Chapter 3

Generalized Linear Models

In this chapter we embed the logistic regression model as well as the classical regression model into the framework of generalized linear models. Generalized linear models (GLMs), which have been proposed by Nelder and Wedderburn (1972), may be seen as a framework for handling several response distributions, some categorical and some continuous, in a unified way. Many of the binary response models considered in later chapters can be seen as generalized linear models, and the same holds for part of the count data models in Chapter 7.

The chapter may be read as a general introduction to generalized linear models; continuous response models are treated as well as categorical response models. Therefore, parts of the chapter can be skipped if the reader is interested in categorical data only. Basic concepts like the deviance are introduced in a general form, but specific forms that are needed in categorical data analysis will also be given in the chapters where the models are considered. Nevertheless, the GLM is useful as a background model for categorical data modelling, and since McCullagh and Nelder's (1983) book everybody working with regression models should be familiar with the basic concept.

3.1 Basic Structure

A generalized linear model is composed from several components. The random component specifies the distribution of the conditional response \( y \) given \( x \), whereas the systematic component specifies the link between the expected response and the covariates.

(1) Random component and distributional assumptions

Given \( x \), the \( y \)’s are (conditionally) independent observations from a simple exponential family. This family has a probability density function or mass function of the form

\[
 f(y_i | \beta, \phi_i) = \exp \left\{ \frac{y_i \beta_i - h(\phi_i)}{\phi_i} + c(y_i, \phi_i) \right\},
\]

where

- \( \beta_i \) is the natural parameter of the family,
- \( \phi_i \) is a scale or dispersion parameter, and
- \( h(\cdot) \) and \( c(\cdot) \) are specific functions corresponding to the type of the family.

As will be outlined later, several distributions like the binomial, normal, or Poisson distribution are members of the simple exponential family.
(2) Systematic component

The systematic component is determined by two structuring components, the linear term and the link between the response and the covariates. The linear part that gives the GLM is the one that specifies that the variables \( x_i \) enter the model in linear form by forming the linear predictor

\[
\eta_i = x_i^T \beta,
\]

where \( \beta \) is an unknown parameter vector of dimension \( p \). The relation between the linear part and the conditional expectation \( \mu_i = E(y_i|\mathbf{x}_i) \) is determined by the transformation

\[
\mu_i = h(\eta_i) = h(x_i^T \beta),
\]

or, equivalently,

\[
\eta_i = g(\mu_i) = x_i^T \beta,
\]

where

- \( h \) is a known one-to-one response function,
- \( g \) is the so-called link function, that is, the inverse of \( h \).

Equations (3.2) and (3.3) reflect equivalent ways to specify how the mean of the response variable is linked to the linear predictor. The response function \( h \) in (3.2) shows how the linear predictor has to be transformed to determine the expected mean. Equation (3.3) shows for which transformation of the mean the model becomes linear. A simple example is the logistic model, where the mean \( \mu_i \) corresponds to the probability of success \( \pi_i \). In this case one has the two forms

\[
\pi_i = \frac{\exp(x_i^T \beta)}{1 + \exp(x_i^T \beta)},
\]

yielding the response function \( h(\eta_i) = \exp(\eta_i)/(1 + \exp(\eta_i)) \), and

\[
\log\left( \frac{\pi_i}{1 - \pi_i} \right) = x_i^T \beta,
\]

where the link function \( g = h^{-1} \) is specified by the logit transformation \( g(\pi) = \log(\pi/(1 - \pi)) \).

Based on the latter form, which corresponds to (3.3), it is seen that a GLM is a linear model for the transformed mean where additionally it is assumed that the response has a distribution in the simple exponential family. A specific generalized linear model is determined by

- the type of the exponential family that specifies the distribution of \( y_i|x_i \),
- the form of the linear predictor, that is, the selection and coding of covariates;
- the response or link function.

Before considering the various models that fit into this framework, let us make some remarks on simple exponential families: In simple exponential families the natural parameter is linked to the mean of the distribution. Thus the parameter \( \theta_i \) may be seen as \( \theta_i = \theta(\mu_i) \), where \( \theta \) is considered as a transformation of the mean. Parameterization of specific distributions must often use different names and also different sets of parameters; for example, \( \lambda_i \) is often used in the case of the Poisson distribution and the exponential distribution. These parameters determine uniquely the mean \( \mu_i \) and therefore the natural parameter \( \theta_i \).

3.2 Generalized Linear Models for Continuous Responses

3.2.1 Normal Linear Regression

The normal linear regression model is usually given with an error term in the form

\[ y_i = x_i^T \beta + \epsilon_i \]

with normal error, \( \epsilon_i \sim N(0, \sigma^2) \). Alternatively, the model may be specified in GLM terminology by

\[ y_i|x_i \sim N(\mu_i, \sigma^2) \quad \text{and} \quad \mu_i = y_i = x_i^T \beta. \]

The form separates the distribution from the systematic component. While \( y_i|x_i \sim N(\mu_i, \sigma^2) \) assumes that the response is normal with the variance not depending on the observation, the link between the mean and the predictor is provided by assuming \( \mu_i = y_i = x_i^T \beta \). Thus, the classical linear model uses the identity as a link function. It is easily seen that the normal distribution is within the exponential family by considering

\[
f(y) = \exp\left\{ \frac{1}{2} \frac{y - \mu}{\sigma}^2 - \log(\sqrt{2\pi} \sigma) \right\} = \exp\left\{ \frac{y - \mu^2/2}{\sigma^2} - \frac{y^2}{2\sigma^2} - \log(\sqrt{2\pi} \sigma) \right\}.
\]

Therefore, the natural parameter and the function \( b \) are given by

\[
\theta(\mu) = \mu, \quad b(\theta) = \theta^2/2 = \mu^2/2, \quad \phi = \sigma^2.
\]

The separation of random and systematic components makes it easy to allow for alternative links between the mean and the predictors. For example, if the response is income or reaction time, the responses are expected to be positive. Then, a more appropriate link that at least ensures that means are positive is

\[
\mu = \exp(\eta) = \exp(x_i^T \beta).
\]

Of course, the influence of the covariates and consequently the interpretation of the parameters differ from those in the classical linear model. In contrast to the linear model,

\[
\mu = x_i^T \beta = x_i \beta_1 + \ldots + x_i \beta_p,
\]

where the change of \( x_i \) by one unit means an additive effect of \( \beta_i \) on the expectation. The modified link

\[
\nu = \exp(x_i \beta_1 + \ldots + x_i \beta_p) = e^{x_i \beta_1} \ldots e^{x_i \beta_p}
\]

specifies that the change of \( x_i \) by one unit has a multiplicative effect on \( \nu \) by the factor \( e^{x_i \beta} \), since \( e^{x_i \beta_1 + \ldots + x_i \beta_p} = e^{x_i \beta_1} \ldots e^{x_i \beta_p} \). In Figure 3.1 the normal regression model is illustrated for one explanatory variable. The left picture shows the linear model and the right picture the log-link model. The straight line and the curve show the means as functions of \( x \); the densities of the response are shown only at three distinct \( x \)-values.

3.2.2 Exponential Distribution

In cases where responses are strictly non-negative, for example, in the analysis of duration time or survival, the normal distribution model is rarely adequate. A classical distribution that is often used when time is the response variable is the exponential distribution

\[
f(y) = \lambda e^{-\lambda y} = \exp(-\lambda y + \log(\lambda)), \quad y \geq 0.
\]
3.2 GENERALIZED LINEAR MODELS FOR CONTINUOUS RESPONSES

where \( \nu = \nu' \mu \). When using the expectation as a parameter, as we did in the specification of the density, we will write \( \Gamma(\nu, \frac{\mu}{\nu}) \).

