planned,} \end{cases}$$

$$\mathit{RISK} = \begin{cases} 1 & \text{risk factors existed,} \\ 0 & \text{no risk factors,} \end{cases}$$

$$\mathit{ANTIB} = \begin{cases} 1 & \text{antibiotics were administered as prophylaxis,} \\ 0 & \text{no antibiotics.} \end{cases}$$

After grouping the individual data of 251 mothers, the data can be represented in the form of a contingency table; see Table 5.2.

If we model the probability for an infection with a logit model

$$\log \frac{\text{P(Infection)}}{\text{P(No Infection)}} = \beta_0 + \beta_1\,\mathit{NPLAN} + \beta_2\,\mathit{RISK} + \beta_3\,\mathit{ANTIB},$$

we obtain the estimated coefficients

$$\hat{\beta}_0 = -1.89\,, \quad \hat{\beta}_1 = 1.07\,, \quad \hat{\beta}_2 = 2.03\,, \quad \hat{\beta}_3 = -3.25\,.$$

The multiplicative effect $\exp(\hat{\beta}_2) = 7.6$ implies that the odds of an infection is seven times higher when risk factors are present, for fixed levels of the other two factors. Such an interpretation of course requires that the chosen model without any interaction terms is adequate. We will return to this question in Example 5.3.

If we select a probit model with the same linear predictor, we obtain the estimated coefficients

$$\hat{\beta}_0 = -1.09\,, \quad \hat{\beta}_1 = 0.61\,, \quad \hat{\beta}_2 = 1.20\,, \quad \hat{\beta}_3 = -1.90\,.$$

Similar to Example 5.1, the absolute values seem to be very different. However, the relative effects, e.g., the ratios $\hat{\beta}_1/\hat{\beta}_2$, are again very similar.

△

Overdispersion

For grouped data, we can estimate the variance within a group via $\bar{y}_i(1-\bar{y}_i)/n_i$, since $\bar{y}_i$ is the ML estimator for π_i based on the data in group i, disregarding the covariate information. In applications, this *empirical* variance is often much larger than the variance $\hat{\pi}_i(1-\hat{\pi}_i)/n_i$ predicted by a binomial regression model with $\hat{\pi}_i = h(\boldsymbol{x}_i'\hat{\boldsymbol{\beta}})$. This phenomenon is called overdispersion, since the data show a higher variability than is presumed by the model. The two main reasons for overdispersion are *unobserved heterogeneity*, which remains unexplained by the observed covariates, and *positive correlations* between the individual binary observations of the response variables, for example, when individual units belong to one *cluster* such as the same household. In either case, the individual binary response variables within a group are then (in most cases positively) correlated. The sum of binary responses is then no longer binomially distributed and has a larger variance according to the variance formula for correlated variables; see in Appendix B.2 Theorem B.2.4. This situation occurs in Sect. 5.2 for Poisson distributed response variables, where a data example of overdispersion is presented.

The easiest way to address the increased variability is through the introduction of a multiplicative overdispersion parameter $\phi > 1$ into the variance formula, i.e., we assume

$$\text{Var}(y_i) = \phi \frac{\pi_i(1-\pi_i)}{n_i}.$$

Estimation of the overdispersion parameter is described in Sect. 5.1.5.

5.1.2 Maximum Likelihood Estimation

The primary goal of statistical inference is the estimation of parameters $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_k)'$ and hypothesis testing for these effects, similar to linear models in Chap. 3. The methodology of this section is based on the likelihood principle: For given data $(y_i, \boldsymbol{x}_i)$, estimation of the parameters relies on the maximization of the likelihood function. Hypotheses regarding the parameters are tested using either likelihood ratio, Wald, or score tests; see Sect. 5.1.3. Appendix B.4.4 provides a general introduction into likelihood-based hypothesis testing.

Due to the (conditional) independence of the response variables, the likelihood $L(\boldsymbol{\beta})$ is given as the product

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} f(y_i \mid \boldsymbol{\beta}) \tag{5.7}$$

of the densities of y_i, which depend on the unknown parameter $\boldsymbol{\beta}$ through $\pi_i = \text{E}(y_i) = h(\boldsymbol{x}_i'\boldsymbol{\beta})$. Maximization of $L(\boldsymbol{\beta})$ or the log-likelihood $l(\boldsymbol{\beta}) = \log(L(\boldsymbol{\beta}))$

then yields the ML estimator $\hat{\boldsymbol{\beta}}$. It turns out that the ML estimator has no closed form as for linear models. Instead we rely on iterative methods, in particular Fisher scoring as briefly described in Appendix B.4.2. In order to compute the ML estimator numerically we require the score function $\boldsymbol{s}(\boldsymbol{\beta})$ and the observed or expected Fisher matrix $\boldsymbol{H}(\boldsymbol{\beta})$ or $\boldsymbol{F}(\boldsymbol{\beta})$.

We consider the case of individual data and describe the necessary steps for deriving ML estimates in the binary logit model:

1. Likelihood

For binary response variables $y_i \sim \mathrm{B}(1, \pi_i)$ with $\pi_i = \mathrm{P}(y_i = 1) = \mathrm{E}(y_i) = \mu_i$, the (discrete) density is given by

$$f(y_i \mid \pi_i) = \pi_i^{y_i}(1 - \pi_i)^{1-y_i}.$$

Since $\pi_i = h(\boldsymbol{x}_i'\boldsymbol{\beta})$, the density depends on $\boldsymbol{\beta}$ for given $\boldsymbol{x}_i$, and we can therefore also denote it as $f(y_i \mid \boldsymbol{\beta})$. The density also defines the likelihood contribution $L_i(\boldsymbol{\beta})$ of the ith observation. Due to the (conditional) independence of the responses y_i, the likelihood $L(\boldsymbol{\beta})$ is given by

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} L_i(\boldsymbol{\beta}) = \prod_{i=1}^{n} \pi_i^{y_i}(1 - \pi_i)^{1-y_i},$$

i.e., the product of the individual likelihood contributions $L_i(\boldsymbol{\beta})$.

2. Log-likelihood

The log-likelihood results from taking the logarithm of the likelihood yielding

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{n} l_i(\boldsymbol{\beta}) = \sum_{i=1}^{n} \{y_i \log(\pi_i) - y_i \log(1 - \pi_i) + \log(1 - \pi_i)\},$$

with the *log-likelihood contributions*

$$\begin{aligned} l_i(\boldsymbol{\beta}) = \log L_i(\boldsymbol{\beta}) &= y_i \log(\pi_i) - y_i \log(1 - \pi_i) + \log(1 - \pi_i) \\ &= y_i \log\left(\frac{\pi_i}{1 - \pi_i}\right) + \log(1 - \pi_i). \end{aligned}$$

For the logit model, we have

$$\pi_i = \frac{\exp(\boldsymbol{x}_i'\boldsymbol{\beta})}{1 + \exp(\boldsymbol{x}_i'\boldsymbol{\beta})} \quad \text{or} \quad \log\left(\frac{\pi_i}{1 - \pi_i}\right) = \boldsymbol{x}_i'\boldsymbol{\beta} = \eta_i$$

and $(1 - \pi_i) = (1 + \exp(\boldsymbol{x}_i'\boldsymbol{\beta}))^{-1}$. Therefore we obtain

$$l_i(\boldsymbol{\beta}) = y_i(\boldsymbol{x}_i'\boldsymbol{\beta}) - \log(1 + \exp(\boldsymbol{x}_i'\boldsymbol{\beta})) = y_i\eta_i - \log(1 + \exp(\eta_i)).$$

3. Score function

To calculate the ML estimator, defined as the maximizer of the log-likelihood $l(\boldsymbol{\beta})$, we require the score function, i.e., the first derivative of $l(\boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}$:

$$\boldsymbol{s}(\boldsymbol{\beta}) = \frac{\partial l(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} \frac{\partial l_i(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} \boldsymbol{s}_i(\boldsymbol{\beta}).$$

The individual contributions are given by $s_i(\boldsymbol{\beta}) = \partial l_i(\boldsymbol{\beta})/\partial \boldsymbol{\beta}$, or more specifically for logistic regression, using the chain rule,

$$\frac{\partial l_i(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \frac{\partial l_i}{\partial \eta_i}\frac{\partial \eta_i}{\partial \boldsymbol{\beta}} = \left[y_i - \frac{1}{1+\exp(\eta_i)}\exp(\eta_i) \right] \boldsymbol{x}_i,$$

with p-dimensional vector $\partial \eta_i / \partial \boldsymbol{\beta} = \boldsymbol{x}_i$. Further substitution of $\pi_i = \exp(\boldsymbol{x}_i'\boldsymbol{\beta})/(1+\exp(\boldsymbol{x}_i'\boldsymbol{\beta}))$ provides

$$\boldsymbol{s}_i(\boldsymbol{\beta}) = \boldsymbol{x}_i(y_i - \pi_i)$$

and the score function

$$\boldsymbol{s}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \boldsymbol{x}_i(y_i - \pi_i). \tag{5.8}$$

Here, $s(\boldsymbol{\beta})$ depends on $\pi_i = \pi_i(\boldsymbol{\beta}) = h(\boldsymbol{x}_i'\boldsymbol{\beta}) = \exp(\boldsymbol{x}_i'\boldsymbol{\beta})/(1+\exp(\boldsymbol{x}_i'\boldsymbol{\beta}))$ and is therefore nonlinear in $\boldsymbol{\beta}$. From $\mathrm{E}(y_i) = \pi_i$ it follows

$$\mathrm{E}(\boldsymbol{s}(\boldsymbol{\beta})) = \boldsymbol{0}.$$

Equating the score function to zero leads to the *ML equations*

$$\boldsymbol{s}(\hat{\boldsymbol{\beta}}) = \sum_{i=1}^{n} \boldsymbol{x}_i \left(y_i - \frac{\exp(\boldsymbol{x}_i'\hat{\boldsymbol{\beta}})}{1+\exp(\boldsymbol{x}_i'\hat{\boldsymbol{\beta}})} \right) = \boldsymbol{0}. \tag{5.9}$$

This p-dimensional, nonlinear system of equations for $\hat{\boldsymbol{\beta}}$ is usually solved iteratively by the Newton–Raphson or Fisher scoring algorithm; see p. 283.

4. Information matrix

For the estimation of the regression coefficients and the covariance matrix of the ML estimator $\hat{\boldsymbol{\beta}}$, we need the *observed information matrix*

$$\boldsymbol{H}(\boldsymbol{\beta}) = -\frac{\partial^2 l(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}'},$$

with the second derivatives $\partial^2 l(\boldsymbol{\beta})/\partial \beta_j \partial \beta_r$ as elements of the matrix $\partial^2 l(\boldsymbol{\beta})/\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}'$, or the *Fisher matrix (expected information matrix)*

$$\boldsymbol{F}(\boldsymbol{\beta}) = \mathrm{E}\left(-\frac{\partial^2 l(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}'}\right) = \mathrm{Cov}(\boldsymbol{s}(\boldsymbol{\beta})) = \mathrm{E}(\boldsymbol{s}(\boldsymbol{\beta})\boldsymbol{s}'(\boldsymbol{\beta})).$$

The last equality holds since $\mathrm{E}(\boldsymbol{s}(\boldsymbol{\beta})) = \mathbf{0}$. To derive the Fisher matrix note that $F(\boldsymbol{\beta})$ is additive, i.e., $\boldsymbol{F}(\boldsymbol{\beta}) = \sum_{i=1}^n \boldsymbol{F}_i(\boldsymbol{\beta})$, where $\boldsymbol{F}_i(\boldsymbol{\beta}) = \mathrm{E}(\boldsymbol{s}_i(\boldsymbol{\beta})\boldsymbol{s}_i(\boldsymbol{\beta})')$ is the contribution of the ith observation. For $\boldsymbol{F}_i(\boldsymbol{\beta})$ we obtain

$$\begin{aligned}\boldsymbol{F}_i(\boldsymbol{\beta}) &= \mathrm{E}\left(\boldsymbol{s}_i(\boldsymbol{\beta})\boldsymbol{s}_i(\boldsymbol{\beta})'\right)\\ &= \mathrm{E}\left(\boldsymbol{x}_i\boldsymbol{x}_i'(y_i - \pi_i)^2\right)\\ &= \boldsymbol{x}_i\boldsymbol{x}_i'\mathrm{E}\,(y_i - \pi_i)^2\\ &= \boldsymbol{x}_i\boldsymbol{x}_i'\mathrm{Var}(y_i)\\ &= \boldsymbol{x}_i\boldsymbol{x}_i'\pi_i(1-\pi_i).\end{aligned}$$

We finally get

$$\boldsymbol{F}(\boldsymbol{\beta}) = \sum_{i=1}^n \boldsymbol{F}_i(\boldsymbol{\beta}) = \sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{x}_i'\pi_i(1-\pi_i).$$

Since $\pi_i = h(\boldsymbol{x}_i'\boldsymbol{\beta})$, the Fisher matrix also depends on $\boldsymbol{\beta}$.

To derive the observed information matrix we use Definition A.29 of Appendix A.8. We obtain $\boldsymbol{H}(\boldsymbol{\beta}) = -\partial^2 l(\boldsymbol{\beta})/\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}' = -\partial\boldsymbol{s}(\boldsymbol{\beta})/\partial\boldsymbol{\beta}'$ through another differentiation of

$$-\boldsymbol{s}(\boldsymbol{\beta}) = \sum_{i=1}^n \boldsymbol{x}_i(\pi_i(\boldsymbol{\beta}) - y_i).$$

Using the chain rule, this yields

$$\boldsymbol{H}(\boldsymbol{\beta}) = -\frac{\partial\boldsymbol{s}(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}'} = \sum_{i=1}^n \boldsymbol{x}_i\frac{\partial\pi_i(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}'} = \sum_{i=1}^n \boldsymbol{x}_i\frac{\partial\eta_i}{\partial\boldsymbol{\beta}'}\frac{\partial\pi_i(\boldsymbol{\beta})}{\partial\eta_i} = \sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{x}_i'\pi_i(\boldsymbol{\beta})(1-\pi_i(\boldsymbol{\beta})).$$

We thereby used

$$\frac{\partial\eta_i}{\partial\boldsymbol{\beta}'} = \left(\frac{\partial\eta_i}{\partial\boldsymbol{\beta}}\right)' = \boldsymbol{x}_i'$$

and

$$\frac{\partial\pi_i(\boldsymbol{\beta})}{\partial\eta_i} = \frac{(1+\exp(\eta_i))\exp(\eta_i) - \exp(\eta_i)\exp(\eta_i)}{(1+\exp(\eta_i))^2} = \pi_i(\boldsymbol{\beta})(1-\pi_i(\boldsymbol{\beta})).$$

The expected and the observed information matrix are, thus, identical for the logit model, i.e., $\boldsymbol{H}(\boldsymbol{\beta}) = \boldsymbol{F}(\boldsymbol{\beta})$. This relationship, however, does not hold for other models, e.g., the *probit* or the *complementary log–log model*. In these models, we usually use the Fisher matrix $\boldsymbol{F}(\boldsymbol{\beta})$, which is typically easier to compute than the observed Fisher matrix $\boldsymbol{H}(\boldsymbol{\beta})$. Its general form will be given in Sect. 5.4.2.

If now instead of individual data with binary response variables $y_i \sim \mathrm{B}(1, \pi_i)$, we rather consider a binomially distributed response $y_i \sim \mathrm{B}(n_i, \pi_i)$ or relative frequencies

$$\bar{y}_i \sim \mathrm{B}(n_i, \pi_i)/n_i, \qquad i = 1, \ldots, n,$$

as, for example, in the case of grouped data, the formulae for $l(\boldsymbol{\beta}), \boldsymbol{s}(\boldsymbol{\beta})$, and $\boldsymbol{F}(\boldsymbol{\beta})$ have to be modified appropriately. Analogous arguments than for individual data yield

$$\begin{aligned} l(\boldsymbol{\beta}) &= \sum_{i=1}^{G} \{y_i \log(\pi_i) - y_i \log(1-\pi_i) + n_i \log(1-\pi_i)\} \\ \boldsymbol{s}(\boldsymbol{\beta}) &= \sum_{i=1}^{G} \boldsymbol{x}_i (y_i - n_i \pi_i) = \sum_{i=1}^{G} n_i \boldsymbol{x}_i (\bar{y}_i - \pi_i) \\ \boldsymbol{F}(\boldsymbol{\beta}) &= \sum_{i=1}^{G} \boldsymbol{x}_i \boldsymbol{x}_i' n_i \pi_i (1-\pi_i). \end{aligned}$$

Iterative Calculation of the ML Estimator

Several iterative algorithms that compute the ML estimator as the solution of the ML equation $\boldsymbol{s}(\hat{\boldsymbol{\beta}}) = \boldsymbol{0}$ can be used for computing $\hat{\boldsymbol{\beta}}$. The most common method is the Fisher scoring algorithm; see Sect. B.4.2 in Appendix B. Given starting values $\hat{\boldsymbol{\beta}}^{(0)}$, e.g., the least squares estimate, the algorithm iteratively performs updates

$$\hat{\boldsymbol{\beta}}^{(t+1)} = \hat{\boldsymbol{\beta}}^{(t)} + \boldsymbol{F}^{-1}(\hat{\boldsymbol{\beta}}^{(t)}) \boldsymbol{s}(\hat{\boldsymbol{\beta}}^{(t)}), \quad t = 0, 1, 2, \ldots. \tag{5.10}$$

Once a convergence criterion is met, for example, $||\hat{\boldsymbol{\beta}}^{(t+1)} - \hat{\boldsymbol{\beta}}^{(t)}||/||\hat{\boldsymbol{\beta}}^{(t)}|| \le \varepsilon$ (with $||\cdot||$ denoting the L_2-norm of a vector), the iterations will be stopped, and $\hat{\boldsymbol{\beta}} \equiv \hat{\boldsymbol{\beta}}^{(t)}$ is the ML estimator. Since $\boldsymbol{F}(\boldsymbol{\beta}) = \boldsymbol{H}(\boldsymbol{\beta})$ in the logit model, the Fisher scoring algorithm corresponds to a Newton method. The iterations Eq. (5.10) can also be expressed in the form of an iteratively weighted least squares estimation; see Sect. 5.4.2 (p. 306).