The variance of the Gamma-distribution is given by \( \nu \sigma^2 = \nu \mu^2 / \nu' \). Thus the variance depends strongly on the expectation, an effect that is often found in practice. The dependence may be characterized by the coefficient of variation \( \nu = \sigma / \mu \), a specific measure of variation that scales the standard deviation by the expectation. For Gamma-distributions, the coefficient of variation for the i-th observation is given by \( \nu_i = \mu_i^2 / (\nu \mu_i) = 1 / \sqrt{\nu} \). Since it does not depend on the observation, one may set \( \nu = \nu_i / \mu_i \). Therefore, the assumption of a Gamma-distribution implies that the coefficient of variation is held constant across observations. It is implicitly assumed that large means are linked to large variances. This is in contrast to the assumption that is often used for normal distributions, when variances are assumed to be constant over observations.

3.2.3 Gamma-Distributed Responses

Since the exponential distribution is a one-parameter distribution, its flexibility is rather restricted. A more flexible distribution model for non-negative responses like duration or insurance claims is the Gamma distribution. With \( \nu > 0 \) denoting the expected value and \( \nu' > 0 \) the shape parameter, the Gamma-distribution has the form

\[
\Gamma(y) = \frac{1}{\Gamma(\nu)(\nu')^{y}} y^{\nu-1} \exp \left( -\frac{\nu}{\nu'} y \right) \exp \left( -\frac{\nu}{\nu'} \log(\nu) + \nu' \log(y) - \nu' \log(\nu') \right).
\]

In exponential family parameterization one obtains the dispersion parameter \( \phi = 1/\nu \) and \( \eta(\mu) = -1/\nu \log(\nu) = -\log(\phi) \). In contrast to the exponential distribution, the dispersion parameter is not fixed. While it is \( \phi = 1 \) for the exponential distribution, it is an additional parameter in the Gamma-distribution. As is seen from Figure 3.3, the parameter \( \nu \) is a shape parameter. For \( 0 < \nu < 1 \), \( f(y) \) decreases monotonically, whereas for \( \nu > 1 \) the density has a mode at \( y = \mu - \mu' y \) and is positively skewed. Usually the Gamma-distribution is abbreviated by \( \Gamma(\nu, \sigma) \).

![Figure 3.2: Gamma distributions for several \( \mu, \nu \).](image)

The canonical link for the Gamma distribution is the same as for the exponential distribution. Figure 3.3 shows the exponential and the Gamma regression models for the log-link function. We can see how the shifting of the mean along the logarithmic function changes the form of the distribution. In contrast to the normal model, where densities are simply shifted, for Gamma-distributed responses, the form of the densities depends on the mean. Moreover, Figure 3.3 shows that densities are positive only for positive \( x \)-values. For the normal model shown in Figure 3.1 the log-link ensures that the mean is positive, but nevertheless the model also allows negative values. Thus, for a strictly positive-valued response the normal model is often not a good choice, but, of course, the adequacy of the model depends on the values of \( x \) that are modeled and the variance of the response.

3.2.4 Inverse Gaussian Distribution

An alternative distribution with a strictly non-negative response, which can be used to model responses like duration, is the inverse Gaussian distribution. In its usual form it is given by the density

\[
f(y) = \left( \frac{1}{2\pi y^3} \right)^{1/2} \exp \left( -\frac{1}{2\nu y} (y - \mu)^2 \right), \quad y > 0,\]

where \( \nu = \nu' \mu \). When using the expectation as a parameter, as we did in the specification of the density we will write \( \Gamma(\nu, \frac{\mu}{\nu}) \).

![Figure 3.1: Normal regression with identity link (left) and with log-link (right).](image)
with abbreviation \( IG(\mu, \lambda) \), where \( \mu, \lambda > 0 \) are the determining parameters. Straightforward derivation yields

\[
f(\mu) = \exp \left( \frac{\gamma(1/2\sigma^2)}{1/\lambda} + 1/\mu - \frac{\lambda}{2} \log(2\pi) - \frac{\lambda}{2} \log(\mu) \right),
\]

and therefore

\[
\theta = \frac{1}{2\mu^2}, \quad b(\theta) = -\mu = -\sqrt{-2\theta}, \quad \phi = 1/\lambda,
\]

\[
c(\mu, \phi) = -\frac{1}{2} \log(2\pi \phi^2) - \frac{3}{2} \log(\mu).
\]

The canonical link function, for which \( \theta(\mu) = 1 \) holds, is given by

\[
\mu(\theta) = -\frac{3}{2\theta^2} \quad \text{or} \quad b(\mu) = -\frac{1}{\sqrt{2\theta}},
\]

which implies the severe restriction \( \eta = \theta^2/\beta > 0 \). A link function without these problems is the log-link function \( g(\mu) = \log(\mu) \) and thus \( \theta(\mu) = \exp(g(\mu)) \).

The inverse Gaussian distribution has several interesting properties, including that the ML estimate of the mean \( \mu \) and the dispersion \( 1/\lambda \), given by

\[
\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} y_i, \quad \frac{1}{\hat{\lambda}} = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{y_i} - \frac{1}{\theta},
\]

are independent. This is similar to the normal distribution, for which the sample mean and the sample variance are independent. Based on the independence, Tweedie (1957) suggested an analog of the analysis of variance for censored data (see also Folks and Chhikara, 1978).

### 3.3 GLMs for Discrete Responses

#### 3.3.1 Models for Binary Data

The simplest case of a discrete response is when only "success" or "failure" is measured with the outcome \( y \in \{0, 1\} \). The Bernoulli distribution has for \( y \in \{0, 1\} \) the probability mass function

\[
f(y) = x^y(1-x)^{1-y} = \exp \left( y \log \left( \frac{x}{1-x} \right) + \log(1-x) \right).
\]

where \( x = \frac{P(y=1)}{1} \) is the probability for "success." With \( \mu = x \) it is an exponential family with

\[
\theta(\mu) = \log(\mu/(1-\mu)), \quad b(\theta) = \log(1+\exp(\theta)) = -\log(1-x), \quad \phi = 1.
\]

The classical link that yields the logit model is

\[
x = \frac{\exp(\eta)}{1 + \exp(\eta)}, \quad g(\mu) = \log \left( \frac{x}{1-x} \right)
\]

(see Chapter 2). Alternatively, any strictly monotone distribution function \( F \) like the normal distribution or extreme value distributions may be used as a response function, yielding \( \tau = F(\eta/\beta) \), with the response and link functions given by \( h(\eta) = F'(\eta), g(\tau) = F^{-1}(\tau) \).

#### 3.3.2 Models for Binomial Data

If experiments that distinguish only between "success" and "failure" are repeated independently, if it natural to consider the number of successes or the proportion as the response variable. For \( m \) trials one obtains the binomially distributed response \( y \in \{0, \ldots, m\} \). The binomial distribution has the parameters \( m \) and the probability \( x \) of success in one trial. For \( y \in \{0, \ldots, m\} \) it has the form

\[
f(y) = \frac{m!}{y!(m-y)!} x^y (1-x)^{m-y} \exp \left( \frac{y}{m} \log \left( \frac{x}{1-x} \right) + \log(1-x) \right) + \log(m)
\]

By considering the proportion of successes \( y = \frac{y}{m} \) instead of the number of successes \( y \), one obtains an exponential family with the same specifications as for binary responses: \( \mu = \frac{\hat{y}}{m} = \pi, \theta(\mu) = \log(\pi/(1-\pi)), \) and \( b(\theta) = (1+\exp(\theta))/\phi - \log(1-\pi) \). Only the dispersion parameter is different, given as \( \phi = 1/m \). The distribution of \( y \) has the usual binomial form

\[
f(y) = \frac{m!}{y!(m-y)!} \pi^y (1-\pi)^{m-y} = \frac{m!}{y!(m-y)!} \pi^y (1-\pi)^{m-y} \exp(\frac{y}{m} \log(\pi/(1-\pi)), \)
\]

but with values \( y \in \{0, 1/m, \ldots, 1\} \). Because the support is different from the usual binomial distribution, it is called the scaled binomial distribution. It consists of a simple rescaling of the number of successes to proportions and therefore changes the support.
For the binomial distribution the specification of the dispersion parameter differs from that for the other distributions considered here. With indices one has for observation $y_i = y_i/m_i$, the dispersion parameter $\phi_i = 1/m_i$, where $m_i$ is the number of replications. Because $m_i$ is fixed, the dispersion is fixed (and known) but may depend on the observations since the number of replications may vary across observations. In contrast to the other distributions, the dispersion depends on $i$.