The Fisher scoring iterations can only converge to the ML solution $\hat{\boldsymbol{\beta}}$ if the Fisher matrix $\boldsymbol{F}(\boldsymbol{\beta})$ is invertible for all $\boldsymbol{\beta}$. As in the linear regression model, this requires that the design matrix $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)'$ has full rank p. Then $\boldsymbol{F}(\boldsymbol{\beta})$ is invertible for the types of regression models that we have considered thus far. For example, in case of the logit model, $\boldsymbol{F}(\boldsymbol{\beta}) = \sum_i \boldsymbol{x}_i \boldsymbol{x}_i' \pi_i (1-\pi_i)$ has full rank because $\boldsymbol{X}'\boldsymbol{X} = \sum_i \boldsymbol{x}_i \boldsymbol{x}_i'$ has full rank p and $\pi_i(1-\pi_i) > 0$ for all $\boldsymbol{\beta} \in \mathbb{R}^p$. Hence, as in the linear regression model, we will always assume that

$$\mathrm{rk}(\boldsymbol{X}) = p.$$

Typically, the algorithm then converges and stops close to the maximum after only a few iterations.

Nevertheless, it is possible that iterations diverge, i.e., that the successive differences $\|\hat{\boldsymbol{\beta}}^{(t+1)} - \hat{\boldsymbol{\beta}}^{(t)}\|$ increase instead of converging towards zero. This is the case when the likelihood does not have a maximum for finite $\boldsymbol{\beta}$, i.e., if at least one component in $\hat{\boldsymbol{\beta}}^{(t)}$ diverges to $\pm\infty$, and no finite ML estimator exists. In general, the non-existence of the ML estimator is observed in very unfavorable data configurations, especially when the sample size n is small in comparison to the dimension p.

Even though several authors have elaborated on conditions of the uniqueness and existence of ML estimators, these conditions are, to some extent, very complex. For practical purposes it is, thus, easier to check the convergence or divergence of the iterative method empirically.

Example 5.3 Caesarian Delivery—Binary Regression

In Example 5.2, we chose a main effects model

$$\eta = \beta_0 + \beta_1 \mathit{NPLAN} + \beta_2 \mathit{RISK} + \beta_3 \mathit{ANTIB}$$

for the linear predictor, i.e., a model without interactions between the covariates. If we want to improve the model fit by introducing interaction terms, we observe the following:

If we only include the interaction *NPLAN* · *ANTIB*, the corresponding estimated coefficient is close to zero. If we include the interactions *RISK* · *ANTIB* or *NPLAN* · *RISK*, we observe the problem of a nonexistent maximum, i.e., the ML estimator diverges. The reason is that we exclusively observed “no infection” for the response variable for both $\mathit{NPLAN} = 0$, $\mathit{RISK} = 0$, $\mathit{ANTIB} = 1$ and $\mathit{NPLAN} = 1$, $\mathit{RISK} = 0$, $\mathit{ANTIB} = 0$. This leads to the divergence towards infinity for the estimated effects of *ANTIB* and *RISK* · *ANTIB* or *NPLAN* and *NPLAN*·*RISK*, and a termination before convergence yields exceptionally high estimated interaction effects and standard errors. Depending on the chosen software, the user may receive a warning or not. In any case, very high estimated regression coefficients and/or standard errors may be a sign for non-convergence of the ML estimator.

It is clear that the problem is dependent on the specific data configuration: If we were to move one observation from the two empty cells over to the “infection” category, then the interactions converge and finite ML estimators exist.

△

Comparison of the ML and Least Squares Estimator

In a linear regression model with normally distributed error terms, we have

$$y_i \sim \mathrm{N}(\mu_i = \boldsymbol{x}_i'\boldsymbol{\beta}, \sigma^2).$$

Apart from constant factors, the score function is then given by

$$\boldsymbol{s}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \boldsymbol{x}_i (y_i - \mu_i),$$

where $\mathrm{E}(y_i) = \mu_i = \boldsymbol{x}_i'\boldsymbol{\beta}$ linearly depends on $\boldsymbol{\beta}$. For the logit model, the score function (5.8) follows the same structure, with $\mathrm{E}(y_i) = \pi_i = \mu_i$. However, $\boldsymbol{s}(\boldsymbol{\beta})$ is nonlinear in $\boldsymbol{\beta}$ since $\pi_i = \mu_i = \exp(\boldsymbol{x}_i'\boldsymbol{\beta})/\{1+\exp(\boldsymbol{x}_i'\boldsymbol{\beta})\}$. The ML or least squares system of equations for the linear model has the form

$$\boldsymbol{s}(\hat{\boldsymbol{\beta}}) = \sum_{i=1}^{n} \boldsymbol{x}_i (y_i - \boldsymbol{x}_i'\hat{\boldsymbol{\beta}}) = \boldsymbol{X}'\boldsymbol{y} - \boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{0},$$

with responses $\boldsymbol{y} = (y_1, \dots, y_n)'$. If the design matrix $\boldsymbol{X}$ has full rank p, we obtain the estimated regression coefficients as the solution of the system of equations $\boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}'\boldsymbol{y}$ in a single step, yielding

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}.$$

In contrast, the solution to the nonlinear system of equations (5.9) has to be obtained numerically in several iterative steps in the logit model. The (observed and expected) information matrix in the linear model is

$$\boldsymbol{F}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \boldsymbol{x}_i'\boldsymbol{x}_i/\sigma^2 = \frac{1}{\sigma^2}\boldsymbol{X}'\boldsymbol{X}.$$

The structure is again very similar to the one in the logit model, but the information matrix does not depend on $\boldsymbol{\beta}$.

Asymptotic Properties of the ML Estimator

Under relatively weak regularity conditions, one can shows that asymptotically (i.e., for $n \to \infty$), the ML estimator exists, is consistent, and follows a normal distribution. This result does not require that the sample size goes to infinity for each distinct location in the covariate space, but it is sufficient that the total sample size goes to infinity, i.e., $n \to \infty$. Then, for a sufficiently large sample size n, $\hat{\boldsymbol{\beta}}$ has an approximate normal distribution

$$\hat{\boldsymbol{\beta}} \overset{a}{\sim} \mathrm{N}(\boldsymbol{\beta}, \boldsymbol{F}^{-1}(\hat{\boldsymbol{\beta}})),$$

with estimated covariance matrix

$$\widehat{\mathrm{Cov}}(\hat{\boldsymbol{\beta}}) = \boldsymbol{F}^{-1}(\hat{\boldsymbol{\beta}})$$

equal to the inverse Fisher matrix evaluated at the ML estimator $\hat{\boldsymbol{\beta}}$. The diagonal element a_{jj} of the inverse Fisher matrix $\boldsymbol{A} = \boldsymbol{F}^{-1}(\hat{\boldsymbol{\beta}})$ is then an estimator of the variance of the jth component $\hat{\beta}_j$ of $\hat{\boldsymbol{\beta}}$, i.e.,

$$\widehat{\mathrm{Var}}(\hat{\beta}_j) = a_{jj},$$

and $\mathrm{se}_j = \sqrt{a_{jj}}$ is the standard error of $\hat{\beta}_j$ or in other words an estimator for the standard deviation $\sqrt{\mathrm{Var}(\hat{\beta}_j)}$. More details regarding the asymptotic properties of the ML estimator can be found in Fahrmeir and Kaufmann (1985).

5.1.3 Testing Linear Hypotheses

Linear hypotheses have the same form as in linear models:

$$H_0 : \boldsymbol{C}\boldsymbol{\beta} = \boldsymbol{d} \quad \text{versus} \quad H_1 : \boldsymbol{C}\boldsymbol{\beta} \neq \boldsymbol{d},$$

with $\boldsymbol{C}$ having full row rank $r \leq p$. We can use the likelihood ratio, the score and the Wald statistics for testing; see Appendix B.4.4. The *likelihood ratio statistic*

$$lr = -2\{l(\tilde{\boldsymbol{\beta}}) - l(\hat{\boldsymbol{\beta}})\}$$

measures the deviation in log-likelihood between the unrestricted maximum $l(\hat{\boldsymbol{\beta}})$ and that of the restricted maximum $l(\tilde{\boldsymbol{\beta}})$ under H_0, where $\tilde{\boldsymbol{\beta}}$ is the ML estimator under the restriction $\boldsymbol{C}\boldsymbol{\beta} = \boldsymbol{d}$. For the special case

$$H_0 : \boldsymbol{\beta}_1 = \boldsymbol{0} \quad \text{versus} \quad H_1 : \boldsymbol{\beta}_1 \neq \boldsymbol{0}, \tag{5.11}$$

where $\boldsymbol{\beta}_1$ is a subset of $\boldsymbol{\beta}$, we test the significance of the effects belonging to $\boldsymbol{\beta}_1$. The computation of $\tilde{\boldsymbol{\beta}}$ then simply requires ML estimation of the corresponding submodel. The numerical complexity is much greater for general linear hypotheses, since maximization has to be carried out under the constraint $\boldsymbol{C}\boldsymbol{\beta} = \boldsymbol{d}$.

The *Wald statistic*

$$w = (\boldsymbol{C}\hat{\boldsymbol{\beta}} - \boldsymbol{d})'[\boldsymbol{C}\boldsymbol{F}^{-1}(\hat{\boldsymbol{\beta}})\boldsymbol{C}']^{-1}(\boldsymbol{C}\hat{\boldsymbol{\beta}} - \boldsymbol{d})$$

measures the distance between the estimate $\boldsymbol{C}\hat{\boldsymbol{\beta}}$ and the hypothetical value $\boldsymbol{d}$ under H_0, weighted with the (inverse) asymptotic covariance matrix $\boldsymbol{C}\boldsymbol{F}^{-1}(\hat{\boldsymbol{\beta}})\boldsymbol{C}'$ of $\boldsymbol{C}\hat{\boldsymbol{\beta}}$.

The *score statistic*

$$u = \boldsymbol{s}'(\tilde{\boldsymbol{\beta}})\boldsymbol{F}^{-1}(\tilde{\boldsymbol{\beta}})\boldsymbol{s}(\tilde{\boldsymbol{\beta}})$$

measures the distance between $\boldsymbol{0} = \boldsymbol{s}(\hat{\boldsymbol{\beta}})$, i.e., the score function evaluated at the ML estimator $\hat{\boldsymbol{\beta}}$, and $\boldsymbol{s}(\tilde{\boldsymbol{\beta}})$, i.e., the score function evaluated at the restricted ML estimator $\tilde{\boldsymbol{\beta}}$.

Wald tests are mathematically convenient when an estimated model is to be tested against a simplified submodel, since it does not require additional estimation of the submodel. Conversely, the score test is convenient when an estimated model is to be tested against a more complex model alternative.

For the special hypothesis Eq. (5.11), the Wald and score statistic are reduced to

$$w = \hat{\boldsymbol{\beta}}_1'\hat{\boldsymbol{A}}_1^{-1}\hat{\boldsymbol{\beta}}_1$$

and

$$u = \boldsymbol{s}_1(\tilde{\boldsymbol{\beta}}_1)'\tilde{\boldsymbol{A}}_1\boldsymbol{s}_1(\tilde{\boldsymbol{\beta}}_1),$$

where $\boldsymbol{A}_1$ represents the submatrix of $\boldsymbol{A} = \boldsymbol{F}^{-1}$ and $\boldsymbol{s}_1(\tilde{\boldsymbol{\beta}}_1)$ represents the subvector of the score function $\boldsymbol{s}(\tilde{\boldsymbol{\beta}})$ that corresponds to the elements of $\tilde{\boldsymbol{\beta}}_1$. The notation "ˆ" or "˜" reflects the respective evaluation at $\hat{\boldsymbol{\beta}}$ or $\tilde{\boldsymbol{\beta}}$.

Under weak regularity conditions, similar to those required for the asymptotic normality of the ML estimators, the three test statistics are asymptotically equivalent under H_0 and approximately follow a χ^2-distribution with r degrees of freedom:

$$lr, w, u \overset{a}{\sim} \chi^2_r.$$

Critical values or p-values are calculated using this asymptotic distribution. For moderate sample sizes, the approximation through the χ^2-distribution is generally sufficient. For a smaller sample size, e.g., $n \leq 50$, the values of the test statistics can, however, differ considerably.

In the special case $H_0 : \beta_j = 0$ versus $H_1 : \beta_j \neq 0$ the Wald statistic equals the squared "t-value"

$$w = t_j^2 = \frac{\hat{\beta}_j^2}{a_{jj}},$$

with a_{jj} as the jth diagonal element of the asymptotic covariance matrix $\boldsymbol{A} = \boldsymbol{F}^{-1}(\hat{\boldsymbol{\beta}})$. Then the test is usually based on t_j which is asymptotically $\mathrm{N}(0, 1)$ distributed. The null hypothesis is then rejected if $|t_j| > z_{1-\alpha/2}$ where $z_{1-\alpha/2}$ is the $(1 - \alpha/2)$-quantile of the $\mathrm{N}(0, 1)$ distribution.

5.1.4 Criteria for Model Fit and Model Choice

Assessing the fit of an estimated model relies on the following idea: When the data have been maximally grouped, we can estimate the group-specific parameter π_i using the mean value $\bar{y}_i$. The use of these mean values as estimators corresponds to the *saturated model*, i.e., the model which contains separate parameters for each group. Thus the saturated model provides the best fit to the data and serves as a benchmark when evaluating the fit of estimated regression models. We now can formally test whether the departure between the estimated model and the saturated model is significant or not. The Pearson statistic and the deviance are the most frequently used goodness-of-fit statistics used for testing such a departure, both requiring that the data have been grouped as much as possible.

The *Pearson statistic* is given by the sum of the squared standardized residuals:

$$\chi^2 = \sum_{i=1}^{G} \frac{(\bar{y}_i - \hat{\pi}_i)^2}{\hat{\pi}_i(1 - \hat{\pi}_i)/n_i},$$

where G represents the number of groups, $\bar{y}_i$ is the relative frequency for group i, $\hat{\pi}_i = h(\boldsymbol{x}_i'\hat{\boldsymbol{\beta}})$ is the probability $\mathrm{P}(y_i = 1)$ estimated by the model, and $\hat{\pi}_i(1-\hat{\pi}_i)/n_i$ is the corresponding estimated variance.

The *deviance* is defined by

$$D = -2\sum_{i=1}^{G}\{l_i(\hat{\pi}_i) - l_i(\bar{y}_i)\},$$

where $l_i(\hat{\pi}_i)$ and $l_i(\bar{y}_i)$ represent the log-likelihood of group i for the estimated and the saturated model, respectively. The Pearson statistic looks similar to conventional chi-square statistics for testing if a random sample comes from a hypothesized discrete distribution: The squared differences between data and estimates are standardized by the variance and then summed up. The deviance compares the

Table 5.3 Patent opposition: estimation results from the logit model

Variable	Coefficient	Standard error	t-value	p-value	95 % Confidence interval	
intercept	201.740	22.321	9.04	<0.001	157.991	245.489
year	−0.102	0.011	−9.10	<0.001	−0.124	−0.080
ncit	0.114	0.022	5.09	<0.001	0.070	0.157
nclaim	0.027	0.006	4.49	<0.001	0.015	0.038
ustwin	−0.403	0.100	−4.03	<0.001	−0.599	−0.207
patus	−0.526	0.113	−4.67	<0.001	−0.747	−0.306
patgsgr	0.197	0.117	1.68	0.094	−0.033	0.427
ncountry	0.098	0.015	6.55	<0.001	0.068	0.127

(maximum of the) log-likelihood of the estimated model to the value of the log-likelihood of the saturated model, i.e., the largest value of the log-likelihood that can be attained. For finite samples, the Pearson and the deviance statistic will differ, but it can be shown that they are asymptotically equivalent for grouped data. If the n_i are sufficiently large for *all* groups $i = 1, \dots, G$, both statistics are approximately χ^2_{G-p}-distributed, where p represents the number of estimated coefficients. Based on this approximate distribution, we can conduct a formal test for model fit by comparing the observed value of the test statistic to the corresponding quantile of the χ^2_{G-p}-distribution. Larger values in the observed test statistic indicate lack of fit and therefore correspond to larger p-values. For a prespecified significance level α a model is rejected if the $(1-\alpha)$-quantile is exceeded or the p-value is smaller than α. However, if n_i is small (especially if $n_i = 1$ as with ungrouped individual data), conducting such a test can be problematic. In this case, large values of χ^2 or D do not necessarily indicate a poor fit.

As already discussed for the coefficient of determination in linear regression (section “Analysis of Variance and Coefficient of Determination” of Sect. 3.2.3), a model choice strategy that tries to make the goodness-of-fit statistics as small as possible will usually result in an overfit model choice. When comparing models with different predictors and parameters, a compromise should be found between a good model fit obtained with a large number of parameters and model complexity. A well-known compromise is Akaike’s information criterion

$$\text{AIC} = -2l(\hat{\boldsymbol{\beta}}) + 2p\,,$$

in which the term $2p$ penalizes complex models with a large number of parameters. When choosing between several models, we prefer those with small AIC values. Rather than the AIC value, one also often considers AIC/n, i.e., the AIC standardized for sample size n. Another alternative is the BIC; see Appendix B.5.4.

Example 5.4 Patent Opposition—Testing and Model Choice

Table 5.3 presents the estimated coefficients for the logit model in Example 5.1 (p. 275), along with the corresponding standard errors, t-values, p-values, and 95 % confidence intervals. For the log-likelihood and the AIC criterion of the estimated model, we have

Table 5.4 Patent opposition: estimation results from the probit model

Variable	Coefficient	Standard error	t-value	p-value	95 % Confidence interval	
intercept	119.204	13.192	9.04	<0.001	93.349	145.060
year	−0.060	0.007	−9.11	<0.001	−0.073	−0.047
ncit	0.068	0.014	5.02	<0.001	0.041	0.094
nclaim	0.016	0.004	4.46	<0.001	0.009	0.023
ustwin	−0.243	0.060	−4.07	<0.001	−0.360	−0.126
patus	−0.309	0.066	−4.72	<0.001	−0.438	−0.181
patgsgr	0.121	0.071	1.71	0.086	−0.017	0.260
ncountry	0.059	0.009	6.51	<0.001	0.041	0.076

Table 5.5 Patent opposition: estimation results of the extended logit model

Variable	Coefficient	Standard error	t-value	p-value	95 % Confidence interval	
intercept	198.131	22.739	8.71	<0.001	153.563	242.699
year	−0.101	0.011	−8.82	<0.001	−0.123	−0.078
ncit	0.113	0.022	5.08	<0.001	0.070	0.157
nclaim	0.026	0.006	4.45	<0.001	0.015	0.038
ustwin	−0.409	0.100	−4.09	<0.001	−0.605	−0.213
patus	−0.539	0.113	−4.77	<0.001	−0.761	−0.318
patgsgr	0.180	0.119	1.52	0.130	−0.053	0.414
ncountry	0.394	0.184	2.14	0.032	0.034	0.754
$ncountry^2$	−0.038	0.024	−1.58	0.113	−0.085	0.009
$ncountry^3$	0.001	0.001	1.50	0.134	−0.000	0.003

$$l(\hat{\boldsymbol{\beta}}) = -1488.560\,, \quad \text{AIC} = 2993.12\,.$$

With a p-value of 0.094, the effect of the variable *patgsgr* is at best marginally significant. If we choose $\alpha = 5\,\%$ as the significance level, the hypothesis $H_0 : \beta_6 = 0$ will not be rejected. This implies that the increased probability of patent objection if the patent comes from Germany, Switzerland, or Great Britain appears nonsignificant.

Table 5.4 contains the corresponding values for the probit model. Even though the estimated coefficients and their standard deviations differ due to the absence of a proper adjustment (see Example 5.1), the t-values and p-values are in good agreement and lead to the same conclusions regarding the significance of the effects. With

$$l(\hat{\boldsymbol{\beta}}) = -1488.407\,, \quad \text{AIC} = 2992.815,$$

we obtain very similar values for the log-likelihood and the AIC criterion. Since the results for the patent data are comparable for the logit and probit model, we only further describe the findings for the logit model.

In order to examine whether or not the effect of the continuous covariate *ncountry* is linear, we included a cubic polynomial

$$\beta_7\, ncountry + \beta_8\, ncountry^2 + \beta_9\, ncountry^3$$

into the linear predictor as in Example 2.8 (p. 35) and estimated this modified logit model. Table 5.5 contains the estimated coefficients, their standard errors, as well as t-values,

Table 5.6 Credit scoring: description of the covariates including summary statistics

Variable	Description	Mean/ frequency in %	Std. dev.	Min/max
acc1	1 = no running account	27.40		
	0 = good or bad running account	72.60		
acc2	1 = good running account	39.40		
	0 = no or bad running account	60.60		
duration	Duration of the credit in months	20.90	12.06	4/72
amount	Credit amount in 1000 Euro,	1.67	1.44	0.13/9.42
moral	Previous payment behavior			
	1 = good	91.10		
	0 = bad	8.90		
intuse	Intended use			
	1 = private	65.70		
	0 = business	34.30		

p-values, and 95 % confidence intervals. The t-values and the p-values corresponding to *ncountry*^{2} and *ncountry*^{3} already indicate that the more conservative linear model may be sufficient and that the nonlinearity is over-interpreted. The log-likelihood and the AIC criterion for the extended model yields

$$l(\hat{\boldsymbol{\beta}}) = -1487.232\,, \quad \text{AIC} = 2994.463\,.$$

This further confirms that we should rather choose the simpler model with linear modeling of the *ncountry* effect. We can also investigate nonlinearity by testing the hypotheses

$$H_0 : (\beta_8, \beta_9) = (0, 0) \qquad \text{versus} \qquad H_1 : (\beta_8, \beta_9) \neq (0, 0).$$

The likelihood ratio test statistic results in

$$lr = -2\{-1488.56 - (-1487.23)\} = 2.66\,.$$

The 95 % quantile of the (approximate) $\chi^2(2)$-distribution is $\chi^2_{95\,\%}(2) = 5.99$, thus H_0 cannot be rejected, which also follows from the p-value of 0.269. In summary, the assumption of a linear effect of covariate *ncountry* cannot be rejected. The Wald test also leads to the same result.

△

Example 5.5 Credit Scoring—Binary Regression

When issuing credit, banks check the "solvency" or "creditworthiness" of clients, i.e., their ability and willingness to pay back the credit in the specified time frame. To evaluate creditworthiness using statistical methods (credit scoring), characteristics of the borrower are requested that reflect his or her personal and economic situation and thus influence the probability of creditworthiness. Binary regression models are suited for such evaluations since they model the probability of a loan default for given characteristics of the client.

We use a data set on $n = 1{,}000$ private credits issued by a German bank published in Fahrmeir, Hamerle, and Tutz (1996). Every client is associated with a binary response y defined as

$$y = \begin{cases} 1 & \text{client is not creditworthy,} \\ 0 & \text{client is creditworthy.} \end{cases}$$

Among a total of 20 characteristics, we use those described in Table 5.6 as covariates.

Table 5.7 Credit scoring: estimation results for the logit model

Variable	Coefficient	Standard error	t-value	p-value	95 % Confidence interval	
intercept	0.487	0.266	1.83	0.067	−0.034	1.007
acc1	0.618	0.175	3.53	<0.001	0.275	0.960
acc2	−1.338	0.201	−6.65	<0.001	−1.732	−0.944
durationo	0.401	0.093	4.29	<0.001	0.218	0.584
amounto	0.066	0.092	0.72	0.474	−0.115	0.247
moral	−0.986	0.251	−3.93	<0.001	−1.478	−0.494
intuse	−0.426	0.158	−2.69	0.007	−0.736	−0.115

Table 5.8 Credit scoring: results for the extended logit model

Variable	Coefficient	Standard error	t-value	p-value	95 % Confidence interval	
intercept	0.474	0.270	1.75	0.079	−0.055	1.004
acc1	0.618	0.176	3.51	<0.001	0.272	0.963
acc2	−1.337	0.202	−6.61	<0.001	−1.734	−0.941
durationo	0.508	0.100	5.07	<0.001	0.312	0.705
duration2o	−0.173	0.079	−2.20	0.028	−0.327	−0.019
amounto	0.035	0.098	0.36	0.720	−0.155	−0.224
amount2o	0.288	0.097	3.07	0.002	0.104	0.471
moral	−0.995	0.255	−3.90	<0.001	−1.495	−0.495
intuse	−0.404	0.160	−2.52	0.012	−0.718	−0.090

We model the probability $\mathrm{P}(y = 1)$ for a weak creditworthiness with a logit model and the linear predictor

$$\eta = \beta_0 + \beta_1\, acc1 + \beta_2\, acc2 + \beta_3\, durationo + \beta_4\, amounto + \beta_5\, moral + \beta_6\, intuse.$$

Since we will later also estimate quadratic orthogonal polynomials (see Example 3.5 on p. 90) for the effects of the continuous covariates *duration* and *amount* we included the linear parts *durationo* and *amounto* of these orthogonal polynomials in the predictor rather than the original covariates. Table 5.7 lists the estimated coefficients, along with their corresponding standard errors, t-values, p-values, and 95 % confidence intervals. The p-value for the effect of *amounto* indicates that the corresponding effect is not significant. The AIC value for this model is 1,043.815.

In a second step of our analysis, we assume a quadratic orthogonal polynomial for the effects of the continuous covariates *duration* and *amount* to detect possible nonlinearity. Table 5.8 contains the corresponding estimated results. All p-values, also those for squared effects, now show significance. Furthermore, the lower AIC value of 1,035.371 indicates an improved model fit.

Figure 5.2 shows the estimated linear effects of credit amount and duration together with the quadratic, nonlinear effects. The "bathtub" shape of the squared effects of the credit amount illustrates that small and large credit increases the risk of not paying back the credit. This effect is missed when reducing the model to linear effects.

We finally apply the likelihood ratio and Wald test in order to test the model having quadratic effects against the submodel with linear effects for credit amount and duration. The likelihood ratio and the Wald statistic yield

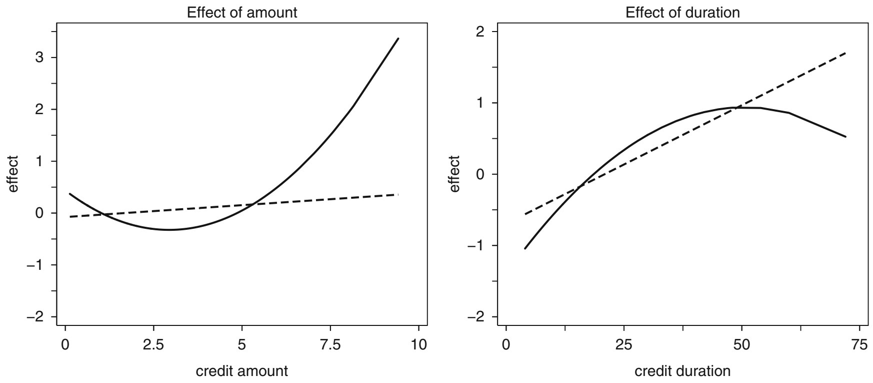

Fig. 5.2 Credit scoring: estimated linear (- - -) and quadratic (—) effects of credit amount and credit duration

$$lr = 12.44\,, \qquad w = 11.47$$

with two degrees of freedom and corresponding p-values of 0.0020 and 0.0032, respectively. Hence, both tests again confirm the specification of the more complex model with quadratic effects.

△

5.1.5 Estimation of the Overdispersion Parameter

As discussed in Sect. 5.1 (p. 279), we may observe overdispersion when working with grouped data. To allow for overdispersion, we assume

$$\text{Var}(y_i) = \phi \frac{\pi_i(1-\pi_i)}{n_i}.$$

The overdispersion parameter ϕ can be estimated as the average Pearson statistic or the average deviance:

$$\hat{\phi}_P = \frac{1}{n-p}\chi^2 \quad \text{or} \quad \hat{\phi}_D = \frac{1}{n-p} D\,.$$

This is analogous to the estimation of the error variance in the linear model, with χ^2 or D replacing the residual sum of squares.

Accordingly, we multiply the estimated covariance matrix with $\hat{\phi}$, i.e., $\widehat{\text{Cov}}(\hat{\boldsymbol{\beta}}) = \hat{\phi}\boldsymbol{F}^{-1}(\hat{\boldsymbol{\beta}})$. Strictly speaking, this approach to treat overdispersion does not correspond to a true likelihood method, but rather to a quasi-likelihood model; see Sect. 5.5.

Since we only need $\pi_i = \mathrm{E}(y_i)$ and $\mathrm{Var}(y_i)$, and not the likelihood itself for the maximum likelihood estimation of $\boldsymbol{\beta}$, both $\boldsymbol{\beta}$ and ϕ can be formally estimated just as if we considered a distribution with scale parameter ϕ, such as a normal or gamma distribution; see Sect. 5.4.2. In fact, the introduction of an overdispersion parameter leads to one of the simplest forms of quasi-likelihood estimation. Even though distributions with variance $\phi\pi_i(1-\pi_i)/n_i$ exist, for example, the beta-binomial distribution, their actual likelihood is not necessary and will also not be used in the estimation process. Other approaches to account for overdispersion are, for example, models with random effects; see Chap. 7. A good additional reference on models with overdispersion is Collett (1991).

5.2 Count Data Regression

Count data are frequently observed when the number of certain events within a fixed time frame or frequencies in a contingency table have to be analyzed. Sometimes, a normal approximation can be sufficient, particularly when the events occur with high frequencies. In situations with only a small number of counts, models for categorial response variables (Chap. 6) can be an alternative. In general, however, discrete distributions recognizing the specific properties of count data are most appropriate. The Poisson distribution is the simplest and most widely used choice, but model modifications and alternatives such as the negative binomial distribution are also used. For details on such extensions, we refer to the specialized literature on count data regression given in the final section of this chapter.

5.2.1 Models for Count Data

Log-Linear Poisson Model

The most widely used model for count data connects the rate $\lambda_i = \mathrm{E}(y_i)$ of the Poisson distribution with the linear predictor $\eta_i = \boldsymbol{x}_i'\boldsymbol{\beta} = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik}$ via

$$\lambda_i = \exp(\eta_i) = \exp(\beta_0)\exp(\beta_1 x_{i1}) \cdot \ldots \cdot \exp(\beta_k x_{ik})$$

or in log-linear form through

$$\log(\lambda_i) = \eta_i = \boldsymbol{x}_i'\boldsymbol{\beta} = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik}. \tag{5.12}$$

The effect of covariates on the rate λ is, thus, exponentially multiplicative similar to the effect on the odds $\pi/(1-\pi)$ in the logit model. The effect on the logarithm of the rate in Eq. (5.12) is linear.

Linear Poisson Model

The direct relationship

$$\lambda_i = \eta_i = \boldsymbol{x}_i'\boldsymbol{\beta}$$

5.3 Log-Linear Poisson Model for Count Data

Data

The response variables y_i take values $\{0, 1, 2, \dots\}$ and are (conditionally) independent given the covariates $x_{i1}, \dots, x_{ik}$.

Model Without Overdispersion

$y_i \sim \text{Po}(\lambda_i)$ with

$$\lambda_i = \exp(\boldsymbol{x}_i'\boldsymbol{\beta}) \quad \text{or} \quad \log(\lambda_i) = \boldsymbol{x}_i'\boldsymbol{\beta}.$$

Model with Overdispersion

$$\text{E}(y_i) = \lambda_i = \exp(\boldsymbol{x}_i'\boldsymbol{\beta}), \quad \text{Var}(y_i) = \phi\lambda_i$$

with overdispersion parameter ϕ.

is useful when the covariates have an additive effect on the rate. Since $\boldsymbol{x}_i'\boldsymbol{\beta}$ must not be nonnegative, this usually implies restrictions for the parameter space of $\boldsymbol{\beta}$.

Overdispersion

The assumption of a Poisson distribution for the responses implies

$$\lambda_i = \text{E}(y_i) = \text{Var}(y_i).$$

For similar reasons as in case with binomial data, a significantly higher empirical variance is frequently observed in applications of Poisson regression. For this reason it is often useful to introduce an overdispersion parameter ϕ by assuming

$$\text{Var}(y_i) = \phi\lambda_i.$$

As for binomial data, there are also more complex modeling approaches for count data which take the additional variability into account. One possibility is the use of the negative binomial distribution, which is closely related to the use of random effects models; see Chap. 7.

Example 5.6 Number of Citations from Patents—Poisson Regression

We illustrate the use of regression models for counts with the patent data described in Example 1.3 (p. 8). In contrast to Examples 5.1 and 5.4, we now choose the number of citations for a patent, variable *ncit*, as the response. We also use the complete data set, i.e., patents which either belong to the biotechnology or to the pharmaceutical industry. We incorporate the remaining variables described in Table 1.4 (p. 8) as covariates. As in Sect. 3.1.2 (p. 92), we center the continuous covariates *yearc*, *ncountryc*, and *nclaimsc* around their means and use these centered covariates in the linear predictor. Based on previous descriptive analysis in Example 2.8, we exclude all observations with *nclaims* $>$ 60 and *ncit* $>$ 15 from further analysis.

As a first step, we examine a log-linear model for the rate $\lambda_i = \mathrm{E}(ncit_i)$ with purely linear predictor

$$\begin{aligned}\log(\lambda_i) = \eta_i &= \beta_0 + \beta_1 yearc_i + \beta_2 ncountryc_i + \beta_3 nclaimc_i + \beta_4 biopharm_i \\ &\quad + \beta_5 ustwin_i + \beta_6 patus_i + \beta_7 patgsgr_i + \beta_8 opp_i\,.\end{aligned}$$

In Example 5.7, we estimate a Poisson model without an overdispersion parameter, as well as a model that includes an overdispersion parameter ϕ in the variance $\mathrm{Var}(ncit_i) = \phi\lambda_i$. To allow for possibly nonlinear effects of the continuous covariates, we further considered polynomial effects for *yearc*, *ncountryc*, and *nclaimsc* and compare the different models using AIC.

△

5.2.2 Estimation and Testing: Likelihood Inference

We again assume that the response variables y_i are (conditionally) independent. The derivations of the likelihood, score function, and the information matrix are analogous to the developments for binary data in Sect. 5.1.

Maximum Likelihood Estimation

For the Poisson distributed response variable, the discrete density (or the likelihood $L_i(\boldsymbol{\beta})$ of the ith observation) is given by

$$f(y_i \mid \boldsymbol{\beta}) = \frac{\lambda_i^{y_i}\exp(-\lambda_i)}{y_i!}, \quad \mathrm{E}(y_i) = \lambda_i\,.$$

It depends on $\boldsymbol{\beta}$ through $\lambda_i = \boldsymbol{x}_i'\boldsymbol{\beta}$ in the linear Poisson model and through $\lambda_i = \exp(\boldsymbol{x}_i'\boldsymbol{\beta})$ in the log-linear Poisson model. The ML estimator for the log-linear Poisson model is obtained in the following steps:

1. Log-likelihood

The *log-likelihood* is given by

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{n} l_i(\boldsymbol{\beta}) = \sum_{i=1}^{n}(y_i\log(\lambda_i) - \lambda_i),$$

apart from the additive constant $-n\log(y_i!)$ (that is independent of $\boldsymbol{\beta}$). The Poisson log-linear model with $\log(\lambda_i) = \boldsymbol{x}_i'\boldsymbol{\beta} = \eta_i$ yields

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{n} l_i(\boldsymbol{\beta}) = \sum_{i=1}^{n} y_i(\boldsymbol{x}_i'\boldsymbol{\beta}) - \exp(\boldsymbol{x}_i'\boldsymbol{\beta}) = \sum_{i=1}^{n}(y_i\eta_i - \exp(\eta_i))\,.$$

2. Score function
Differentiating according to the chain rule $\partial l_i(\boldsymbol{\beta})/\partial\boldsymbol{\beta} = (\partial l_i/\partial\eta_i)\cdot\partial\eta_i/\partial\boldsymbol{\beta} = \partial l_i/\partial\eta_i\cdot\boldsymbol{x}_i$, we obtain the *score function*

$$\boldsymbol{s}(\boldsymbol{\beta}) = \sum_{i=1}^{n}\boldsymbol{x}_i(y_i - \exp(\eta_i)) = \sum_{i=1}^{n}\boldsymbol{x}_i(y_i - \lambda_i).$$

3. Fisher information
Using the same arguments as in the logit model, we obtain the *Fisher information*

$$\boldsymbol{F}(\boldsymbol{\beta}) = \mathrm{E}(\boldsymbol{s}(\boldsymbol{\beta})\boldsymbol{s}'(\boldsymbol{\beta})) = \sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i'\lambda_i,$$

utilizing $\mathrm{E}(y_i - \lambda_i)^2 = \mathrm{Var}(y_i) = \lambda_i$.