An alternative way of looking at binomial data is by considering them as grouped observations, that is, grouping of replications (see Section 3.5). For the special case $m = 1$ there is no difference between the binomial and the rescaled binomial distributions. Of course, the binary case may be treated as a special case of the binomial case. Consequently, the link and response functions are treated in the same way as in the binary case.

### 3.3.3 Poisson Model for Count Data

Discrete responses often take the form of counts, for example, the number of insurance claims or case numbers in epidemiology. Contingency tables may be seen as counts that occur as entries in the cells of the table. A simple distribution for count data is the Poisson distribution, which for integer values $y \in \{0, 1, \ldots\}$ and parameter $\lambda > 0$ has the form

$$f(y) = \frac{\lambda^y e^{-\lambda}}{y!} \propto \exp\{y \log(\lambda) - \lambda - \log(y!))\}.$$ 

With expectation $\mu = \lambda$ the parameters of the exponential family are given by $\theta(\mu) = \log(\mu)$, $h(\theta) = \exp(\theta) = e^\theta \phi = 1$. A sensible choice of the link function should account for the restriction $\lambda > 0$. Thus a widely used link function is the log-link yielding

$$\log(\lambda) = \mu(0) = \mu$$ or $$\lambda = \exp(\mu).$$ 

The distribution is shown for three distinct $\lambda$-values in Figure 3.5. It is seen that differing mean imply different shapes of the distribution. While the distribution of the response is skewed for low means, it is nearly symmetric for large values of the mean.

![Figure 3.5: Poisson regression with log-link.](image)

### 3.3.4 Negative Binomial Distribution

An alternative distribution for count data is the negative binomial distribution, which has mass function

$$f(y) = \frac{\Gamma(y + \nu)}{\Gamma(y + 1)\Gamma(\nu)} \left(\frac{\nu}{\mu + \nu}\right)^y \left(\frac{\mu}{\mu + \nu}\right)^\nu \quad y = 0, 1, \ldots,$$ 

where $\nu, \mu > 0$ are parameters. We will use the abbreviation NB$(\nu, \mu)$. The distribution may be motivated in several ways. It may be seen as a mixture of Poisson distributions in the so-called Gamma-Poisson model. The model assumes that the parameter $\lambda$ of the Poisson distribution is itself a random variable that is Gamma-distributed with $\lambda \sim \Gamma(\nu, \frac{\mu}{\nu})$ with shape parameter $\nu$ and expectation $\mu$. Given $\lambda$, it is assumed that $y$ is Poisson-distributed, $y | \lambda \sim P(\lambda)$. Then the marginal distribution of $y$ is given by (3.5). Since it is often more appropriate to assume that the total counts result from heterogeneous sources with individual parameters, the negative binomial model is an attractive alternative to the Poisson model. From the variance of the Gamma-distribution $\nu^2/\nu$, it is seen that for $\nu \to \infty$ the mixture of Poisson distributions shrinks to just one Poisson distribution and one obtains the Poisson model as the limiting case. The expectation and variance of the negative binomial are given by

$$E(y) = \mu, \quad \text{var}(y) = \mu + \nu \mu^2/\nu.$$ 

Thus, for $\nu \to \infty$, one obtains $E(y) = \mu$, which is in accordance with the Poisson distribution. The parameter $\nu$ may be seen as an additional dispersion parameter that yields a larger variation for small values. Thus it is more appropriate to consider $1/\nu$ as an indicator for the amount of variation.

For integer-valued $\nu$ the negative binomial is also considered in the form

$$f(y) = \frac{(\nu + y - 1)^y}{\nu^y y!} \quad y = 0, 1, \ldots,$$ 

where $\nu, \mu \in (0, 1)$ may be seen as an alternative parameter with a simple interpretation. If independent Bernoulli variables with probability of occurrence $\nu$ are considered, then the negative binomial distribution (3.6) reflects the probability for the number of trials that are necessary in addition to $\nu$ to obtain $\nu$ hits. The most familiar case is $\nu = 1$, where one considers the number of trials (plus one) that are necessary until the first hit occurs. The corresponding geometric distribution is a standard distribution, for example, in fertility studies where the number of trials until conception is modeled.

Within the exponential family framework one obtains with $\pi = \nu/(\nu + \nu)$ from (3.5)

$$f(y) = \exp\left\{ \log(y) + (y/\nu) \log(1 - \nu)/\nu + \log\left(\frac{\Gamma(y + 1)}{\Gamma(y + 1)\Gamma(\nu)}\right) \right\}.$$ 

For fixed $\nu$ one has a simple exponential family for the scaled response $y = y/\nu$ and the dispersion $\phi = 1/\nu$. Since $y/\nu$ is considered as the response, one has expectation $\mu = E[y] = \mu/\nu$ and therefore $\theta(\mu) = \log(1 - \nu) = \log(\nu)/(\mu + 1)$ and $h(\theta) = -\log(1 - \exp(\theta))$. The canonical link model that fulfills $\theta(\mu) = \eta$ is given by

$$\log\left(\frac{\mu}{\mu + 1}\right) = \eta \quad \text{or} \quad \mu = \frac{\exp(\eta)}{1 - \exp(\eta)}.$$
The canonical link may cause problems because, for \( n \to 0 \), one has \( \mu \to \infty \). For the log-link, \( \log(\mu) = \eta = 0 \); or \( \mu = \exp(\eta) \); however, the predictor \( \eta \) is not restricted.

The negative binomial response \( y = 0 \) is scaled by the specified parameter \( \nu \). Thus, when treated within the framework of GLMs, the parameter has to be fixed in advance.

3.4 Further Concepts

3.4.1 Means and Variances

The distribution of the responses is assumed to be in the exponential family \( f(y; \theta, \phi) = \exp\left(\{y \theta - \eta(\theta)\} + \phi \psi(\eta(\theta))\right) \). In the previous sections examples have been given for the dependence of the natural parameters \( \lambda_{i} \) on \( \mu_{i} \) and the parameters that characterize the distribution. For example, for the Bernoulli distribution one obtains \( \lambda_{i} = \theta(\mu_{i}) \) in the form \( \lambda_{i} = \log(\mu_{i}/(1 - \mu_{i})) \), and since \( \mu_{i} = \pi_{i} \), one has \( \lambda_{i} = \log(\pi_{i}/(1 - \pi_{i})) \).

In general, in the exponential families the mean is directly related to the function \( \eta(\lambda_{i}) \) in the form

\[
\mu_{i} = \theta^{-1} = \theta^{-1}(\lambda_{i}) = \theta^{-1}(\theta(\mu_{i}))
\]

and for the variances one obtains

\[
\sigma_{i}^{2} = \text{var}(y_{i}) - \phi \psi(\eta(\lambda_{i})) = \phi \psi(\eta(\lambda_{i}))(\theta(\mu_{i})).
\] (3.7)

Thus the variances are composed from the dispersion parameter \( \phi \) and the so-called variance function \( \psi^{2}(\lambda_{i}) \). As is seen from (3.7) and (3.8) in GLMs there is a strict link between the mean \( \mu_{i} \) and the variance since both are based on derivatives of \( \eta(\theta) \). Because \( \lambda_{i} \) depends on the mean through the functional form \( \lambda_{i} = \theta(\mu_{i}) \), the variance function is a function of the mean, that is, \( \psi(\mu_{i}) = \psi^{2}(\lambda_{i})/\theta^{2}(\mu_{i}) \), and the variance can be written as \( \sigma_{i}^{2} = \phi \psi(\mu_{i}) \). For the normal distribution one obtains \( \psi(\mu_{i}) = 1 \), and for the Poisson \( \psi(\mu_{i}) = \mu_{i} \) (see Table 3.1).