4. Numerical computation
Due to $\lambda_i = \exp(\boldsymbol{x}_i'\boldsymbol{\beta})$, we obtain the nonlinear equation system

$$\boldsymbol{s}(\hat{\boldsymbol{\beta}}) = \boldsymbol{0}$$

for $\hat{\boldsymbol{\beta}}$. The numerical computation of $\hat{\boldsymbol{\beta}}$ is again carried out using Fisher scoring Eq. (5.10), inserting the corresponding expressions for $\boldsymbol{s}(\boldsymbol{\beta})$ and $\boldsymbol{F}(\boldsymbol{\beta})$. Similar to linear and binary regression, we assume

$$\mathrm{rk}(\boldsymbol{X}) = p$$

for the design matrix $\boldsymbol{X} = (\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n)'$. The remarks made in Sect. 5.1.2 also apply for the convergence or divergence of the iterations in the Poisson model.

Under moderate regularity assumptions and for large n (more precisely for $n \to \infty$), we have the asymptotic result

$$\hat{\boldsymbol{\beta}} \overset{a}{\sim} \mathrm{N}(\boldsymbol{\beta}, \boldsymbol{F}^{-1}(\hat{\boldsymbol{\beta}}))$$

with estimated covariance matrix $\widehat{\mathrm{Cov}}(\hat{\boldsymbol{\beta}}) = \boldsymbol{F}^{-1}(\hat{\boldsymbol{\beta}})$.

Testing Linear Hypotheses

We use the same test statistics as in binary regression models for testing linear hypotheses $\boldsymbol{C}\boldsymbol{\beta} = \boldsymbol{d}$, where the appropriate expressions for $l(\boldsymbol{\beta}), \boldsymbol{s}(\boldsymbol{\beta})$, and $\boldsymbol{F}(\boldsymbol{\beta})$ associated with the Poisson model are to be inserted. In addition, the same statements regarding the asymptotic or approximate χ^2-distribution of the test statistics apply.

5.2.3 Criteria for Model Fit and Model Choice

The criteria discussed in Sect. 5.1.2 for binary regression models can be transferred to the Poisson case. Since $\text{Var}(y_i) = \lambda_i$ for the Poisson distribution, we obtain the Pearson statistic

$$\chi^2 = \sum_{i=1}^{G} \frac{(y_i - \hat{\lambda}_i)^2}{\hat{\lambda}_i / n_i}.$$

The Poisson log-likelihood must be inserted into the definition of the deviance and the AIC. Note that by convention $0 \cdot \log(0) = 0$ (for $y_i = 0$).

5.2.4 Estimation of the Overdispersion Parameter

As previously stated, in situations where we allow for possible overdispersion with the assumption $\text{Var}(y_i \mid \boldsymbol{x}_i) = \phi \lambda_i$, the overdispersion parameter ϕ can be estimated as the average Pearson statistic or the average deviance:

$$\hat{\phi}_P = \frac{1}{n-p} \chi^2 \quad \text{or} \quad \hat{\phi}_D = \frac{1}{n-p} D \,.$$

This is analogous to the estimation of the error variance in the linear model, with χ^2 or D replacing the residual sum of squares.

We then have to multiply the estimated covariance matrix with $\hat{\phi}$, i.e., $\widehat{\text{Cov}}(\hat{\boldsymbol{\beta}}) = \hat{\phi} \boldsymbol{F}^{-1}(\hat{\boldsymbol{\beta}})$. Strictly speaking, this approach to the estimation of overdispersion does not correspond to a true likelihood method, but rather to a quasi-likelihood model; see Sect. 5.5.

Example 5.7 Number of Citations from Patents—Overdispersion

Table 5.9 shows estimation results for the log-linear Poisson model of Example 5.6 having no overdispersion (i.e., $\phi = 1$) and only linear effects (AIC $= 19{,}092.25$, deviance $= 12{,}085.31$, Pearson-$\chi^2 = 14{,}091.66$).

The p-values indicate significance of all covariates, with the exception of *ustwin*. Since overdispersion is very common with Poisson models, we reanalyze the model by estimating the overdispersion parameter as the mean Pearson statistic or mean deviance. We obtain

$$\hat{\phi}_P = \frac{1}{n-p} \chi^2 = 2.935 \quad \text{resp.} \quad \hat{\phi}_D = \frac{1}{n-p} D = 2.518 \,,$$

with $n = 4{,}809$, $p = 9$. In contrast to the Poisson model, the estimated variance or standard deviation of the estimated regression coefficients needs to be multiplied by $\hat{\phi}$ and $\hat{\phi}^{1/2}$, respectively, while the point estimates are the same as for the pure Poisson model without overdispersion. Table 5.10 lists the results for $\hat{\phi}_P$. In comparison to Table 5.9, we see that the standard errors increase by the factor $\hat{\phi}_P^{1/2} = 1.71$. This adjustment causes an increase of the p-values, such that the effect of variable *patus* is now insignificant, while the analysis without overdispersion resulted in a p-value that was significant.

In order to detect possibly nonlinear effects of the centered continuous covariates *yearc*, *ncountryc*, and *nclaimc*, we construct polynomials of degree three. The centering

Table 5.9 Number of citations from patents: model with linear effects and $\phi = 1$

Variable	Coefficient	Standard error	t-value	p-value	95 % Confidence interval	
intercept	0.158	0.033	4.85	<0.001	0.094	0.222
yearc	−0.072	0.003	−24.17	<0.001	−0.078	−0.066
ncountryc	−0.028	0.004	−6.60	<0.001	−0.036	−0.020
nclaimc	0.018	0.001	14.16	<0.001	0.016	0.021
biopharm	0.239	0.032	7.42	<0.001	0.176	0.302
ustwin	0.002	0.026	0.09	0.926	−0.048	0.053
patus	−0.078	0.027	−2.84	0.005	−0.132	−0.024
patgsgr	−0.198	0.032	−6.24	<0.001	−0.260	−0.136
opp	0.372	0.025	14.81	<0.001	0.322	0.421

Table 5.10 Number of citations from patents: model with linear effects and overdispersion

Variable	Coefficient	Standard error	t-value	p-value	95 % Confidence interval	
intercept	0.158	0.056	2.83	0.005	0.049	0.267
yearc	−0.072	0.005	−14.11	<0.001	−0.082	−0.062
ncountryc	−0.028	0.007	−3.85	<0.001	−0.042	−0.014
nclaimc	0.018	0.002	8.26	<0.001	0.014	0.022
biopharm	0.239	0.055	4.33	<0.001	0.131	0.347
ustwin	0.002	0.044	0.05	0.957	−0.084	0.088
patus	−0.078	0.047	−1.66	0.098	−0.170	0.014
patgsgr	−0.198	0.054	−3.64	<0.001	−0.305	−0.091
opp	0.372	0.043	8.64	<0.001	0.287	0.456

is conducted as described in section "Continuous Covariates" of Sect. 3.1.3. The model obtains AIC $= 18,786.32$, deviance$= 11,767.37$, Pearson-$\chi^2 = 13,815.96$, $\hat{\phi}_D = 2.45$, $\hat{\phi}_P = 2.88$. Compared to the model with only linear effects the fit is considerably improved. Table 5.11 shows the results. The variable *ustwin* remains nonsignificant while *patus* is now weakly significant. The other variables all remain significant (on a level of 5 %) with the exception of some of the polynomial terms. This indicates that lower-order polynomials might be sufficient to model the nonlinearity of the covariate effects. Figure 5.3 compares linear and nonlinear effects of the continuous covariates. We leave the interpretation of the results to the reader; see Example 2.13 (p. 54) on how to interpret the (nonlinear) effects in Poisson regression models.

△

5.3 Models for Nonnegative Continuous Response Variables

The classical linear model

$$y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + \varepsilon_i, \quad \mathrm{E}(\varepsilon_i) = 0, \quad \mathrm{Var}(\varepsilon_i) = \sigma^2$$

Table 5.11 Number of citations from patents: extended model with overdispersion

Variable	Coefficient	Standard error	t-value	p-value	95 % Confidence interval	
intercept	0.17115	0.05558	3.08	0.002	0.06221	0.28009
yearc	−0.09924	0.00966	−10.27	<0.001	−0.11818	−0.08031
$yearc^2$	−0.00974	0.00226	−4.31	<0.001	−0.01417	−0.00531
$yearc^3$	−0.00011	0.00030	−0.37	0.715	−0.00070	0.00048
ncountryc	0.01552	0.01322	1.17	0.241	−0.01040	0.04143
$ncountryc^2$	−0.00213	0.00206	−1.03	0.301	−0.00618	0.00191
$ncountryc^3$	−0.00157	0.00044	−3.60	<0.001	−0.00243	−0.00071
nclaimc	0.02611	0.00352	7.42	<0.001	0.01922	0.03301
$nclaimc^2$	−0.00046	0.00036	−1.30	0.195	−0.00116	0.00024
$nclaimc^3$	0.00000	0.00000	0.18	0.855	−0.00002	0.00002
biopharm	0.15504	0.05564	2.79	0.005	0.04598	0.26410
ustwin	−0.00288	0.04338	−0.07	0.947	−0.08791	0.08215
patus	−0.09502	0.04715	−2.02	0.044	−0.18743	−0.00260
patgsgr	−0.20185	0.05446	−3.71	<0.001	−0.30859	−0.09511
opp	0.37154	0.04253	8.74	<0.001	0.28819	0.45489

is well suited for analyzing regression data when the errors ε_i have (at least approximately) a normal distribution. In this case, the response variables y_i, for given covariate vector $\boldsymbol{x}_i$, are (conditionally) independent and follow a normal distribution with

$$y_i \sim \mathrm{N}(\mu_i, \sigma^2), \qquad \mu_i = \mathrm{E}(y_i) = \boldsymbol{x}_i'\boldsymbol{\beta}.$$

In many applications, the response variable cannot be negative, for example, in case of life times, claim sizes, and costs. Such responses are also usually highly non-normal, often following a (right) skewed distribution.

Lognormal Model

To enable the application of linear models, the response variable y is often transformed logarithmically such that a usual linear model with normal errors can be assumed for $\tilde{y} = \log(y)$, i.e.,

$$\tilde{y}_i = \boldsymbol{x}_i'\boldsymbol{\beta} + \varepsilon_i \quad \text{or} \quad \tilde{y}_i \sim \mathrm{N}(\boldsymbol{x}_i'\boldsymbol{\beta}, \sigma^2).$$

This implies that the original variable y follows a log-normal distribution (see Definition B.6 in Appendix B.1) with

$$\mathrm{E}(y_i) = \exp(\boldsymbol{x}_i'\boldsymbol{\beta} + \sigma^2/2), \qquad \mathrm{Var}(y_i) = \exp(2\boldsymbol{x}_i'\boldsymbol{\beta} + \sigma^2)(\exp(\sigma^2) - 1). \quad (5.13)$$

We can obtain "plug-in" estimators for Eq. (5.13), using the least squares estimates $\hat{\boldsymbol{\beta}}$ and the estimated variance $\hat{\sigma}^2$ for the linear model. When estimating $\hat{\mu}_i = \exp(\hat{\eta}_i) = \exp(\boldsymbol{x}_i\hat{\boldsymbol{\beta}} + \hat{\sigma}^2/2)$, considerable bias can be induced by the nonlinear back-transformation with the exponential function.

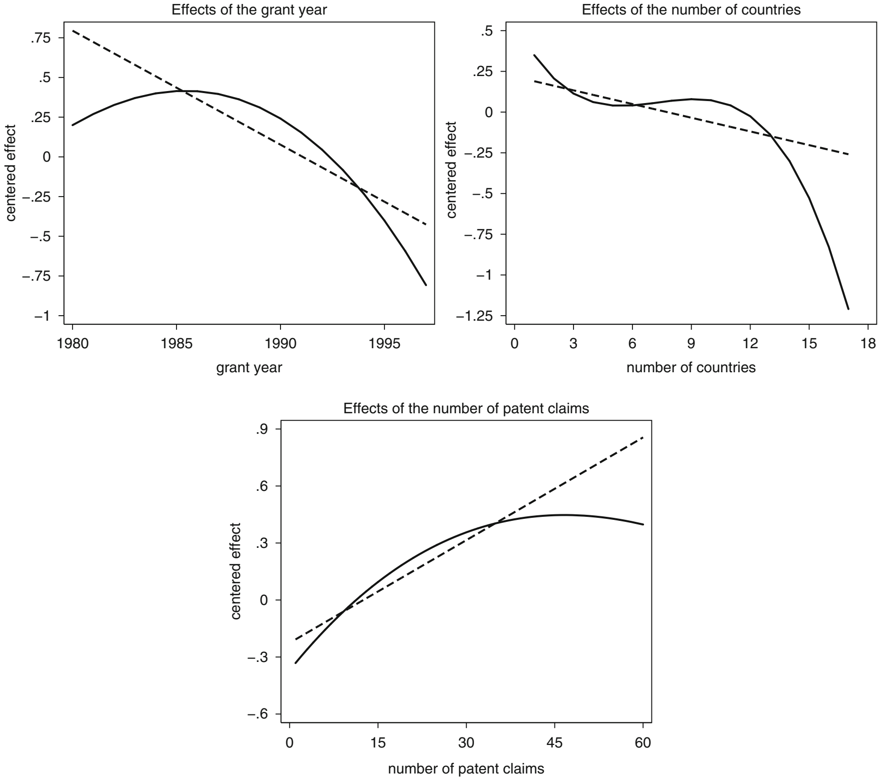

Fig. 5.3 Number of citations from patents: linear (- - -) and nonlinear (—) effects of the continuous covariates *year*, *ncountry*, and *nclaim*

Gamma Regression

To circumvent this difficulty, the assumption of a gamma distribution (see Definition B.9 in Appendix B.1), with expectation $\text{E}(y_i) = \mu_i$ and scale parameter ν for the response variables y_i, can be a valuable alternative. The variance is then given by

$$\text{Var}(y_i) = \sigma_i^2 = \mu_i^2/\nu.$$

For the nonnegative, gamma-distributed response, we have $\text{E}(y_i) = \mu_i > 0$. A direct linear link

$$\mu_i = \boldsymbol{x}_i'\boldsymbol{\beta}$$

is again problematic, since we have to comply with the condition $\boldsymbol{x}_i'\boldsymbol{\beta} > 0$. Thus a *multiplicative exponential model*

$$\mu_i = \exp(\eta_i) = \exp(\boldsymbol{x}_i'\boldsymbol{\beta}) = \exp(\beta_0)\exp(\beta_1 x_{i1}) \cdot \ldots \cdot \exp(\beta_k x_{ik}), \qquad (5.14)$$

with response function $h(\eta) = \exp(\eta)$, is often better suited than the linear link function.

Another possible choice for the response or link function is the *reciprocal*

$$\mu_i = \frac{1}{\eta_i} = \frac{1}{\boldsymbol{x}_i'\boldsymbol{\beta}}.$$

Since $\boldsymbol{x}_i'\boldsymbol{\beta} > 0$ has to be fulfilled, the choice again implies restrictions for $\boldsymbol{\beta}$. Even though the reciprocal response function is the so-called *natural* or *canonical response function* for the gamma distribution (see Sect. 5.4), the multiplicative exponential model (5.14) is usually more adequate for both modeling and interpretation.

5.4 Generalized Linear Models

5.4.1 General Model Definition

The linear model and the regression models for non-normal response variables described in the preceding sections have common properties that can be summarized in a unified framework:

1. The mean $\mu = \mathrm{E}(y)$ of the response y is connected with the linear predictor $\eta = \boldsymbol{x}'\boldsymbol{\beta}$ by a response function h or by a link function $g = h^{-1}$:

$$\mu = h(\eta) \quad \text{or} \quad \eta = g(\mu).$$

2. The distribution of the response variables (normal, binomial, Poisson, and gamma distribution) can be written in the form of a *univariate exponential family*:

5.4 Exponential Family

The density of a univariate exponential family for the response variable y is defined by

$$f(y \mid \theta) = \exp\left(\frac{y\theta - b(\theta)}{\phi}w + c(y, \phi, w)\right).$$

The log-density is given by

$$\log f(y \mid \theta) = \frac{y\theta - b(\theta)}{\phi}w + c(y, \phi, w).$$

The parameter θ is called the *natural* or *canonical* parameter. For the function $b(\theta)$ it is required that $f(y \mid \theta)$ can be normalized and the first and second derivative $b'(\theta)$ and $b''(\theta)$ exist. The second parameter ϕ is a dispersion parameter, while w is a known value (usually a weight).

As a consequence, both the definition of GLMs and the corresponding statistical inference can be presented in a unified framework. More generally, the resulting concepts can also be applied to regression problems with distributions that do not belong to the exponential family. To treat individual data and grouped data simultaneously, we introduce the weight factor w. For individual data, we set $w = 1$, whereas in the case when the response is summarized as a group mean, w is rather set to the corresponding group size. In the case when the sum of the individual responses of group i is selected for the response variable y_i, the weight equals $1/n_i$.

The Bernoulli and Poisson distributions do not include a dispersion parameter, i.e., $\phi = 1$. For the normal distribution, we have $\phi = \sigma^2$. The parameter θ represents the parameter of main interest that is connected to the linear predictor $\eta = \boldsymbol{x}'\boldsymbol{\beta}$, while the parameter ϕ is often considered a "nuisance parameter" of secondary interest. The term $c(y, \phi, a)$ does not depend on θ. It can be shown that

$$\mathrm{E}(y) = \mu = b'(\theta), \quad \mathrm{Var}(y) = \phi\, b''(\theta)/w.$$

Example 5.8 Bernoulli, Poisson, and Normal Distribution

1. *Bernoulli distribution:* A Bernoulli variable has probability mass function or (discrete) density

$$f(y \,|\, \pi) = \mathrm{P}(Y = y) = \pi^y (1-\pi)^{1-y}, \quad y = 0, 1,$$

where $\mathrm{P}(Y = 1) = \pi = \mathrm{E}(Y) = \mu$ and $\mathrm{Var}(Y) = \pi(1-\pi)$. Taking the logarithm yields

$$\log(f(y \,|\, \pi)) = y\log(\pi) - y\log(1-\pi) + \log(1-\pi).$$

If we define $\theta = \log(\pi) - \log(1-\pi) = \log(\pi/(1-\pi))$ as the natural parameter and take $\log(1-\pi) = -\log(1+\exp(\theta))$ into account, we obtain the density in the form of an exponential family:

$$f(y \,|\, \theta) = \exp(y\theta - \log(1+\exp(\theta)),$$

with $b(\theta) = \log(1+\exp(\theta))$, $\phi = 1$ and $c = 0$. Differentiation results in $b'(\theta) = \exp(\theta)/(1+\exp(\theta))$ and $b''(\theta) = \exp(\theta)/(1+\exp(\theta))^2$. Solving $\theta = \log(\pi/(1-\pi))$ for π results in

$$\pi = \exp(\theta)/(1+\exp(\theta))\,,$$

so that

$$\mathrm{E}(y) = b'(\theta) = \pi, \quad \mathrm{Var}(y) = b''(\theta) = \pi(1-\pi)$$

holds.

2. *Poisson distribution:* A Poisson variable has the (discrete) density

$$f(y \,|\, \lambda) = \mathrm{P}(Y = y) = \frac{\lambda^y \exp(-\lambda)}{y!}, \quad y = 0, 1, \ldots$$

The logarithm of this density results in

$$\log(f(y \,|\, \lambda)) = y\log(\lambda) - \lambda - \log(y!).$$

With $\theta = \log(\lambda)$ as the natural parameter, we obtain

$$\log(f(y \,|\, \theta)) = y\theta - \exp(\theta) - \log(y!).$$

Table 5.12 Univariate exponential families

(a) Density

$$f(y\,|\,\theta,\phi,w) = \exp\left(\frac{y\theta - b(\theta)}{\phi}w + c(y,\phi,w)\right)$$

(b) Exponential family parameters

Distribution		$\theta(\mu)$	$b(\theta)$	ϕ
Normal	$\mathrm{N}(\mu,\sigma^2)$	μ	$\theta^2/2$	σ^2
Bernoulli	$\mathrm{B}(1,\pi)$	$\log(\pi/(1-\pi))$	$\log(1+\exp(\theta))$	1
Poisson	$\mathrm{Po}(\lambda)$	$\log(\lambda)$	$\exp(\theta)$	1
Gamma	$\mathrm{G}(\mu,\nu)$	$-1/\mu$	$-\log(-\theta)$	ν^{-1}
Inverse Gaussian	$\mathrm{IG}(\mu,\sigma^2)$	$-1/(2\mu^2)$	$-(-2\theta)^{1/2}$	σ^2

(c) Expectation and variance

Distribution	$E(y) = b'(\theta)$	$b''(\theta)$	$\mathrm{Var}(y) = b''(\theta)\phi/w$
Normal	$\mu = \theta$	1	σ^2/w
Bernoulli	$\pi = \frac{\exp(\theta)}{1+\exp(\theta)}$	$\pi(1-\pi)$	$\pi(1-\pi)/w$
Poisson	$\lambda = \exp(\theta)$	λ	λ/w
Gamma	$\mu = -1/\theta$	μ^2	$\mu^2\nu^{-1}/w$
Inverse Gaussian	$\mu = (-2\theta)^{-1/2}$	μ^3	$\mu^3\sigma^2/w$

It follows that $b(\theta) = \exp(\theta) = \lambda$, $\phi = 1$ and $c(y,\phi) = -\log(y!)$. With $b'(\theta) = b''(\theta) = \exp(\theta) = \lambda$, we obtain

$$\mathrm{E}(y) = \mu = \lambda, \quad \mathrm{Var}(y) = \lambda,$$

i.e., the equality of expectation and variance that is characteristic for Poisson variables.

3. *Normal distribution:* The density of the normal distribution is

$$f(y\,|\,\mu) = \frac{1}{(2\pi\sigma^2)^{1/2}}\exp\left[-\frac{1}{2\sigma^2}(y-\mu)^2\right],$$

where $\mu = \mathrm{E}(y)$ is the parameter of interest and $\sigma^2 = \mathrm{Var}(y)$ is the nuisance parameter. The density can be written in the form of an exponential family

$$\begin{aligned} f(y\,|\,\mu) &= \exp\left[-\frac{y^2}{2\sigma^2} + \frac{y\mu}{\sigma^2} - \frac{\mu^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2)\right] \\ &= \exp\left[\frac{y\mu - \mu^2/2}{\sigma^2} - \frac{y^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2)\right] \end{aligned}$$

with $\theta = \mu$, $\phi = \sigma^2$, $b(\theta) = \mu^2/2 = \theta^2/2$ and $c(y,\phi) = -y^2/(2\sigma^2) - 0.5\log(2\pi\sigma^2)$. It follows, as expected, that

$$b'(\theta) = \theta = \mu = \mathrm{E}(y), \quad b''(\theta) = 1$$

and thus

$$\mathrm{Var}(y) = \phi b''(\theta) = \sigma^2.$$

$\triangle$

Similarly, one can derive the properties for the other distributions; see the summary in Table 5.12.

5.5 Generalized Linear Model

Distributional Assumptions

For given covariates $\boldsymbol{x}_i = (1, x_{i1}, \ldots, x_{ik})'$, the response variables are (conditionally) independent and the (conditional) density of y_i belongs to an exponential family with

$$f(y_i \mid \theta_i) = \exp\left(\frac{y_i\theta_i - b(\theta_i)}{\phi} w_i + c(y_i, \phi, w_i)\right).$$

The parameter θ_i is the natural parameter and ϕ is a common dispersion parameter, independent of i. For $\mathrm{E}(y_i) = \mu_i$ and $\mathrm{Var}(y_i)$, we have

$$\mathrm{E}(y_i) = \mu_i = b'(\theta_i), \quad \mathrm{Var}(y_i) = \sigma_i^2 = \phi\, b''(\theta_i)/w_i.$$

The weight parameter w_i is 1 for ungrouped data ($i = 1, \ldots, n$). In the case when the *sum* of the individual responses of group i is selected for the response variable y_i, the weight equals $1/n_i$ for grouped data ($i = 1, \ldots, G$). Note $w_i = n_i$ when the group mean, rather than the sum, is selected.

Structural Assumptions

The (conditional) mean μ_i is connected to the linear predictor $\eta_i = \boldsymbol{x}_i'\boldsymbol{\beta} = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik}$ through

$$\mu_i = h(\eta_i) = h(\boldsymbol{x}_i'\boldsymbol{\beta}) \quad \text{or} \quad \eta_i = g(\mu_i),$$

where

h is a (one-to-one and twice differentiable) *response function* and
g is the *link function*, i.e., the inverse $g = h^{-1}$.

In summary, a specific GLM is completely determined by the type of the exponential family (Gaussian, binomial, Poisson, gamma, inverse Gaussian), the choice of the link or response function, and the definition and selection of covariates.

The choice of an appropriate response or link function is, as presented in the preceding examples, dependent on the type of the response variable. Every exponential family has a unique *canonical* (or *natural*) link function, defined by $\theta_i = \eta_i = \boldsymbol{x}_i'\boldsymbol{\beta}$. According to Table 5.12, the linear model $\mu_i = \eta_i = \boldsymbol{x}_i'\boldsymbol{\beta}$ corresponds to the natural link function for the normal distribution, whereas the logit model is obtained in binary regression models, and the log-linear model results for Poisson models.

5.6 Maximum Likelihood Estimation in GLMs

Definition

The ML estimator $\hat{\boldsymbol{\beta}}$ maximizes the (log-)likelihood and is defined as the solution

$$s(\hat{\boldsymbol{\beta}}) = \mathbf{0}$$

of the score function given by

$$\boldsymbol{s}(\boldsymbol{\beta}) = \sum \boldsymbol{x}_i \frac{h'(\eta_i)}{\sigma_i^2}(y_i - \mu_i) = \boldsymbol{X}'\boldsymbol{D}\boldsymbol{\Sigma}^{-1}(\boldsymbol{y} - \boldsymbol{\mu}),$$

where $\boldsymbol{D} = \text{diag}(h'(\eta_1), \ldots, h'(\eta_n))$, $\boldsymbol{\Sigma} = \text{diag}(\sigma_1^2, \ldots, \sigma_n^2)$ and $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)'$ is the vector of expectations $\text{E}(y_i) = \mu_i = h(\eta_i)$.
The Fisher matrix is

$$\boldsymbol{F}(\boldsymbol{\beta}) = \sum \boldsymbol{x}_i \boldsymbol{x}_i' \tilde{w}_i = \boldsymbol{X}'\boldsymbol{W}\boldsymbol{X}$$

where $\boldsymbol{W} = \text{diag}(\tilde{w}_1, \ldots, \tilde{w}_n)$ is the diagonal matrix of working weights

$$\tilde{w}_i = \frac{(h'(\eta_i))^2}{\sigma_i^2}.$$

Numerical Computation

The ML estimator $\hat{\boldsymbol{\beta}}$ is obtained iteratively using Fisher scoring in form of iteratively weighted least squares estimates

$$\hat{\boldsymbol{\beta}}^{(t+1)} = (\boldsymbol{X}'\boldsymbol{W}^{(t)}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{W}^{(t)}\tilde{\boldsymbol{y}}^{(t)}, \quad t = 0, 1, 2, \ldots$$

with working weights and observations given in Eqs. (5.17) and (5.16).

For canonical link functions, the log-likelihood is always concave so that the ML estimator is always unique (if it exists). Moreover, it can be shown that the expected and observed information matrix coincide, i.e., $F(\boldsymbol{\beta}) = H(\boldsymbol{\beta})$.

5.4.2 Likelihood Inference

Inference in GLMs is again based on the likelihood principle. Let

$$\boldsymbol{X} = \begin{pmatrix} \boldsymbol{x}_1' \\ \vdots \\ \boldsymbol{x}_n' \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1k} \\ \vdots & & & \vdots \\ 1 & x_{n1} & \cdots & x_{nk} \end{pmatrix}$$

be the design matrix with $\mathrm{rk}(\boldsymbol{X}) = p$. In Sect. 5.8.2 we derive the log-likelihood $l(\boldsymbol{\beta})$, score function $\boldsymbol{s}(\boldsymbol{\beta})$, and Fisher information $\boldsymbol{F}(\boldsymbol{\beta})$; see Box 5.6 for a summary. Based on the score function and the Fisher information, Sect. 5.8.2 also shows that the ML estimator for $\boldsymbol{\beta}$ can be iteratively obtained as

$$\hat{\boldsymbol{\beta}}^{(t+1)} = \left(\boldsymbol{X}'\boldsymbol{W}^{(t)}\boldsymbol{X}\right)^{-1} \boldsymbol{X}'\boldsymbol{W}^{(t)}\tilde{\boldsymbol{y}}^{(t)}, \quad t = 0, 1, 2, \ldots. \tag{5.15}$$

Here, $\tilde{\boldsymbol{y}}^{(t)} = \left(\tilde{y}_1\left(\hat{\eta}_1^{(t)}\right), \ldots, \tilde{y}_n\left(\hat{\eta}_n^{(t)}\right)\right)'$ is a vector of "working observations" with elements

$$\tilde{y}_i\left(\hat{\eta}_i^{(t)}\right) = \hat{\eta}_i^{(t)} + \frac{\left(y_i - h\left(\hat{\eta}_i^{(t)}\right)\right)}{h'\left(\hat{\eta}_i^{(t)}\right)}, \tag{5.16}$$

where $\hat{\eta}_i^{(t)} = \boldsymbol{x}_i'\hat{\boldsymbol{\beta}}^{(t)}$ is the actual predictor, h is the response function, and $h'(\eta) = \partial h(\eta)/\partial\eta$ is the derivative of h with respect to η. The matrix

$$\boldsymbol{W}^{(t)} = \mathrm{diag}\left(\tilde{w}_1\left(\hat{\eta}_1^{(t)}\right), \ldots, \tilde{w}_n\left(\hat{\eta}_n^{(t)}\right)\right)$$

is a diagonal matrix of the "working weights"

$$\tilde{w}_i\left(\hat{\eta}_i^{(t)}\right) = \frac{\left(h'\left(\hat{\eta}_i^{(t)}\right)\right)^2}{\sigma_i^2\left(\hat{\eta}_i^{(t)}\right)}, \tag{5.17}$$

where $\sigma_i^2\left(\hat{\eta}_i^{(t)}\right)$ is the (conditional) variance $\mathrm{Var}(y_i)$ evaluated at $\eta = \hat{\eta}_i^{(t)}$. The required quantities to compute the weighted least squares estimator can be easily obtained from Table 5.12. A key role in the iterations Eq. (5.15) plays the Fisher matrix $\boldsymbol{F}(\boldsymbol{\beta}) = \boldsymbol{X}'\boldsymbol{W}\boldsymbol{X}$. Since the elements $\tilde{w}_i$ of the diagonal matrix $\boldsymbol{W}$ depend on the covariates $\boldsymbol{x}_i$ and on $\boldsymbol{\beta}$, invertibility of $\boldsymbol{F}(\boldsymbol{\beta})$ in Eq. (5.15) does not follow from the invertibility of $\boldsymbol{X}'\boldsymbol{X}$ (or equivalently the full rank of $\boldsymbol{X}$) in general. However, usually (almost) all of the weights are positive such that $\boldsymbol{F}(\boldsymbol{\beta})$ is invertible, which we assume in the following. Then, according to the stopping criterion, the algorithm

5.7 Asymptotic Properties of the ML Estimator

Let $\hat{\boldsymbol{\beta}}_n$ denote the ML estimator based on a sample of size n. Under regularity conditions, $\hat{\boldsymbol{\beta}}_n$ is consistent and asymptotically normal:

$$\hat{\boldsymbol{\beta}}_n \overset{a}{\sim} \mathrm{N}(\boldsymbol{\beta}, \boldsymbol{F}^{-1}(\boldsymbol{\beta})).$$

This result holds even if the estimator $\boldsymbol{F}(\hat{\boldsymbol{\beta}})$ replaces $\boldsymbol{F}(\boldsymbol{\beta})$.

typically converges close to a maximum after a number of iterative steps. With the natural link function, it can be shown that the achieved maximum is unique. However, this statement does not hold in general and therefore several different starting values should be used to help ensure the global maximum is achieved.

Asymptotic Properties of ML Estimates

As in section "Asymptotic Properties of the Least Squares Estimator" of Sect. 3.2.3 we index the model quantities with the number of observations n. For regressors with compact support, $(\boldsymbol{X}_n'\boldsymbol{X}_n)^{-1} \to \mathbf{0}$ or $\lambda_{\min}(\boldsymbol{X}_n'\boldsymbol{X}_n) \to \infty$ are sufficient for asymptotic normality and weak consistency in case of models with canonical link function (where $\lambda_{\min}$ denotes the smallest eigenvalue of $\boldsymbol{X}_n'\boldsymbol{X}_n$). Compare also section "Asymptotic Properties of the Least Squares Estimator" of Sect. 3.2.3 for a brief discussion and some examples of these conditions. For non-canonical link function, stronger conditions on the limiting behavior of $\boldsymbol{X}_n'\boldsymbol{X}_n$ have to be imposed. If, in the case of stochastic regressors, the observations $(y_i, \boldsymbol{x}_i)$ are independent and identically distributed, e.g., $(y, \boldsymbol{x})$, and comply with the assumptions of a general linear model, asymptotic normality follows under mild regularity conditions on the marginal distribution of $\boldsymbol{x}$.

Under the same assumptions, $\hat{\boldsymbol{\beta}}_n$ asymptotically exists with probability 1, i.e.,

$$\lim_{n\to\infty} \mathrm{P}(\hat{\boldsymbol{\beta}}_n \text{ exists}) = 1.$$

Details and general proofs can be found in Fahrmeir and Kaufmann (1985).

The inverse of the Fisher information matrix $\boldsymbol{F}(\boldsymbol{\beta})$, evaluated at the ML estimator $\hat{\boldsymbol{\beta}}$, is the asymptotic or approximate covariance matrix $\boldsymbol{A} = \boldsymbol{F}^{-1}(\hat{\boldsymbol{\beta}})$ of $\hat{\boldsymbol{\beta}}$. The diagonal element a_{jj} is an estimator for the variance $\sigma_j^2 = \mathrm{Var}(\hat{\beta}_j)$ of the jth component and $\sqrt{a_{jj}}$ for the standard deviation σ_j.

Estimation of the Scale or Overdispersion Parameter

Recall that $\mathrm{Var}(y_i) = \phi\, b''(\theta_i)/w_i$ for general exponential families. Denote by $v(\mu_i) = b''(\theta_i)$ the so-called variance function; see Table 5.12 for the specific expression of b''. Note that $b''(\theta_i)$ implicitly depends on μ_i through the relation $b'(\theta_i) = \mu_i$.

Using the variance function the dispersion parameter can then be estimated consistently by

$$\hat{\phi} = \frac{1}{G-p} \sum_{i=1}^{G} \frac{(y_i - \hat{\mu}_i)^2}{v(\hat{\mu}_i)/n_i},$$

where p denotes the number of regression parameters, $\hat{\mu}_i = h(\boldsymbol{x}_i' \hat{\boldsymbol{\beta}})$ is the estimated expectation, $v(\hat{\mu}_i)$ is the estimated variance function, and the data should be grouped as much as possible. We then substitute $\hat{\phi}$ for ϕ in every term containing ϕ, as, for example, in $\boldsymbol{F}(\hat{\boldsymbol{\beta}})$.