The link between the mean and variance includes the dispersion parameter \( \phi \). However, the latter is not always an additional parameter. It is fixed for the exponential, Bernoulli, binomial, and Poisson distributions. Only for the normal, Gamma, negative binomial, and inverse Gaussian is it a parameter that may be chosen data-dependently. In all these cases the dispersion has the general form

\[
\phi = \phi_{i},
\]

where \( \phi_{i} \) is known with \( \alpha_{i} = 1/\alpha_{i} \), for the binomial distribution and \( \alpha_{i} = 1 \) otherwise. The parameter \( \phi \) is the actual dispersion that is known (\( \phi = 1 \) for exponential, Bernoulli, binomial, Poisson) or an additional parameter. The only case where \( \phi_{i} \neq 1 \) is the binomial distribution, which may be considered as replications of Bernoulli variables.

### Means and Variances

\[
\begin{align*}
\mu_{i} &= \theta^{-1}(\lambda_{i}) \\
\sigma_{i}^{2} &= \phi \psi^{2}(\lambda_{i}) = \phi \psi(\mu_{i})
\end{align*}
\]
links are given, and one obtains, for example,

\[
g(\mu) = \mu \\
g(\mu) = \log(\mu(1 - \mu)) \\
g(\mu) = -1/\mu
\]

for the normal distribution,
for the Bernoulli distribution,
for the Gamma-distribution.

The last example shows that the canonical link might not always be the best choice because

\[-1/\mu = -x_0^T \beta \quad \text{or} \quad \mu = 1/x_0^T \beta \]

implies severe restrictions on \( \beta \) arising from the restriction that \( \mu \) has to be non-negative.

### 3.4.3 Extensions Including Offsets

When modeling insurance claims or numbers of cases \( y_i \) within a time interval, the time interval may depend on \( x_i \) and therefore may vary across observations. With \( \Delta_i \) denoting the underlying time interval for observation \( y_i \), one may assume Poisson-distributed responses \( y_i \sim P(\lambda_i) \), where \( \lambda_i \) is the underlying intensity for one unit of time, which may be any time unit like minutes, days, or months. A sensible approach to modeling will not specify the expectation of \( y_i \), which is \( \lambda_i \), but the intensity \( \lambda_i \) in dependence on covariates and include the time intervals as known constants. The model

\[
\lambda_i = \exp(x_0^T \beta)
\]

yields for the expectation \( \mu_i = E(y_i) \) and \( \Delta_i \),

\[
\mu_i = \Delta_i \lambda_i = \exp(\log(\lambda_i) + x_0^T \beta).
\]

Since \( y_i \) follows a Poisson distribution one has a GLM but with a known additive constant in the predictor. Constants of this type are called offsets; they are not estimated but considered as fixed and known. In the special case where all observations are based on the same length of the time interval, that is, \( \Delta_i = \Delta \), the offset \( \log(\lambda_i) \) is omitted because it cannot be distinguished from the intercept within \( x_0^T \beta \). For more examples see Section 7.4.

### 3.5 Modeling of Grouped Data

In the previous sections observations have been given in the ungrouped form \((y_i, x_i), i = 1, \ldots, n\). Often, for example, if covariates are categorical or in experimental studies, several of the covariate values \( x_{1i}, \ldots, x_{ni} \) will be identical. Thus the responses for fixed covariate vectors may be considered as replications with identical mean. By relabeling the data one obtains the form

\[
(y_{ij}, x_{ij}), \quad i = 1, \ldots, N, \quad j = 1, \ldots, n_i,
\]

where observations \( y_{1i}, \ldots, y_{ni} \) have a fixed covariate vector \( x_i \) with \( n_i \) denoting the sample size at covariate value \( x_i \), yielding the total sum of observations \( n = n_1 + \cdots + n_N \). Since means depend on covariates, one has

\[
\mu_{ij} = \mu_i = E(y_{ij}) = h(x_0^T \beta), \quad j = 1, \ldots, n_i,
\]

and also the natural parameter \( \theta_i = \theta(\mu_i) \) depends on \( x_i \) only. Let the dispersion parameter \( \phi_i = \phi(\mu_i) \) be constant over replications \( y_{ij}, j = 1, \ldots, n_i \). Then one obtains for the mean over individual responses at covariate value \( x_{ij}, y_{ij} = \frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij} \), the density or mass function

\[
L(y_i) = \exp \left\{ \frac{(y_i - h(\theta_i))}{\phi_i} + c(y_{1i}, \ldots, y_{ni}, \theta) \right\}, \quad (3.9)
\]

where \( c(y_{1i}, \ldots, y_{ni}, \theta) \) is the constant term in the mean and variance.

### 3.6 Maximum Likelihood Estimation

For GLMs the most widely used method of estimation is maximum likelihood. The basic principle is to construct the likelihood of the unknown parameters for the sample data, where the likelihood represents the joint probability or probability density of the observed data, considered as a function of the unknown parameters. Maximum likelihood (ML) estimation for all GLMs has a common form. This is due to the assumption that the responses come from an exponential family. The essential feature of the simple exponential family with density

\[
f(y_i; \theta, \phi) = \exp\left(\frac{(y_i - h(\theta))}{\phi} + c(y_i, \theta)\right)
\]

is that the mean and variance are given by

\[
E(y_i) = \frac{\partial h(\theta)}{\partial \theta}, \quad \text{var}(y_i) = \phi \frac{\partial^2 h(\theta)}{\partial \theta^2},
\]

where the parameterization is in the canonical parameter \( \theta \). As will be seen, the likelihood and its logarithm, the log-likelihood, are determined by the assumed mean and variance.

#### Log-Likelihood and Score Function

From the exponential family one obtains for independent observations \( y_{1i}, \ldots, y_{ni} \) the log-likelihood

\[
l(\beta) = \sum_{i=1}^{n} l(\beta_i) = \sum_{i=1}^{n} \sum_{j=1}^{n_i} \log(\phi_i) - \phi_i h(\theta_i) + c(y_{1i}, \ldots, y_{ni}, \theta),
\]

where the term \( c(y_{1i}, \ldots, y_{ni}, \theta) \) is omitted because it does not depend on \( \theta \) and therefore not on \( \beta \). For the maximization of the log-likelihood one computes the derivative \( s(\beta) = \partial l(\beta)/\partial \beta \),
which is called the score function. For the computation it is useful to consider the parameters as resulting from transformations in the form \( \theta_i = \theta(h_{i1}), \mu_i = h_{i2}, \eta_i = x_i^T \beta \). One has the transformation structure

\[
\begin{align*}
\eta_i &= \frac{1}{g} h_{i1} \\
\mu_i &= \frac{1}{g} h_{i2} \\
\theta_i &= \theta(h_{i1})
\end{align*}
\]

yielding \( \theta_i = \theta(\mu_i) = \theta(h_{i1}) \). Then the score function \( s(\beta) = \partial(\beta)/\partial \beta \) is given by

\[
s(\beta) = \frac{\partial(\beta)}{\partial \beta} = \sum_{i=1}^{n} \left( \frac{\partial h_{i1}(\theta_i)}{\partial \theta_i} \frac{\partial \theta_i}{\partial \mu_i} \right) \frac{\partial \mu_i}{\partial \beta}
\]

With \( \mu_i = \mu(\theta_i) \) denoting the transformation of \( \theta_i \) into \( \mu_i \), one obtains

\[
\begin{align*}
\frac{\partial \theta_i}{\partial \mu_i} &= \frac{(y_i - \theta'(\theta_i)) / \theta_i}{(y_i - \theta'(\theta_i)) / \theta_i} \\
\frac{\partial \mu_i}{\partial \beta} &= \frac{\partial \theta_i}{\partial \beta} \left( \frac{\partial \theta_i}{\partial \theta_i} \right)^{-1} = \left( \frac{\partial \theta_i}{\partial \theta_i} \right)^{-1} \phi_i / \text{var}(y_i), \\
\frac{\partial \theta_i}{\partial \beta} &= x_i,
\end{align*}
\]

and therefore the score function

\[
s(\beta) = \sum_{i=1}^{n} s_i(\beta) = \sum_{i=1}^{n} x_i \frac{\partial h_{i1}(\theta_i)}{\partial \theta_i} \left( \frac{y_i - \mu_i}{\text{var}(y_i)} \right),
\]

With \( \sigma^2 = \phi \text{var}(\mu_i) = \text{var}(y_i) \), the estimation equation \( s(\beta) = 0 \) has the form

\[
\sum_{i=1}^{n} x_i \left( \frac{y_i - \mu_i}{\text{var}(y_i)} \right) = 0.
\]

(3.10) in the response (or link) function is found in the specification of the mean \( \mu_i = h(\alpha x_i^T \beta) \) and in the derivative \( \partial h_{i1}(\theta_i) / \partial \theta_i \), whereas higher moments of the distribution of \( y_i \) only the variance \( \sigma^2 = \phi \text{var}(\mu_i) \) is needed. Since \( \phi = \text{disp} \), the dispersion parameter \( \phi \) may be canceled out and the estimate \( \hat{\beta} \) does not depend on \( \phi \).