Testing Linear Hypotheses

For testing linear hypotheses

$$H_0 : \boldsymbol{C}\boldsymbol{\beta} = \boldsymbol{d} \qquad \text{versus} \qquad H_1 : \boldsymbol{C}\boldsymbol{\beta} \neq \boldsymbol{d},$$

where $\boldsymbol{C}$ has a full row rank $r \leq p$, we can use the likelihood ratio statistic lr, the Wald statistic w, and the score statistic u as discussed in more detail for binary responses in Sect. 5.1.3; see also Appendix B.4.4 (p. 662) for a general presentation of likelihood-based hypothesis testing. In the corresponding definitions, the specific formulae for the chosen GLM have to be used for $l(\boldsymbol{\beta})$, $s(\boldsymbol{\beta})$, and $F(\boldsymbol{\beta})$. Under conditions similar to those for the asymptotic results on ML estimation, we have $lr, w, s \overset{a}{\sim} \chi^2_r$, allowing for the computation of appropriate critical values and (approximate) p-values.

Criteria for Model Fit and Model Selection

The *Pearson statistic*

$$\chi^2 = \sum_{i=1}^{G} \frac{(y_i - \hat{\mu}_i)^2}{v(\hat{\mu}_i)/w_i}$$

and the *deviance*

$$D = -2 \sum_{i=1}^{G} \{l_i(\hat{\mu}_i) - l_i(\bar{y}_i)\}$$

are the two most common global statistics to verify the fit of a model relative to the saturated model. Here, $\hat{\mu}_i$ and $v(\hat{\mu}_i)$ are the estimated expectations and variance functions, respectively, and the ith log-likelihood contribution of the saturated model is $l_i(\bar{y}_i)$, where $\bar{y}_i$ replaces μ_i. This results in the maximum possible value of the log-likelihood. For both model fit statistics, the data should be grouped as much as possible. When n_i is sufficiently large in *all* groups $i = 1, \ldots, G$, both statistics are approximately or asymptotically (for $n \to \infty$) $\phi\chi^2(G-p)$-distributed, where p denotes the number of estimated parameters. In this situation, we can use both statistics for formal testing of model fit, i.e., for comparing the estimated model fit to that of the saturated model. For small n_i, especially for $n_i = 1$, such formal tests

5.8 Testing Linear Hypotheses

Hypotheses

$$H_0 : \boldsymbol{C}\boldsymbol{\beta} = \boldsymbol{d} \qquad \text{versus} \qquad H_1 : \boldsymbol{C}\boldsymbol{\beta} \neq \boldsymbol{d}\,.$$

Test Statistics

1. *Likelihood ratio statistic:* $lr = -2\{l(\tilde{\boldsymbol{\beta}}) - l(\hat{\boldsymbol{\beta}})\}$
2. *Wald statistic:* $w = (\boldsymbol{C}\hat{\boldsymbol{\beta}} - \boldsymbol{d})'[\boldsymbol{C}\boldsymbol{F}^{-1}(\hat{\boldsymbol{\beta}})\boldsymbol{C}']^{-1}(\boldsymbol{C}\hat{\boldsymbol{\beta}} - \boldsymbol{d})$
3. *Score statistic:* $u = \boldsymbol{s}'(\tilde{\boldsymbol{\beta}})\boldsymbol{F}^{-1}(\tilde{\boldsymbol{\beta}})\boldsymbol{s}(\tilde{\boldsymbol{\beta}})$
where $\tilde{\boldsymbol{\beta}}$ is the ML estimator under H_0.

Test Decision

For large n and under H_0, we have the asymptotic results

$$lr, w, u \overset{a}{\sim} \chi^2_r$$

where r is the (full) row rank of $\boldsymbol{C}$. We reject H_0 when

$$lq, w, u > \chi^2_r(1-\alpha).$$

can be problematic, even with a large sample size n. Large values of χ^2 or D then will not necessarily indicate a poor model fit.

The AIC for model selection is defined generally as

$$\text{AIC} = -2l(\hat{\boldsymbol{\beta}}) + 2p\,.$$

If the model contains a dispersion parameter ϕ, as is the case with the normal distribution, its maximum likelihood estimator should be substituted into the respective model. Accordingly, the total number of parameters must be increased to $p + 1$.

5.5 Quasi-likelihood Models

For GLMs, the response is assumed to be a member of the exponential family, e.g., the Gaussian, Poisson, or binomial distribution. This distributional assumption, in combination with the mean structure $\text{E}(y) = \mu = h(\boldsymbol{x}'\boldsymbol{\beta})$, implies a specific variance structure $\text{Var}(y) = \phi\, b''(\mu)/w$, where the variance function $v(\mu) = b''(\mu)$ is determined by the exponential family. If the empirical variance does not comply with the estimated variance $\hat{\phi}\, b''(\hat{\mu})/w$, the distribution of the data will be incorrectly specified, i.e., the data do not agree with the chosen distribution from the exponential family.

Quasi-likelihood models allow for a separate specification of the mean and the variance structure. Furthermore, it is not necessary that these specifications correspond to a proper likelihood function. It suffices to specify a correct expectation structure $\mathrm{E}(y) = h(\boldsymbol{x}'\boldsymbol{\beta})$, together with a "working" variance structure σ_i^2, and to define parameter estimates as the roots of a quasi-score function or *generalized estimating equation* (GEE) that has the same form as in usual GLMs; see the formula for $\boldsymbol{s}(\boldsymbol{\beta})$ in Box 5.6.

We then start directly with the specification of a *generalized estimating function*

$$\boldsymbol{s}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \boldsymbol{x}_i \frac{h'(\eta_i)}{\sigma_i^2}(y_i - \mu_i). \tag{5.18}$$

Similar as in the score function of Box 5.6 that was obtained as the derivative of the log-likelihood of a GLM, we assume that the expectation $\mu_i = h(\boldsymbol{x}_i'\boldsymbol{\beta})$ of y_i given $\boldsymbol{x}_i$ is correctly specified with

$$\mathrm{E}(y_i) = \mu_i = h(\boldsymbol{x}_i'\boldsymbol{\beta}).$$

We then have

$$\mathrm{E}(\boldsymbol{s}(\boldsymbol{\beta})) = \sum_{i=1}^{n} \boldsymbol{x}_i \frac{h'(\eta_i)}{\sigma_i^2}(\mathrm{E}(y_i) - \mu_i) = \mathbf{0},$$

as for a real score function, a property that is crucial for the consistency of parameter estimates.

In contrast, it is not necessary that the specified variance σ_i^2 in Eq. (5.18) equals the true variance $\mathrm{Var}(y_i)$, but it can rather be specified with the help of a given quasi-variance function $v(\mu)$, i.e., $\sigma^2(\mu) = \phi\, v(\mu)/w$. We then call $\sigma^2(\mu) = \phi\, v(\mu)/w$ the *working variance*.

The simplest form of a (working) variance function results from overdispersion in binomial and Poisson models with $w_i = n_i$ and

$$\sigma_i^2(\pi_i) = \phi \frac{\pi_i(1-\pi_i)}{n_i}$$

or

$$\sigma_i^2(\lambda_i) = \phi\lambda_i,$$

respectively. In this case, the quasi-score function (5.18) is identical to the score function of a binomial or Poisson model up to a constant factor $1/\phi$, but it no longer corresponds to the derivative of a log-likelihood function.

The (working) variance function is often parameterized with another parameter θ as

$$\sigma^2(\mu) = \phi v(\mu; \theta).$$

An important special case is

$$v(\mu;\theta) = \mu^{\theta},$$

where we obtain the variance function of the Gaussian, Poisson, gamma, and of the inverse Gaussian distribution, for $\theta = 0, 1, 2, 3$, respectively.

The quasi-ML estimator $\hat{\boldsymbol{\beta}}$ is defined as the root of the quasi-score function, i.e., as the solution to the *generalized estimating equation* (GEE)

$$\boldsymbol{s}(\hat{\boldsymbol{\beta}}) = \boldsymbol{0}.$$

As in case of ML estimation, the solution is obtained iteratively. The quasi-Fisher information matrix $\boldsymbol{F}(\boldsymbol{\beta}) = \mathrm{E}(-\partial \boldsymbol{s}(\boldsymbol{\beta})/\partial \boldsymbol{\beta}')$ becomes

$$\boldsymbol{F}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i' \tilde{w}_i,$$

with working variance σ_i^2 included in the working weights $\tilde{w}_i = (h'(\eta_i))^2/\sigma_i^2$. However, $\boldsymbol{F}(\boldsymbol{\beta})$ differs from $\boldsymbol{V}(\boldsymbol{\beta}) = \mathrm{Cov}(\boldsymbol{s}(\boldsymbol{\beta})) = \mathrm{E}(\boldsymbol{s}(\boldsymbol{\beta})\boldsymbol{s}'(\boldsymbol{\beta}))$ in general. In fact, we have

$$\boldsymbol{V}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i' \tilde{w}_i \frac{\sigma_{0i}^2}{\sigma_i^2}.$$

Thus only in the case when the working variances equal the true variances σ_{0i}^2 we obtain $\boldsymbol{F}(\boldsymbol{\beta}) = \boldsymbol{V}(\boldsymbol{\beta})$ as in ML estimation.

Under regularity assumptions, quasi-ML estimators are consistent and asymptotically normal

$$\boldsymbol{\beta} \overset{a}{\sim} \mathrm{N}\left(\boldsymbol{\beta}, \hat{\boldsymbol{F}}^{-1} \hat{\boldsymbol{V}} \hat{\boldsymbol{F}}^{-1}\right)$$

with estimates $\hat{\boldsymbol{F}} = \boldsymbol{F}(\hat{\boldsymbol{\beta}})$ and

$$\hat{\boldsymbol{V}} = \sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i' (h'(\hat{\eta}_i))^2 \frac{(y_i - \hat{\mu}_i)^2}{\sigma_i^4(\hat{\boldsymbol{\beta}})}$$

for $\boldsymbol{F}(\boldsymbol{\beta})$ and $\boldsymbol{V}(\boldsymbol{\beta})$. Compared with the asymptotic properties of the ML estimator, only the asymptotic covariance matrix $\mathrm{Cov}(\hat{\boldsymbol{\beta}}) = \hat{\boldsymbol{F}}^{-1}$ has to be corrected to the "sandwich" matrix $\hat{\boldsymbol{A}} = \hat{\boldsymbol{F}}^{-1} \hat{\boldsymbol{V}} \hat{\boldsymbol{F}}^{-1}$. Thus, quasi-likelihood models allow consistent and asymptotically normal estimation of $\boldsymbol{\beta}$ but with some loss of (asymptotic) efficiency. To keep this loss minimal, the working variance structure should be not far off the true variance structure.

5.6 Bayesian Generalized Linear Models

The Bayesian approach for linear models discussed in Sect. 4.4 can, in principle, be applied to GLMs. However, the application is more complicated both mathematically and numerically. Fully Bayesian inference usually requires the use of MCMC simulation techniques that are more complex than the corresponding techniques for linear models. This section gives a brief overview of Bayesian inference in GLMs. We limit the discussion to models without dispersion parameters, specifically focusing on binomial and Poisson models. A more complete discussion of Bayesian GLMs and extensions can be found in Dey, Gosh, and Mallick (2000), as well as in the corresponding sections in Fahrmeir and Tutz (2001).

In Bayesian GLMs, we assume a prior density $p(\boldsymbol{\beta})$ for the parameter vector $\boldsymbol{\beta}$ which is considered a random variable. Similar to Bayesian linear models discussed in Sect. 4.4 we assume a multivariate Gaussian prior, i.e.,

$$\boldsymbol{\beta} \sim \mathrm{N}(\boldsymbol{m}, \boldsymbol{M}), \tag{5.19}$$

where $\boldsymbol{m}$ is the prior mean vector and $\boldsymbol{M}$ the prior covariance matrix. A typical choice is $\boldsymbol{m} = \mathbf{0}$ and $\boldsymbol{M} = \boldsymbol{I}$ thereby assuming independence among the regression coefficients. A noninformative prior is obtained by $\boldsymbol{m} = \mathbf{0}$ and the limit $\boldsymbol{M}^{-1} = \mathbf{0}$. Other choices such as a combination of informative and noninformative priors, Bayesian ridge and LASSO or spike and slab priors, that have been discussed extensively for Bayesian linear models, can be used as well. We restrict the discussion here to the normal prior (5.19) because the only difficulty compared to linear models is inference regarding the regression coefficients. Inference for the hyperparameters is typically based on identical MCMC updating steps as for linear models. The reason is that their full conditionals are independent of the specific observation model. For instance, the Bayesian LASSO assumes

$$\beta_1, \ldots, \beta_k \mid \tau_1^2, \ldots, \tau_k^2 \sim \mathrm{N}(\mathbf{0}, \mathrm{diag}(\tau_1^2, \ldots, \tau_k^2)).$$

While updating the regression coefficients in the Bayesian LASSO might be problematic because the full conditional is not Gaussian (see below), updating the variances τ_j^2 proceeds exactly as described in Sect. 4.4.2.

We now discuss the difficulties involved with Bayesian inference for non-Gaussian data. According to Bayes' theorem, inference relies on the posterior density $p(\boldsymbol{\beta} \mid \boldsymbol{y})$, given the (conditionally independent) response variables $\boldsymbol{y} = (y_1, \ldots, y_n)'$ and covariates. Suppressing the notational dependence on covariates, this yields

$$p(\boldsymbol{\beta} \mid \boldsymbol{y}) = \frac{L(\boldsymbol{\beta} \mid \boldsymbol{y})\, p(\boldsymbol{\beta})}{\int L(\boldsymbol{\beta} \mid \boldsymbol{y})\, p(\boldsymbol{\beta})\, d\boldsymbol{\beta}} \propto L(\boldsymbol{\beta} \mid \boldsymbol{y})\, p(\boldsymbol{\beta}), \tag{5.20}$$

where

$$L(\boldsymbol{\beta} \mid \boldsymbol{y}) = \prod_{i=1}^{n} f_i(y_i \mid \boldsymbol{\beta})$$

is the likelihood of a given GLM, for example, a binomial logit model or a log-linear Poisson model. The posterior mean is defined as

$$\mathrm{E}(\boldsymbol{\beta} \mid \boldsymbol{y}) = \int \boldsymbol{\beta}\, p(\boldsymbol{\beta} \mid \boldsymbol{y}) d\boldsymbol{\beta}$$

and the corresponding posterior covariance matrix

$$\text{Cov}(\boldsymbol{\beta}\,|\,\boldsymbol{y}) = \int (\boldsymbol{\beta} - \text{E}(\boldsymbol{\beta}\,|\,\boldsymbol{y}))(\boldsymbol{\beta} - \text{E}(\boldsymbol{\beta}\,|\,\boldsymbol{y}))'\, p(\boldsymbol{\beta}\,|\,\boldsymbol{y})\, d\boldsymbol{\beta}$$

provides a measure for the precision of the posterior mean. At first glance, it seems straightforward to put Bayesian inference into effect. However, the integrations involved are problematic, as their analytical solution is only possible in a few special cases. Numerical integration methods are applicable, as long as the dimension of $\boldsymbol{\beta}$ remains relatively small (about $\leq$ 5); extensions to more complex models are described in the following chapters, yet remain widely intractable. Hence, we have two options: First, posteriori mode or MAP (maximum a posteriori) estimation, for which we have to maximize the numerator in Eq. (5.20), or second, fully Bayesian inference with MCMC techniques.

5.6.1 Posterior Mode Estimation

The *posterior mode* $\hat{\boldsymbol{\beta}}_{MAP}$ maximizes the posterior density $p(\boldsymbol{\beta}\,|\,\boldsymbol{y})$ or the log-posterior

$$\log(p(\boldsymbol{\beta}\,|\,\boldsymbol{y})) = l(\boldsymbol{\beta}) + \log\, p(\boldsymbol{\beta})\,,$$

where $l(\boldsymbol{\beta})$ is the log-likelihood of the given GLM. For a Gaussian prior

$$\boldsymbol{\beta} \sim \text{N}(\boldsymbol{m}, \boldsymbol{M}), \qquad \boldsymbol{M} \text{ positive definite,}$$

we obtain the special case

$$\log(p(\boldsymbol{\beta}\,|\,\boldsymbol{y})) = l(\boldsymbol{\beta}) - \frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{m})'\boldsymbol{M}^{-1}(\boldsymbol{\beta} - \boldsymbol{m})\,,$$

where terms independent of $\boldsymbol{\beta}$ have been left out. Now $\log(p(\boldsymbol{\beta}\,|\,\boldsymbol{y}))$ can also be viewed as a *penalized (log-)likelihood*, where the penalty term $(\boldsymbol{\beta} - \boldsymbol{m})'\boldsymbol{M}^{-1}(\boldsymbol{\beta} - \boldsymbol{m})$ penalizes large deviations from the prior mean $\boldsymbol{m}$. This penalization potentially overcomes the problem of non-existence or divergence of the ML estimators. We also refer to the estimator $\hat{\boldsymbol{\beta}}_{MAP}$ as a penalized ML estimator.

For the limiting case $\boldsymbol{M}^{-1} \to \boldsymbol{0}$ of a flat prior

$$p(\boldsymbol{\beta}) \propto const,$$

the penalty disappears, which results in the posterior mode estimator equaling the (unpenalized) ML estimator.

The *ridge estimator* with *shrinkage* parameter $\lambda = 1/(2\tau^2)$ results as a special case with $\boldsymbol{m} = \boldsymbol{0}$ and $\boldsymbol{M} = \tau^2\text{diag}(0, 1, \ldots, 1)$; see also section "Bayesian Ridge Regression" of Sect. 4.4.2. The penalty then simplifies to

$$\lambda\, \boldsymbol{\beta}'\text{diag}(0, 1, \ldots, 1)\boldsymbol{\beta} = \lambda(\beta_1^2 + \beta_2^2 + \ldots + \beta_k^2)\,,$$

and the parameter λ regularizes shrinkage of the ML estimator $\hat{\boldsymbol{\beta}}$ towards $\mathbf{0}$ and therefore stabilizes the ML estimator in cases of large variability.