For the canonical link the estimation equation simplifies. Since \( \eta_i = x_i^T \beta \), the score function reduces to

\[
s(\beta) = \sum_{i=1}^{n} \frac{\partial h_{i1}(\eta_i)}{\partial \eta_i} \left( \frac{y_i - \mu_i}{\text{var}(y_i)} \right)
\]

and one obtains

\[
s(\beta) = \sum_{i=1}^{n} x_i (y_i - \mu_i) \phi_i.
\]

In particular, one has \( \partial h_{i1}(\eta_i)/\partial \eta_i = \text{var}(y_i)/\phi_i \) if the canonical link is used. In matrix notation the score function is given by

\[
s(\beta) = X^T D \Sigma^{-1}(y - \mu),
\]

where \( X^T \) is the design matrix, \( D = \text{Diag}(\partial h_{i1}(\eta_i)/\partial \eta_i, \ldots, \partial h_{ni}(\eta_i)/\partial \eta_i) \) is the diagonal matrix of derivatives, \( \Sigma = \text{Diag}(\sigma^2_1, \ldots, \sigma^2_n) \) is the covariance matrix, and \( y^T = (y_1, \ldots, y_n), \mu^T = (\mu_1, \ldots, \mu_n) \) are the vectors of observations and means. Sometimes it is useful to combine \( D \) and \( \Sigma \) into the weight matrix \( W = D \Sigma^{-1} D^T \), yielding \( s(\beta) = X^T W D^{-1}(y - \mu) \) and \( F(\beta) = X^T W X \).

### 18. Maximum Likelihood Estimation

The information matrix is the maximum likelihood theory the information matrix determines the asymptotic variance. The observed information matrix is given by

\[
F_{obs}(\beta) = -\frac{\partial^2 l(\beta)}{\partial \beta \partial \beta^T} = \left( \frac{\partial^2 l(\beta)}{\partial \beta \partial \beta^T} \right)^{-1}.
\]

In explicit form shows that it depends on the observations and therefore is random. The expected information or Fisher matrix, which is not random, is given by

\[
F(\beta) = \mathbb{E}(F_{obs}(\beta)).
\]

For the derivation it is essential that \( \mathbb{E}(s(\beta)) = 0 \) and that \( \mathbb{E}(-\partial^2 l(\beta)/\partial \beta \partial \beta^T) = \mathbb{E}((\partial h_{i1}(\eta_i)/\partial \beta \partial \beta^T), \ldots, (\partial h_{ni}(\eta_i)/\partial \beta \partial \beta^T)) \), which holds under general assumptions (see, for example, Cox and Hinkley, 1974). Thus one obtains

\[
F(\beta) = \mathbb{E} \left( \sum_{i=1}^{n} s_i(\beta) s_i(\beta)^T \right) = \mathbb{E} \left( \sum_{i=1}^{n} x_i x_i^T \left( \frac{\partial h_{i1}(\eta_i)}{\partial \eta_i} \right)^2 \frac{\text{var}(y_i)^2}{\text{var}(y_i)^2} \right) \]

\[
- \sum_{i=1}^{n} x_i x_i^T \left( \frac{\partial h_{i1}(\eta_i)}{\partial \eta_i} \right)^2 / \sigma^2,
\]

where \( \sigma^2 = \phi \text{var}(\mu) \). By using the design matrix \( X \) one obtains the information matrix \( F(\beta) \) in the form

\[
F(\beta) = X^T W X,
\]

where \( W = \text{Diag} \left( \frac{\text{var}(y_1)}{\phi}, \ldots, \frac{\text{var}(y_n)}{\phi} \right) \) is a diagonal weight matrix that has the matrix form \( W = D \Sigma^{-1} D^T \).

For the canonical link the corresponding simpler form is

\[
F(\beta) = \sum_{i=1}^{n} x_i x_i^T / \sigma^2 = X^T X / \sigma^2.
\]

In this case it is well known that the covariance of the estimator \( \hat{\beta} \) is given by \( \text{cov}(\hat{\beta}) = \sigma^2 (X^T X)^{-1}, F(\beta)^{-1} \). For GLMs the result holds only asymptotically (\( n \to \infty \)). With

\[
\text{cov}(\hat{\beta}) = \text{cov} \left( \begin{bmatrix} x_1 & \ldots & x_n \end{bmatrix} \right)^{-1},
\]

where \( W \) means the evaluation of \( W \) at \( \hat{\theta} \); that is, \( \partial \ln(\eta_i)/\partial \theta_i \) is replaced by \( \partial \ln(\eta_i)/\partial \theta_i, \theta_i = x_i^T \hat{\beta}, \sigma^2_1 - \text{var}(\mu_i), \mu_i = h(\eta_i) \).

It should be noted that in the grouped observations case the form of the likelihood, score function, and Fisher matrix are the same; only the summation index \( n \) has to be replaced by \( IV \).
The unifying concept of GLMs may be seen in the common form of the log-likelihood, the score function (which determines the estimation equation), and the information matrix (which determines the variances of estimators). Specific models result from specific choices of

- the link or response function, yielding the derivative matrix \( D \), which contains \( \frac{\partial h(y)}{\partial \beta} \);
- the distribution, yielding the covariance matrix \( \Sigma \), which contains \( \sigma^2 \); and
- the explanatory variables, which determine the design matrix \( X \).

In GLMs these constituents may be chosen freely. In principle, any link function can be combined with any distribution and any set of explanatory variables. Of course there are combinations of links and distributions that are more sensible than others.

### 3.7 Inference

Main questions in inference concern

- the adequacy of the model or goodness-of-fit of the model,
- the relevance of explanatory variables,
- the explanatory value of the model.

In the following these questions are considered in a different order. First the deviance is introduced, which measures the discrepancy between the observations and the fitted model. The deviance is a tool for various purposes. The relevance of the explanatory variables may be investigated by comparing the deviance of two models, the model that contains the variable in question and the model where this variable is omitted. Moreover, for grouped observations the deviance may be used as a goodness-of-fit statistic.

#### 3.7.1 The Deviance

When fitting a GLM one wants some measure for the discrepancy between the fitted model and the observations. The deviance is a measure for the discrepancy that is based on the likelihood ratio statistic for comparing nested models. The nested models that are investigated are the GLM that is under investigation and the most general possible model. This so-called saturated model fits the data exactly by assuming as many parameters as observations.

Let \( (y, \mu, \phi) \) denote the maximum of the log-likelihood of the model where \( y^T = (y_1, \ldots, y_n) \) represents the data, \( \mu^T = (\mu_1, \ldots, \mu_n) \), \( \mu_i = h(x_i^T \beta) \) represent the fitted values based on the ML estimate \( \hat{\beta} \), and the dispersion of observations has the form \( \phi \sigma^2 \), with known \( \sigma^2 \). For the saturated model that matches the data exactly one has \( \mu = y \) and the log-likelihood is given by \( l(y, \mu, \phi) \). With \( \theta(\mu), \theta(y) \) denoting the canonical parameters of the GLM under investigation and the saturated model, respectively, the deviance is given by

\[
D(y, \mu, \phi) = -2 \log \{ l(y, \mu, \phi) - l(y, \mu, \phi) \}
\]

\[
= -2 \sum_{i=1}^n \left[ y_i \theta(y_i) - \theta(y_i) - (b(\theta(y_i)) - b(\theta(\mu_i))) / a_i \right]
\]

\( D(y, \mu, \phi) \) is known as deviance of the model under consideration while \( D^2(y, \mu, \phi) = D(y, \mu, \phi) / \phi \) is the so-called scaled deviance. The deviance is linked to the likelihood ratio statistic.
TABLE 3.2: Deviances for normal distributions.

<table>
<thead>
<tr>
<th>Distribution</th>
<th>Deviance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal</td>
<td>$D(y, \mu) = \sum_{i=1}^{n} (y_i - \mu_i)^2$</td>
</tr>
<tr>
<td>Gamma</td>
<td>$D(y, \mu) = \sum_{i=1}^{n} \log \left( \frac{y_i}{\mu_i} \right) + \left( \frac{y_i - \mu_i}{\mu_i} \right)$</td>
</tr>
<tr>
<td>Inverse Gaussian</td>
<td>$D(y, \mu) = \sum_{i=1}^{n} (y_i - \mu_i)^T \hat{\Sigma}^{-1} (y_i - \mu_i)$</td>
</tr>
<tr>
<td>Bernoulli</td>
<td>$D(y, \mu) = \sum_{i=1}^{n} (y_i \log(\mu_i) + (1 - y_i) \log(1 - \mu_i))$</td>
</tr>
<tr>
<td>Poisson</td>
<td>$D(y, \mu) = \sum_{i=1}^{n} \log \left( \frac{y_i}{\mu_i} \right) + \log(1 - y_i)$</td>
</tr>
</tbody>
</table>

\[ \lambda = -2 \left[ l(y, \mu, \phi) - l(y, \hat{\mu}, \hat{\phi}) \right], \] which compares the current model to the saturated model by $D(y, \mu) = \phi \lambda$.

Simple derivation yields the deviances given in Table 3.2. For the normal model, the deviance is identical to the error or residual sum of squares SSE and the scaled deviance takes the form $\text{SSE}/n^2$. For the Bernoulli distribution, one has $\theta(\mu_i) = \log(\mu_i/(1 - \mu_i))$ and one obtains $D(y, \mu_i) = 2 \sum d(y_i, x_i)$, where $d(y_i, x_i) = -\log(1 - y_i - x_i)$ (for more details see Section 4.2). In the cases of the Poisson and the Gamma deviances, the last term given in brackets is omitted if the model includes a constant term because then the sum over the terms is zero.

The deviance as a measure of discrepancy between the observations and the fitted model may be used in an informal way to compare the fit of two models. For example, two models with the same predictor but differing link functions can be compared by considering which one has the smaller deviance. However, there is no simple way to interpret the difference between the deviances of these models. This is different if the deviance of discrepancy is used for nested models, for example, to investigate the relevance of terms in the linear predictor. The comparison of models with and without the term in question allows one to make a decision based on significance tests with a known asymptotic distribution. The corresponding analysis of deviance (see Section 3.7.2) generalizes the analysis of variance, which is in common use for normal linear models.

For ungrouped data some care has to be taken in the interpretation as a goodness-of-fit measure. As an absolute measure of goodness-of-fit, which allows one to decide if the model has satisfactory fit or not, the deviance for ungrouped observations is appropriate only in special cases. For the interpretation of the value of the deviance it would be useful to have a benchmark in the form of an asymptotic distribution. Since the deviance may be derived as a likelihood ratio statistic, it is tempting to assume that the deviance is asymptotically $\chi^2$-distributed. However, in general, the deviance does not have an asymptotic $\chi^2$-distribution in the limit for $n \to \infty$. Standard asymptotic theory of likelihood ratio statistics for nested models assumes that the ranks of the design matrices that build the two models and therefore the degrees of freedom are fixed for increasing sample size. In the present case this theory does not apply because the degrees of freedom of the saturated model increase with $n$. This is already seen in the case of the normal distribution, where $D(y, \mu) = \sum (y_i - \hat{\mu})^2$ $\to \chi^2(n - p)$. For $n \to \infty$, the limiting distribution does not have a $\chi^2$-distribution with fixed degrees of freedom. Similar effects occur for binary data.

3.7.2 Analysis of Deviance and the Testing of Hypotheses

Let us consider the nested models $\hat{M} \subset M$, where $M$ is a given GLM with $\mu = \text{h}(\text{x} \beta)$, and $\hat{M}$ is a submodel that is characterized by the linear restriction $C \beta = \xi$, where $C$ is a known $(s \times p)$ matrix with rank $(C) = s \leq p$ and $\xi$ is an $s$-dimensional vector. This means that $\hat{M}$ corresponds to the null hypothesis $H_0 : C \beta = \xi$, which specifies a simpler structure of the predictor.

**Analysis of Deviance**

With $\hat{\mu}^T = (\hat{\mu}_1, \ldots, \hat{\mu}_s)$ denoting the fitted values for the restricted model $\hat{M}$ and $\hat{\mu}^T = (\hat{\mu}_1, \ldots, \hat{\mu}_s)$ denoting the fit for model $M$, one obtains the corresponding deviances

\[ D(M) = -2 \phi(l(y, \mu, \phi) - l(y, \hat{\mu}, \phi)), \]

\[ D(\hat{M}) = -2 \phi(l(y, \mu, \phi) - l(y, \hat{\mu}, \phi)). \]

The difference of deviances

\[ D(\hat{M}) = D(M) - D(\hat{M}) = -2 \phi(l(y, \mu, \phi) - l(y, \hat{\mu}, \phi)) \] (3.11)

compares the fits of models $M$ and $\hat{M}$. The difference of scaled deviances $D(\hat{M})/\phi$ is equivalent to the likelihood ratio statistic for testing $H_0$. Similar to the partitioning of the sum of squares in linear regression, one may consider the partitioning of the deviance of the restricted model $\hat{M}$ into

\[ D(\hat{M}) = D(M) + D(\hat{M}) \]

$D(\hat{M})$ gives the increase in discrepancy between the data and the fit if model $\hat{M}$ is fitted instead of the less restrictive model $M$. For normal distributions this corresponds to the partitioning of the sum of squares

\[ \text{SSE}(\hat{M}) = \text{SSE}(\hat{M}) + \text{SSE}(M) \]

(see Section 1.4.6), which, for the special model where $\hat{M}$ contains only an intercept, reduces to $\text{SSE}(\hat{M}) = \text{SSE}(M)$. In the normal case one obtains for $\text{SSE}(\hat{M}), \text{SSE}(M)$ as $\sigma^2 \chi^2(n - p)$ distribution. If $\hat{M}$ holds $p$ terms which the dimension of the predictor in $M$. If $\hat{M}$ holds, $\text{SSE}(\hat{M})$ and $\text{SSE}(M)$ are independent with $\text{SSE}(\hat{M}) \sim \sigma^2 \chi^2(s)$ and $\text{SSE}(M) \sim \sigma^2 \chi^2(n - s - p)$. For testing $H_0$ one uses the $F$-statistic

\[ \frac{\text{SSE}(\hat{M}) - \text{SSE}(M)}{\hat{\phi}^2} \sim F(s, n - p), \]

where $\hat{\phi}^2 = \text{SSE}(M)/(n - p)$. In the general case of GLMs one uses

\[ D(\hat{M}) = D(M) \phi \]

\[ \phi \]

which under mild restrictions is asymptotically χ²(s)-distributed. This means that the difference

\[ D(M) - D(M) = D(M|M) \]

has an asymptotically a χ²(s)-distribution.