Estimation of the posterior mode proceeds analogously to ML estimation. With a Gaussian prior distribution, the score function $\boldsymbol{s}(\boldsymbol{\beta})$ becomes the penalized score function

$$\boldsymbol{s}_p(\boldsymbol{\beta}) = \frac{\partial \log(p(\boldsymbol{\beta} \mid \boldsymbol{y}))}{\partial \boldsymbol{\beta}} = \boldsymbol{s}(\boldsymbol{\beta}) - \boldsymbol{M}^{-1}(\boldsymbol{\beta} - \boldsymbol{m})$$

and the Fisher information matrix $\boldsymbol{F}(\boldsymbol{\beta})$ becomes

$$\boldsymbol{F}_p(\boldsymbol{\beta}) = -\mathrm{E}\left(-\frac{\partial^2 \log(p(\boldsymbol{\beta} \mid \boldsymbol{y}))}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'}\right) = \boldsymbol{F}(\boldsymbol{\beta}) + \boldsymbol{M}^{-1}.$$

Computation is carried out with a modified Fisher scoring algorithm or IWLS algorithm, in which $\boldsymbol{s}_p(\boldsymbol{\beta})$ and $\boldsymbol{F}_p(\boldsymbol{\beta})$ replace $\boldsymbol{s}(\boldsymbol{\beta})$ and $\boldsymbol{F}(\boldsymbol{\beta})$, respectively.

Under regularity assumptions, for $n \to \infty$, $\hat{\boldsymbol{\beta}}_{MAP}$ has an asymptotic (or approximate) normal distribution with

$$\hat{\boldsymbol{\beta}}_{MAP} \overset{a}{\sim} \mathrm{N}\left(\boldsymbol{\beta}, \boldsymbol{F}_p^{-1}(\hat{\boldsymbol{\beta}}_{MAP})\right),$$

and, as a consequence, the posterior mode $\hat{\boldsymbol{\beta}}_{MAP}$ and the (expected) curvature $\boldsymbol{F}_p^{-1}(\hat{\boldsymbol{\beta}}_{MAP})$ are good approximations of the posterior mean $\mathrm{E}(\boldsymbol{\beta} \mid \boldsymbol{y})$ and of the posterior covariance matrix $\mathrm{Cov}(\boldsymbol{\beta} \mid \boldsymbol{y})$, respectively.

5.6.2 Fully Bayesian Inference via MCMC Simulation Techniques

Fully Bayesian inference relies on MCMC techniques (see Appendix B.5) for drawing random numbers from the posterior $p(\boldsymbol{\beta} \mid \boldsymbol{y})$. Posterior means, medians, quantiles, variances, etc. are then approximated with their empirical analogues. For the Gaussian prior $\boldsymbol{\beta} \sim \mathrm{N}(\boldsymbol{m}, \boldsymbol{M})$ and also for the limiting case $\boldsymbol{M}^{-1} \to \mathbf{0}$ of a non-informative prior $p(\boldsymbol{\beta}) \propto const$, we have

$$p(\boldsymbol{\beta} \mid \boldsymbol{y}) \propto \exp\left(l(\boldsymbol{\beta}) - \frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{m})'\boldsymbol{M}^{-1}(\boldsymbol{\beta} - \boldsymbol{m})\right).$$

With the exception of some special cases, no closed analytical form exists for the normalizing constant of this posterior. We therefore use MCMC techniques for drawing samples $\boldsymbol{\beta}^{(t)}, t = 1, \ldots, T$, from $p(\boldsymbol{\beta} \mid \boldsymbol{y})$. Dellaportas and Smith (1993) recommend a Gibbs sampler based on adaptive rejection sampling, which is implemented in the software WinBUGS. We prefer to draw the entire parameter vector $\boldsymbol{\beta}^{(t)}$, in every iteration step t, with a Metropolis–Hastings (MH) algorithm based on IWLS proposals; see Gamerman (1997) and Lenk and DeSarbo (2000). IWLS proposals $q(\boldsymbol{\beta}^* \mid \boldsymbol{\beta}^{(t)})$, for the update $\boldsymbol{\beta}^{(t+1)}$, rely on a normal distribution

5.9 Bayesian GLMs

Posterior Distribution

$$p(\boldsymbol{\beta} \mid \boldsymbol{y}) = \frac{L(\boldsymbol{\beta} \mid \boldsymbol{y})p(\boldsymbol{\beta})}{\int L(\boldsymbol{\beta} \mid \boldsymbol{y})p(\boldsymbol{\beta})d\boldsymbol{\beta}} \propto L(\boldsymbol{\beta} \mid \boldsymbol{y})p(\boldsymbol{\beta}),$$

where $L(\boldsymbol{\beta} \mid \boldsymbol{y})$ is the likelihood of a GLM and $p(\boldsymbol{\beta})$ is the prior distribution.

Posterior Mode

The *posterior mode* $\hat{\boldsymbol{\beta}}_{MAP}$ maximizes the posterior density $p(\boldsymbol{\beta} \mid \boldsymbol{y})$. With a normal prior distribution

$$\boldsymbol{\beta} \sim \mathrm{N}(\boldsymbol{m}, \boldsymbol{M}),$$

this is equivalent to maximizing the *penalized log-likelihood*

$$\log(p(\boldsymbol{\beta} \mid \boldsymbol{y})) = l(\boldsymbol{\beta}) - \frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{m})'\boldsymbol{M}^{-1}(\boldsymbol{\beta} - \boldsymbol{m}),$$

with $l(\boldsymbol{\beta}) = \log L(\boldsymbol{\beta} \mid \boldsymbol{y})$. The iterative calculation of $\hat{\boldsymbol{\beta}}_{MAP}$ via IWLS relies on the penalized score function and Fisher information matrix

$$\boldsymbol{s}_p(\boldsymbol{\beta}) = \boldsymbol{s}(\boldsymbol{\beta}) - \boldsymbol{M}^{-1}(\boldsymbol{\beta} - \boldsymbol{m}) \qquad \boldsymbol{F}_p(\boldsymbol{\beta}) = \boldsymbol{F}(\boldsymbol{\beta}) + \boldsymbol{M}^{-1}.$$

Fully Bayesian Inference

Fully Bayesian inference is accomplished using an MH algorithm with IWLS proposals for drawing random numbers from the posterior density $p(\boldsymbol{\beta} \mid \boldsymbol{y})$.

Let $\boldsymbol{\beta}^{(t)}$ be the actual state of the Markov chain. We then draw the IWLS proposal $\boldsymbol{\beta}^*$ from a normal distribution density $q(\boldsymbol{\beta}^*|\boldsymbol{\beta}^{(t)})$ with

$$\boldsymbol{\beta}^* \sim \mathrm{N}(\boldsymbol{\mu}^{(t)}, (\boldsymbol{X}'\boldsymbol{W}^{(t)}\boldsymbol{X} + \boldsymbol{M}^{-1})^{-1}).$$

The Fisher matrix $\boldsymbol{F}_p^{(t)} = \boldsymbol{X}'\boldsymbol{W}^{(t)}\boldsymbol{X} + \boldsymbol{M}^{-1}$ is evaluated at the current state $\boldsymbol{\beta}^{(t)}$ and

$$\boldsymbol{\mu}^{(t)} = (\boldsymbol{F}_p^{(t)})^{-1}(\boldsymbol{X}'\boldsymbol{W}^{(t)}\tilde{\boldsymbol{y}}^{(t)} + \boldsymbol{M}^{-1}\boldsymbol{m}),$$

with $\boldsymbol{W}^{(t)} = \boldsymbol{W}(\boldsymbol{\beta}^{(t)})$ and the current working observations $\tilde{\boldsymbol{y}}^{(t)}$ (defined in the same way as for ML estimation). The probability of acceptance is then given by

$$\alpha(\boldsymbol{\beta}^* \mid \boldsymbol{\beta}^{(t)}) = \min\left\{\frac{L(\boldsymbol{\beta}^*)\, p(\boldsymbol{\beta}^*)\, q(\boldsymbol{\beta}^{(t)} \mid \boldsymbol{\beta}^*)}{L(\boldsymbol{\beta}^{(t)})\, p(\boldsymbol{\beta}^{(t)})\, q(\boldsymbol{\beta}^* \mid \boldsymbol{\beta}^{(t)})}\right\},$$

with the likelihood $L(\boldsymbol{\beta})$ of the GLM evaluated at the proposed and current value, $\boldsymbol{\beta}^*$ and $\boldsymbol{\beta}^{(t)}$, respectively.

Table 5.13 Number of citations from patents: Bayesian Poisson model

Variable	Coefficient	Standard deviation	2.5 % Quantile	97.5 % Quantile
intercept	0.156	0.031	0.090	0.218
yearc	−0.072	0.003	−0.077	−0.066
ncountryc	−0.028	0.004	−0.036	−0.020
nclaimc	0.018	0.001	0.015	0.021
biopharm	0.240	0.032	0.180	0.301
ustwin	0.003	0.025	−0.043	0.054
patus	−0.078	0.029	−0.133	−0.019
patgsgr	−0.199	0.032	−0.262	−0.138
opp	0.372	0.025	0.321	0.422

having expectation and covariance matrix that are a (first) approximation of the posterior mode estimator and of the respective covariance matrix. Refer to Box 5.9 for details.

Example 5.9 Number of Citations from Patents—Bayesian Inference

We illustrate Bayesian inference through reanalyzing the log-linear Poisson model of Example 5.7 (p. 297) with a flat prior $p(\boldsymbol{\beta}) \propto const$ for $\boldsymbol{\beta}$. With this choice, the posterior mean obtained from a fully Bayesian model specification and the ML estimator, which is identical to the posterior mode, should not differ too much from each other. Table 5.13 contains the posterior mean estimates, as well as the posterior standard deviations and quantiles, for the Bayesian Poisson model with purely linear effects. Table 5.14 contains the corresponding results obtained with nonlinear effects for the continuous covariates. The results are based on `bayesreg objects` of the software `BayesX`. We find good agreement with our previous results based on ML inference. Note, however, that the existing overdispersion has not been taken into account so that the standard deviations are below the standard errors of Table 5.11 (p. 299).

The fact that the results of ML and Bayes inference differ only slightly from each other in this example provokes the following question: What is the advantage of the comparably computer intensive Bayesian estimator based on MCMC methods? One advantage is that, in addition to point estimates and confidence intervals, we are also able to estimate the entire posterior density $p\left(\boldsymbol{\beta} \mid \boldsymbol{y}\right)$ based on the sampled random numbers. Figure 5.4 shows kernel density estimates for the posterior densities $p\left(\beta_j \mid \boldsymbol{y}\right)$ of the covariate effects for *ustwin*, *patus*, and *patgsgr*, as well as corresponding normal densities with adjusted expectations and variances. The posterior densities are all close to normality, which should be expected due to the comparably large sample size in this example. In general, Bayesian inference with MCMC is especially important for more complex regression models, if asymptotic normality approximations of likelihood inference are not reliable.

△

5.6.3 MCMC-Based Inference Using Data Augmentation

For a number of response distributions alternative sampling schemes, based on the representation of the models as latent linear models, can be developed. For binary response models, the connection to latent linear models has been pointed out in Sect. 5.1 on p. 274.

Table 5.14 Number of citations from patents: extended Bayesian Poisson model with nonlinear effects

Variable	Coefficient	Standard deviation	2.5 % Quantile	97.5 % Quantile
intercept	0.17022	0.03225	0.10759	0.23645
yearc	−0.09897	0.00556	−0.10975	−0.08807
yearc^2	−0.00973	0.00131	−0.01242	−0.00723
yearc^3	−0.00011	0.00017	−0.00046	0.00022
ncountryc	0.01581	0.00804	−0.00061	0.03318
ncountryc^2	−0.00215	0.00124	−0.00456	0.00025
ncountryc^3	−0.00158	0.00026	−0.00211	−0.00109
nclaimc	0.02609	0.00208	0.02205	0.03016
nclaimc^2	−0.00047	0.00021	−0.00085	−0.00006
nclaimc^3	0.00000	0.00000	−0.00000	−0.00001
biopharm	0.15549	0.03254	0.09443	0.21985
ustwin	−0.00294	0.02458	−0.05460	0.04415
patus	−0.09475	0.02757	−0.14652	−0.04188
patgsgr	−0.20191	0.03207	−0.26613	−0.13890
opp	0.37127	0.02506	0.32115	0.42044

We illustrate the alternative sampling approach for binary probit models. Conditional on the covariates and the parameters, y_i follows a Bernoulli distribution $y_i \sim \mathrm{B}(1, \pi_i)$ with conditional mean $\pi_i = \Phi(\eta_i)$ where Φ is the cumulative distribution function of a standard normal distribution. On p. 274, the probit model was equivalently defined using latent variables

$$\tilde{y}_i = \boldsymbol{x}_i'\boldsymbol{\beta} + \varepsilon_i = \eta_i + \varepsilon_i$$

with normally distributed errors $\varepsilon_i \sim \mathrm{N}(0, 1)$. The connection between the binary responses and the latent variables is $y_i = 1$ if $\tilde{y}_i > 0$, and $y_i = 0$ if $\tilde{y}_i \leq 0$.

The idea is to use the latent variable representation rather than the original formulation for parameter estimation. This approach was first introduced in a paper by Albert and Chib (1993). The main idea is to consider the latent variables as additional parameters in the model, and to base posterior inference on the extended parameter space. Correspondingly, additional sampling steps for updating the $\tilde{y}_i$'s are required. Fortunately, sampling the $\tilde{y}_i$'s is relatively easy and fast because the full conditionals are truncated normal distributions (see Definition B.5 in Appendix B.1). More specifically, $\tilde{y}_i \mid \cdot \sim \mathrm{TN}_{0,\infty}(\eta_i, 1)$ if $y_i = 1$ and $\tilde{y}_i \mid \cdot \sim \mathrm{TN}_{-\infty,0}(\eta_i, 1)$ if $y_i = 0$. Efficient algorithms for drawing random numbers from a truncated normal distribution can be found in Geweke (1991) or Robert (1995) and are implemented in many major statistics packages. The advantage of defining a probit model through the latent variables $\tilde{y}_i$ is that the full conditionals for the regression coefficients $\boldsymbol{\beta}$ are Gaussian with covariance matrix and mean given by

$$\boldsymbol{\Sigma}_{\beta} = \left(\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{M}^{-1}\right)^{-1}, \qquad \boldsymbol{\mu}_{\beta} = \boldsymbol{\Sigma}_{\beta}\left(\boldsymbol{X}'\tilde{\boldsymbol{y}} + \boldsymbol{M}^{-1}\boldsymbol{m}\right). \tag{5.21}$$

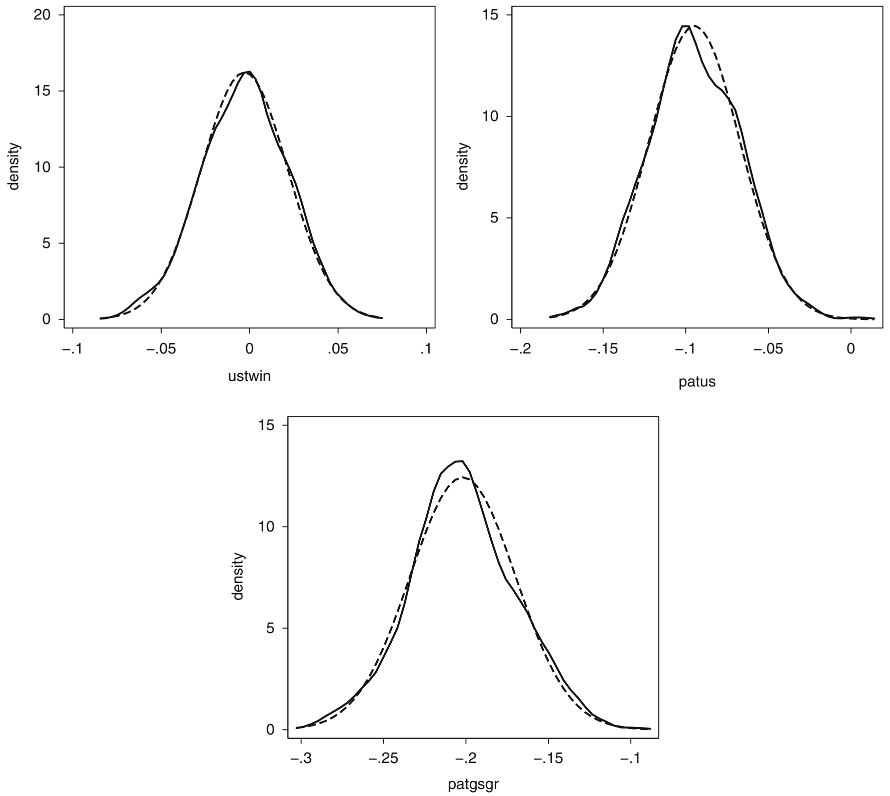

Fig. 5.4 Number of citations from patents: estimated posteriori densities for the effects of *ustwin*, *patus*, and *patgsgr* (*solid line*) together with a normal approximation (*dashed line*)

Hence, we can avoid costly MH steps as is the case with IWLS proposals. Instead, we can resort to the simple Gibbs sampler that was developed for Gaussian responses with slight modifications. Updating of $\boldsymbol{\beta}$ can be done exactly as described in Sect. 4.4.1 (p. 234) using the current values $\tilde{y}_i$ of the latent variables as (pseudo) responses. Another distinct advantage of the Gibbs step is that it works even for high-dimensional parameter vectors, whereas the MH steps with IWLS proposals may break down because acceptance rates typically go down as the parameter dimension increases. The price we pay for the simplicity is the additional update step to draw the latent variables which may be time-consuming in large samples. Then the MH algorithm with IWLS proposals may be faster.