When using the χ²-approximation the deviance has to be scaled by 1/p. For the binomial
(\( p_i = 1/n_i, \phi = 1 \)) Bernoulli, exponential, and Poisson (\( p_i = 1 \)) distributions one may use the difference \( D(M) - D(M) \) directly, whereas for the normal, Gamma, and Inverse Gamma the dispersion parameter has to be estimated. In the normal regression case \( \frac{1}{2} D(M|M) | \phi \) has a \( \chi^2(n - p) \) distribution. In the general case the approximation by the F-distribution may be used if \( \phi \) is consistent for \( \phi_i \), has approximately a scaled \( \chi^2 \) distribution, and \( D(M) - D(M) \)

and \( \phi \) are approximately independent (Jorgensen, 1987). In analogy to the ANOVA table, in normal regression one obtains a table for the analysis of deviance (see Table 3.3).

### Table 3.3: Analysis of deviance table.

<table>
<thead>
<tr>
<th>df</th>
<th>cont. deviance</th>
<th>df</th>
</tr>
</thead>
<tbody>
<tr>
<td>D(M)</td>
<td>n - p + s</td>
<td>D(M)</td>
</tr>
<tr>
<td>D(M)</td>
<td>n - p</td>
<td>D(M</td>
</tr>
</tbody>
</table>

It should be noted that only the difference of deviances \( D(M|M) \) has asymptotically a \( \chi^2(s) \) distribution. The degrees of freedom of \( D(M) \) have the basic structure “number of observations minus number of fitted parameters.” In \( M \), by considering an additional s-dimensional restriction, the effective parameters in the model are reduced to \( p - s \), yielding \( df = n - (p - s) = n - p + s \). In the case of grouped data, the deviances \( D(M) \) and \( D(M) \) themselves are asymptotically distributed with \( D(M) - \chi^2(N - p + s) \) and \( \chi^2(N - p) \), where \( N \) denotes the number of grouped observations (see Section 3.8). While \( D(M) \) and \( D(M) \) are different for grouped and ungrouped data, the difference \( D(M) - D(M) \) is the same.

Next we give a summary of results on distributions for the classical linear case and the deviances within the GLM framework. For the classical linear model one has

\[ SSE(M) = SSE(M|M) + SSE(M) \]

\[ s \chi^2(n + p + s) \]

if \( M \) holds

\[ s \chi^2(n + p) \]

if \( M \) holds

For grouped data within the GLM framework, one has the asymptotic distributions

\[ D(M|M) = D(M|M) + D(M) \]

\[ \phi \chi^2(N - p + s) \]

if \( M \) holds

\[ \phi \chi^2(N - p) \]

if \( M \) holds

The approach may be used to test sequences of nested models,

\[ M_1 \subset M_2 \subset \ldots \subset M_m, \]

by using the successive differences \( (D(M_j) - D(M_{j+1})) | \phi \). The deviance of the most restrictive model is given as sum of these differences:

\[ D(M_1) = (D(M_1) - D(M_2)) + (D(M_2) - D(M_3)) + \ldots + (D(M_{m-1}) - D(M_m)) + D(M_m) \]

\[ = D(M_1|M_2) + \ldots + D(M_{m-1}|M_m) + D(M_m). \]

3.7.3 Alternative Test Statistics for Linear Hypotheses

The analysis of deviance tests if a model can be reduced to a model that has a simpler structure in the covariates. The simplified structure is specified by the null hypothesis \( H_0 \) of the pair of hypotheses

\[ H_0: C \beta = \xi \text{ against } H_1: C \beta \neq \xi, \]

where rank(C) = s. Alternative test statistics that can be used are the Wald test and the score test.

**Wald Test**

The Wald statistic has the form

\[ w = (C \hat{\beta} - \xi)^T [C F^{-1}(C) C^{-1}]^{-1} (C \hat{\beta} - \xi). \]

It uses the weighted distance between the unrestricted estimate \( C \hat{\beta} \) of \( C \beta \) and its hypothetical value \( \xi \) under \( H_0 \). The weight is derived from the distribution of the difference \( (C \hat{\beta} - \xi) \), for which one obtains asymptotically \( \text{cov}(C \beta - \xi) = C F^{-1}(C) C^{-1} \). Therefore, \( w \) is the squared length of the standardized estimate \( (C F^{-1}(C) C^{-1})^{1/2} (C \hat{\beta} - \xi) \), and one obtains for \( w \) under \( H_0 \) an asymptotic \( \chi^2(s) \) distribution.

An advantage of the Wald statistic is that it is based on the ML estimates of the full model. Therefore, it is not necessary to compute an additional fit under \( H_0 \). This is why most programs packages give significance tests for single parameters in terms of the Wald statistic. When a single parameter is tested with \( H_0: \beta_j = 0 \), the corresponding matrix \( C \) is \( C = [0, 1, \ldots, 1, 0] \). Then the Wald statistic has the simple form

\[ w = \frac{\beta_j}{\hat{\beta}_j}, \]

where \( \beta_j \) is the jth diagonal element of the estimated inverse Fisher matrix \( F^{-1} \). Since \( w \) is asymptotically \( \chi^2(1) \) distributed, one may also consider the square root,

\[ z = \sqrt{w} = \frac{\beta_j}{\sqrt{\hat{\beta}_j}}, \]

which follows asymptotically a standard normal distribution. Thus, for single parameters, program packages usually give the standard error \( \sqrt{\hat{\beta}_j} \) and the p-value based on it.

### Score Statistic

The score statistic is based on the following consideration: The score function \( s(\beta) \) for the unrestricted model is the vector of if it is evaluated at the unrestricted ML estimate \( \hat{\beta} \). If, however, \( \beta \) is replaced by the ML estimate \( \hat{\beta} \) under \( H_0 \), \( s(\beta) \) will be significantly different from zero if \( H_0 \) is not true. Since the covariance of the score function is approximately the Fisher matrix, one uses the score statistic,

\[ \nu = s(\beta)|F^{-1}(\hat{\beta}) s(\beta), \]

which is the squared weighted score function evaluated at \( \hat{\beta} \).
3.8 Goodness-of-Fit for Grouped Observations

It has already been mentioned that for grouped observations the deviance has an asymptotic $χ^2$ distribution. Hence, it may be used to test the model fit.

3.8.1 The Deviance for Grouped Observations

The analysis of deviance and alternative tests provide an instrument that helps to decide if a more parsimonious model $M$ may be chosen instead of the more general model $\hat{M}$, where $\hat{M} \subset M$. The test statistics may be seen as tools to investigate the fit of model $\hat{M}$ given model $M$. However, they are of limited use for investigating if a model is appropriate for the given data, that is, the model fit compared to the data. The only possibility would be to choose $\hat{M}$ as the saturated model. But then the deviance has no fixed distribution in the limit.

A different situation occurs if replications are available. If, for a fixed covariate vector $x_i$, independent replications $y_{i1}, \ldots, y_{im}$ are observed, the mean across replications $\bar{y}_i = \frac{1}{m} \sum y_{ij}$ again represents a GLM and the deviance for the means $\bar{y}_i, \ldots, \bar{y}_n$ may be used.

For grouped data with response $y_i$, $i = 1, \ldots, N$, the essential difference is that the scale parameter is given as $\phi_i = \bar{y}_i/n_i$, where $\phi$ is the dispersion for the single observations. Since
models $\hat{M} \in M$ and corresponding fits $\hat{\mu}, \hat{\mu}$, the difference $D(\hat{\mu}, \hat{\mu})$ for grouped data are the same as the difference $D(\hat{\mu}, \hat{\mu}) - D(\hat{\mu}, \hat{\mu})$ for ungrouped observations. Therefore, $(D(\hat{\mu}, \hat{\mu}) - D(\hat{\mu}, \hat{\mu}))/\sigma^2$ is asymptotically $\chi^2$-distributed, where $\sigma$ is the difference between the number of parameters in $M$ and $M$.