Summarizing, we obtain the following Gibbs sampler:

1. Define initial values $\tilde{\boldsymbol{y}}^{(0)}$ and $\boldsymbol{\beta}^{(0)}$. Set $t = 1$.
2. Sample $\tilde{\boldsymbol{y}}^{(t)}$ by drawing $\tilde{y}_i^{(t)}$, $i = 1, \dots, n$, from $\mathrm{TN}_{0,\infty}(\eta_i^{(t-1)}, 1)$ if $y_i = 1$ and $\tilde{y}_i \mid \cdot \sim \mathrm{TN}_{-\infty,0}(\eta_i^{(t-1)}, 1)$ if $y_i = 0$.

3. Sample $\boldsymbol{\beta}^{(t)}$ by drawing from the Gaussian full conditional with covariance matrix and mean given in Eq. (5.21) thereby replacing $\tilde{\boldsymbol{y}}$ by the actual state $\tilde{\boldsymbol{y}}^{(t)}$.
4. Stop if $t = T$, otherwise set $t = t + 1$ and go to 2.

We finally note that the data augmentation trick is not limited to binary probit models. Similar algorithms have been developed, e.g., for binary logit models, multi-categorical logit or probit models as outlined in Chap. 6, and Poisson regression. References to the literature are given in Sect. 5.8.

5.7 Boosting Generalized Linear Models

In Sect. 4.3, we introduced a versatile method for obtaining regularized estimates in linear regression with the particular advantage of implicit variable selection (boosting). In fact, the approach can be immediately transferred to the context of GLMs with rather minor modifications. When considering the generic algorithm in Box 4.4 (p. 226), the basic ingredients of a boosting algorithm are:

- The specification of a lack-of-fit criterion via a loss function
- The specification of base learning procedures

A suitable loss function in GLMs is given by the negative log-likelihood such that

$$\rho(\boldsymbol{\eta}) = -l(\boldsymbol{\eta}) = -\sum_{i=1}^{n} \frac{y_i\theta_i - b(\theta_i)}{\phi} w_i .$$

The negative gradients are then still given by

$$u_i = -\frac{\partial}{\partial\eta}\rho(y_i, \eta)\big|_{\eta=\hat{\eta}_i^{(t-1)}}$$

and are, for GLMs, computed as

$$u_i = \frac{h'(\eta_i)^{(t-1)}}{(\sigma_i^2)^{(t-1)}}\left(y_i - \mu_i^{(t-1)}\right).$$

In contrast, no modifications are required for the base learning procedures, and we can still rely on least squares fits applied to the working responses u_i. In summary, boosting can immediately be adapted to generalized response types by providing a suitable loss function. While the negative log-likelihood is a natural choice, different loss functions can in principle be considered. For example, in case of binary regression, the exponential loss

$$\rho(\boldsymbol{\eta}) = \sum_{i=1}^{n} \exp(-y_i\eta_i)$$

(with $y_i \in \{-1, 1\}$ instead of $y_i \in \{0, 1\}$) is sometimes used as an alternative popular in the classification literature; see Friedman, Hastie, and Tibshirani (2000) for details.

5.8 Bibliographic Notes and Proofs

5.8.1 Bibliographic Notes

Nelder and Wedderburn (1972) introduced GLMs as a general class of models for response variables with densities belonging to the exponential family of distributions. Linear, logit, probit, and Poisson models could therefore be subsumed under one conceptual umbrella, leading to important new stimulations for statistical model building, methodological developments, and applications. The book McCullagh and Nelder (1989) gives a detailed outline of GLMs; a more compact introduction can be found in Fahrmeir and Tutz (2001, Chap. 2). Collett (1991) and Tutz (2011, Chaps. 2–5) provide a detailed exposition of binary regression models. These books also give a good overview of methods for model diagnosis that are based on residuals, developed analogously to the linear model of Chap. 3. In the econometrics literature GLMs are usually treated within the field of "microeconometrics." Standard textbooks on microeconometrics are Cameron and Trivedi (2005) and Winkelmann (2010a). Kleiber and Zeileis (2008) discuss econometrics models including GLMs in the software package R.

Several aspects motivated modifications and additions to basic GLMs. For example, the response distribution may be difficult to model with univariate exponential families (as assumed in Sect. 5.4) in some applications. This especially applies to the following regression situation:

- *Regression Models for Count Data:* Although Poisson regression, as illustrated in Sect. 5.2, is the standard model for count data regression, the Poisson distribution is often too simplistic in applications. Cameron and Trivedi (1998) and Winkelmann (2010b) describe enhanced regression modeling for count data. An overview of available count data models in the software package R is given in Zeileis, Kleiber, and Jackman (2008).
- *Life Time (Survival) and Duration Time Models:* Life times, duration times, and waiting times up to a certain event appear in many areas of application. Statistical analyses are then often complicated by incomplete data due to censoring, e.g., when life spans are not terminated until the end of a study period. The Cox model is the most popular regression model for (censored) life times and is closely related to Poisson regression. Standard textbooks on survival and duration time models are Collett (2003), Klein and Moeschberger (2005), Hosmer, Lemeshow, and May (2008), and Therneau and Grambsch (2000).
- *GLMs for Location, Scale, and Shape:* For continuous response variables, we can model the effect of covariates not only on the mean but also on the variance, skewness, or kurtosis; see Sect. 2.9.1 for a brief introduction and Rigby and Stasinopoulos (2005) for more details.
- *Multivariate Response Variables:* If the response $\boldsymbol{y} = (y_1, \dots, y_c)$ consists of several scalar responses $y_1, \dots, y_c$, this yields *multivariate regression*. Anderson (2003) gives an introduction to multivariate linear regression as an extension of the linear regression model. Components $y_1, \dots, y_c$ that do not have a

normal distribution face the difficulty of finding an appropriate joint distribution. *Copula* concepts offer appropriate possibilities (Joe, 1997) especially for continuous components. Other approaches are quasi-likelihood or marginal models (Fahrmeir and Tutz, 2001, Chap. 3), and models with latent variables (Skrondal and Rabe-Hesketh, 2004).

One of the most important extensions of GLMs is the inclusion of nonparametric and semiparametric approaches that allow for flexible modeling of nonlinear covariate effects. The resulting model class, e.g., generalized additive models (GAM), has already been introduced in Chap. 2 and will be discussed in more detail in Chaps. 8 and 9.

In their original definition, GLMs are especially suited for the regression analysis of cross-sectional data. Mixed models (Chap. 7) are a popular tool for the analysis of *clustered* or *longitudinal data*. Depending on the goals of a longitudinal study, *autoregressive* (or *conditional*) models including temporally lagged values of the response variable as additional covariates, or *marginal models* based on quasi-likelihood approaches, can be a reasonable alternative for the analysis; see Diggle, Heagerty, Liang, and Zeger (2002), Fahrmeir and Tutz (2001, Chap. 6), and the additional comments in Sect. 7.8.

In the early 1990s, *Bayesian GLMs* and corresponding extensions have seen a fast development parallel to the spread of MCMC simulation techniques; see Dey et al. (2000) and corresponding sections in the following chapters. IWLS proposals for updating the regression coefficients are due to Gamerman (1997); see also Lenk and DeSarbo (2000) for a slightly modified approach. Estimating Bayesian GLMs using data augmentation similar as described for probit models works for a variety of response distributions; see Holmes and Held (2006) and Frühwirth-Schnatter and Frühwirth (2010) for logit models, Frühwirth-Schnatter and Wagner (2006) and Frühwirth-Schnatter, Frühwirth, Held, and Rue (2009) for Poisson and gamma regression.

GLMs with errors in variables have been developed for data situations where covariates cannot be observed exactly, but only subject to measurement errors. For details on models of this type, we refer to Carroll, Ruppert, Stefanski, and Crainiceanu (2006).

5.8.2 Proofs

Derivation of the ML Estimator in GLMs (Sect. 5.4.2)

The ML estimator in GLMs is derived with the following steps:

1. Log-likelihood

The log-likelihood contribution of an observation $(y_i, \boldsymbol{x}_i)$ (up to an additive constant) is given by

$$l_i(\boldsymbol{\beta}) = \log(f(y_i \mid \boldsymbol{\beta})) = \frac{y_i\theta_i - b(\theta_i)}{\phi} w_i. \tag{5.22}$$

Thereby, the log-likelihood depends on the regression parameters $\boldsymbol{\beta}$ through the natural parameter θ_i of the exponential family via

$$\mu_i = b'(\theta_i) = h(\boldsymbol{x}_i'\boldsymbol{\beta}).$$

Due to the (conditional) independence of y_i,

$$l(\boldsymbol{\beta}) = \sum l_i(\boldsymbol{\beta})$$

is the complete log-likelihood of the sample. To treat individual data ($i = 1, \ldots, n$) and grouped data ($i = 1, \ldots, G$) simultaneously, we omit n or G from the upper limit of the summation signs.

2. Score function

The score function $\boldsymbol{s}(\boldsymbol{\beta}) = \partial l(\boldsymbol{\beta})/\partial\boldsymbol{\beta}$ is obtained by applying the chain rule to the individual score function contributions:

$$\boldsymbol{s}_i(\boldsymbol{\beta}) = \partial l_i(\boldsymbol{\beta})/\partial\boldsymbol{\beta} = \frac{\partial\eta_i}{\partial\boldsymbol{\beta}}\frac{\partial\mu_i}{\partial\eta_i}\frac{\partial\theta_i}{\partial\mu_i}\frac{\partial(y_i\theta_i - b(\theta_i))}{\partial\theta_i}\frac{w_i}{\phi}.$$

The first contribution is simply given by

$$\frac{\partial\eta_i}{\partial\boldsymbol{\beta}} = \boldsymbol{x}_i.$$

The second contribution

$$\frac{\partial\mu_i}{\partial\eta_i} = \frac{\partial h(\eta_i)}{\partial\eta_i} = h'(\eta_i)$$

depends on the response function h and is therefore specific to a given model. In the following we use the shortcut:

$$d_i = h'(\eta_i).$$

The third term is obtained by reversing the nominator and denominator, which yields

$$\frac{\partial\mu_i}{\partial\theta_i} = \frac{\partial b'(\theta_i)}{\partial\theta_i} = b''(\theta_i) = \frac{w_i\,\mathrm{Var}(y_i)}{\phi} = \frac{w_i\sigma_i^2}{\phi}$$

and therefore

$$\frac{\partial\theta_i}{\partial\mu_i} = \frac{\phi}{w_i\sigma_i^2}.$$

Finally, we have

$$\frac{\partial(y_i\theta_i - b(\theta_i))}{\partial\theta_i} = y_i - b'(\theta_i) = y_i - \mu_i.$$

Putting these pieces together yields the score function as

$$\boldsymbol{s}(\boldsymbol{\beta}) = \sum \boldsymbol{x}_i d_i \frac{\phi}{w_i \sigma_i^2}(y_i - \mu_i)\frac{w_i}{\phi} = \sum \boldsymbol{x}_i \frac{d_i}{\sigma_i^2}(y_i - \mu_i)$$

From $\mathrm{E}(y_i) = \mu_i$, it follows that $\mathrm{E}(\boldsymbol{s}(\boldsymbol{\beta})) = \boldsymbol{0}$ holds.

To express the score function more compactly in matrix notation we define the vectors

$$\boldsymbol{y} = (y_1, \dots, y_n)', \quad \boldsymbol{\mu} = (\mu_1, \dots, \mu_n)',$$

and the diagonal matrices

$$\boldsymbol{D} = \mathrm{diag}(d_1, \dots, d_n), \quad \boldsymbol{\Sigma} = \mathrm{diag}(\sigma_1^2, \dots, \sigma_n^2).$$

Then we obtain

$$\boldsymbol{s}(\boldsymbol{\beta}) = \boldsymbol{X}'\boldsymbol{D}\boldsymbol{\Sigma}^{-1}(\boldsymbol{y} - \boldsymbol{\mu})$$

where $\boldsymbol{D}$ and $\boldsymbol{\Sigma}$ are both dependent on $\boldsymbol{\beta}$.

3. Information matrix

To derive the Fisher matrix $\boldsymbol{F}(\boldsymbol{\beta}) = \mathrm{E}(\boldsymbol{s}(\boldsymbol{\beta})\boldsymbol{s}'(\boldsymbol{\beta}))$, we note that

$$\boldsymbol{F}(\boldsymbol{\beta}) = \sum \mathrm{E}(\boldsymbol{s}_i(\boldsymbol{\beta})\boldsymbol{s}_i'(\boldsymbol{\beta})).$$

We obtain

$$\begin{aligned}
\mathrm{E}\left(\boldsymbol{s}_i(\boldsymbol{\beta})\boldsymbol{s}_i'(\boldsymbol{\beta})\right) &= \mathrm{E}\left(\boldsymbol{x}_i\boldsymbol{x}_i' \frac{d_i^2}{(\sigma_i^2)^2}(y_i - \mu_i)^2\right) \\
&= \boldsymbol{x}_i\boldsymbol{x}_i' \frac{d_i^2}{(\sigma_i^2)^2}\mathrm{E}(y_i - \mu_i)^2 \\
&= \boldsymbol{x}_i\boldsymbol{x}_i' \frac{d_i^2}{(\sigma_i^2)^2}\mathrm{Var}(y_i) \\
&= \boldsymbol{x}_i\boldsymbol{x}_i' \frac{d_i^2}{\sigma_i^2}.
\end{aligned}$$

This yields

$$\boldsymbol{F}(\boldsymbol{\beta}) = \sum \boldsymbol{x}_i\boldsymbol{x}_i'\tilde{w}_i, \tag{5.23}$$

with the "working weights"

$$\tilde{w}_i = \frac{d_i^2}{\sigma_i^2} = \left(h'(\eta_i)\right)^2 \frac{w_i}{b''(\theta_i)\phi}$$

also depending on $\boldsymbol{\beta}$. In matrix notation the Fisher matrix can be written as

$$\boldsymbol{F}(\boldsymbol{\beta}) = \boldsymbol{X}'\boldsymbol{W}\boldsymbol{X}$$

with the diagonal matrix $\boldsymbol{W} = \mathrm{diag}(\ldots, \tilde{w}_i, \ldots)$ of working weights. Note that $\boldsymbol{W} = \boldsymbol{D}^2 \boldsymbol{\Sigma}^{-1}$.

4. Numerical computation of the ML estimator
Computation of the ML estimator $\hat{\boldsymbol{\beta}}$ is usually based on the Fisher scoring algorithm

$$\hat{\boldsymbol{\beta}}^{(t+1)} = \hat{\boldsymbol{\beta}}^{(t)} + \boldsymbol{F}^{-1}(\hat{\boldsymbol{\beta}}^{(t)})\boldsymbol{s}(\hat{\boldsymbol{\beta}}^{(t)}), \quad t = 0, 1, 2, \ldots.$$

Inserting the formulae for $\boldsymbol{s}(\boldsymbol{\beta})$ and $\boldsymbol{F}(\boldsymbol{\beta})$ we obtain

$$\begin{aligned}
\hat{\boldsymbol{\beta}}^{(t+1)} &= \hat{\boldsymbol{\beta}}^{(t)} + (\boldsymbol{X}'\boldsymbol{W}(\hat{\boldsymbol{\beta}}^{(t)})\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{D}(\hat{\boldsymbol{\beta}}^{(t)})\boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}}^{(t)})^{-1}(\boldsymbol{y} - \boldsymbol{\mu}(\hat{\boldsymbol{\beta}}^{(t)})) \\
&= (\boldsymbol{X}'\boldsymbol{W}(\hat{\boldsymbol{\beta}}^{(t)})\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{W}(\hat{\boldsymbol{\beta}}^{(t)})\boldsymbol{X}\hat{\boldsymbol{\beta}}^{(t)} \\
&\quad + (\boldsymbol{X}'\boldsymbol{W}(\hat{\boldsymbol{\beta}}^{(t)})\boldsymbol{X})^{-1}\boldsymbol{X}\boldsymbol{W}(\hat{\boldsymbol{\beta}}^{(t)})'\boldsymbol{D}(\hat{\boldsymbol{\beta}}^{(t)})^{-1}(\boldsymbol{y} - \boldsymbol{\mu}(\hat{\boldsymbol{\beta}}^{(t)})) \\
&= (\boldsymbol{X}'\boldsymbol{W}(\hat{\boldsymbol{\beta}}^{(t)})\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{W}(\hat{\boldsymbol{\beta}}^{(t)})\left[\boldsymbol{\eta}(\hat{\boldsymbol{\beta}}^{(t)}) + \boldsymbol{D}(\hat{\boldsymbol{\beta}}^{(t)})^{-1}(\boldsymbol{y} - \boldsymbol{\mu}(\hat{\boldsymbol{\beta}}^{(t)}))\right].
\end{aligned}$$

Hence the iterations can be expressed as an *iteratively weighted least squares estimator*

$$\hat{\boldsymbol{\beta}}^{(t+1)} = (\boldsymbol{X}'\boldsymbol{W}^{(t)}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{W}^{(t)}\tilde{\boldsymbol{y}}^{(t)}, \quad t = 0, 1, 2, \ldots$$

where $\tilde{\boldsymbol{y}}^{(t)} = (\ldots, \tilde{y}_i(\hat{\boldsymbol{\beta}}^{(t)}), \ldots)'$ is a "working response vector" with elements

$$\tilde{y}_i(\hat{\boldsymbol{\beta}}^{(t)}) = \boldsymbol{x}_i'\hat{\boldsymbol{\beta}}^{(t)} + d_i^{-1}(\hat{\boldsymbol{\beta}}^{(t)})(y_i - \hat{\mu}_i(\hat{\boldsymbol{\beta}}^{(t)})),$$

and $\boldsymbol{W}^{(t)}$ is the weight matrix, evaluated at $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}^{(t)}$. Replacing in $\tilde{w}_i$ and $\tilde{y}_i$ the d_i by $h'(\boldsymbol{x}_i'\hat{\boldsymbol{\beta}}^{(t)})$ and writing the expressions in terms of $\hat{\eta}_i^{(t)} = \boldsymbol{x}_i'\hat{\boldsymbol{\beta}}^{(t)}$ rather than $\hat{\boldsymbol{\beta}}^{(t)}$ we obtain the formulae (5.17) and (5.16) as stated in Sect. 5.4.2.