For the normal linear model alternative tests for the lack-of-fit are available. The partitioning (of ungrouped) least squares data $(y_{ij}, x_{ij}), j = 1, \ldots, n_i$, yields

$$\sum_{i=1}^N n_i (\bar{y}_{ij} - \hat{\mu})^2 = \sum_{i=1}^N n_i (\bar{y}_{ij} - \hat{\mu})^2 + \sum_{i=1}^N n_i (\bar{y}_{ij} - \hat{\mu})^2,$$

which has the form

$$D(\hat{\mu}) = D(\hat{\mu}/M) + D(M),$$

where $\hat{M}$ stands for the linear model and $M$ for a model where only $y_{ij} = \mu + \epsilon_{ij}$ with $\epsilon_{ij} \sim N(0, \sigma^2)$ is assumed. Since in computing $D(\hat{\mu}/M)$ no assumption on linearity is assumed, it is also called the pure error sum of squares and $D(\hat{\mu}/M)$ the lack of fit sum of squares. By use of the mean squares one obtains for $H_0 : \beta = X\beta$ against $H_1 : \beta \neq X\beta$ the F-statistic

$$F = \frac{\frac{1}{N} \sum_{i=1}^N n_i (\bar{y}_{ij} - \hat{\mu})^2 / (N - p)}{\frac{1}{N} \sum_{i=1}^N (y_{ij} - \bar{y}_{ij})^2 / (n - N)}.$$

Linearity is dismissed if $F > F_1(\alpha, N - p, n - N)$. When using the F-statistic, not all levels of covariates need to have repeated observations, only some of the $n_i$'s have to be larger than 1. However, the test is still based on the assumptions that responses are normally distributed and have variance $\sigma^2$. Note that $D(\hat{\mu}|M)$ is the deviance for grouped observations.

### 3.8.2 Pearson Statistic

An alternative measure for the discrepancy between the data and the model is the Pearson statistic:

$$\chi^2_p = \sum_{i=1}^N n_i (\bar{y}_{ij} - \hat{\mu})^2 / v(\hat{\mu}/n_i),$$

where $\hat{\mu}$ is the mean for grouped observations, $\hat{\mu}$ is the estimated mean, and $v(\hat{\mu})$ is the corresponding variance function that is linked to the variance by $v(\mu) = v(\mu)|v(\mu)$. If fixed cells asymptotics applies ($N$ fixed, $n_i \to \infty$, $n_i / n_i \to c_i$, $c_i > 0$), $\chi^2_p$ is asymptotically $\chi^2$-distributed. This is an approximately $\chi^2(N - p)$-distribution of the dispersion parameter $\phi$ has to be known and fixed since the estimation of $\phi$ is based on this statistic. Replacing $\phi$ by the dispersion estimate from grouped observations $\hat{\phi}_N = \chi^2_p(N - p)$ would yield the trivial result $\chi^2_p = N - p$.

### 3.9 Computation of Maximum Likelihood Estimates

Maximum likelihood estimates are obtained by solving the equation $s(\hat{\beta}) = 0$. In general, there is no closed form of the estimate available, and iterative procedures have to be applied. In matrix notation the score function is given by

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

where $X$, $D$, and $W$ depend on $\beta$ and are suppressed (see Section 3.6).

The Newton-Raphson method is an iterative method for solving non-linear equations. Starting with an initial guess $\beta^{(0)}$, the solution is found by successive improvement. Let $\beta^{(k)}$ denote the estimate in the kth step, where $k = 0$ is the initial estimate. If $s(\beta^{(k)}) \neq 0$, one considers the linear Taylor approximation

$$s(\beta) = s^{(0)}(\beta) + \frac{\partial s(\beta)}{\partial \beta}(\beta - \beta^{(k)}).$$

Instead of solving $s(\beta) = 0$ one solves $s^{(0)}(\beta) = 0$, yielding

$$\beta = \beta^{(k)} - \frac{\partial s(\beta)}{\partial \beta}(\beta - \beta^{(k)}).$$

Since $\partial s(\beta)/\partial \beta = \partial^2 s(\beta)/\partial \beta^2$, one obtains with the Hessian matrix $H(\beta) = \partial^2 s(\beta)/\partial \beta^2$ the new estimate

$$\beta^{(k+1)} = \beta^{(k)} - H(\beta)^{-1}s(\beta^{(k)}),$$

by using the observed information matrix $F_0 X_0 X_0' F_0$. Iterations are carried out until the changes between successive steps are smaller than a specified threshold $\epsilon$. Iteration is stopped if

$$\|\beta^{(k+1)} - \beta^{(k)}\| / \|\beta^{(k)}\| < \epsilon.$$
Convergence is usually fast, with the number of correct decimals in the approximation roughly doubling at each iteration.

An alternative method is the Newton method with Fisher scoring. The essential difference is that the observed information matrix \( F_{\theta} \) is replaced by the expected information \( F(\theta) = E F_{\theta} | Z \) by \( E F_{\theta} \), yielding

\[
\hat{\theta}^{(k+1)} = \hat{\theta}^{(k)} + F(\hat{\theta}^{(k)})^{-1} \hat{F}(\hat{\theta}^{(k)}).
\]

(3.13)

The iterative scheme (3.13) may alternatively be seen as an iterative weighted least-squares fitting procedure. Let pseudo- or working observations be given by

\[
\hat{\eta}(\hat{\theta}) = \eta^{T} \hat{\theta} + \left( \frac{\partial \hat{h}(z)}{\partial \eta} \right)^{-1} \left( \hat{y} - \hat{r}(\hat{\theta}) \right)
\]

and \( \tilde{\eta}(\theta)^{T} = \hat{\eta}(\hat{\theta}) \cdots \hat{\eta}(\hat{\theta}) \) denote the vector of pseudo-observations given by \( \hat{\eta}(\hat{\theta}) = X \hat{\beta} + D (\hat{\theta})^{-1} (y - \mu) \). One obtains by simple substitution

\[
\tilde{\beta}^{(k+1)} = (X^{T} W(\tilde{\beta}^{(k)}) X)^{-1} X^{T} W(\tilde{\beta}^{(k)}) \hat{\eta}(\tilde{\beta}^{(k)}).
\]

Thus \( \tilde{\beta}^{(k+1)} \) has the form of a weighted least-squares estimate for the working observations \( \hat{\eta}(\hat{\theta}) \), \( i = 1, \ldots, n \), with the weight \( W(\tilde{\beta}^{(k)}) \) depending on the iteration.

For a canonical link one obtains \( \hat{F}_{\beta}(\theta) = X \beta + W \phi^{-1} \phi^{-1} \mu \) and score functions \( s(\beta) = X \phi^{-1} \mu \) and therefore

\[
\tilde{\beta}^{(k+1)} = \hat{\beta}^{(k)} + (X^{T} W X)^{-1} X^{T} \phi^{-1} (y - \mu),
\]

which corresponds to least-squares fitting

\[
\tilde{\beta}^{(k+1)} = (X^{T} W X)^{-1} X^{T} W \hat{\eta}(\hat{\theta}^{(k)}),
\]

with \( \hat{\eta}(\hat{\theta}) = X \hat{\beta} + W \phi^{-1} \mu \). If \( \phi = \theta \), one obtains

\[
\tilde{\beta}^{(k+1)} = \hat{\beta}^{(k)} + (X^{T} \Sigma X)^{-1} X^{T} \theta (y - \mu),
\]

which corresponds to least-squares fitting

\[
\tilde{\beta}^{(k+1)} = (X^{T} W X)^{-1} X^{T} W \hat{\eta}(\hat{\theta}^{(k)}).
\]

3.10 Hat Matrix for Generalized Linear Models

With weight matrix \( W(\beta) = D(\beta) \Sigma(\beta)^{-1} D(\beta)^{T} \), the iterative fitting procedure has the form

\[
\hat{\beta}^{(k+1)} = (X^{T} W(\hat{\beta}^{(k)}) X)^{-1} X^{T} W(\hat{\beta}^{(k)}) \hat{\eta}(\hat{\beta}^{(k)}).
\]

At convergence one obtains

\[
\hat{\beta} = (X^{T} W(\beta) X)^{-1} X^{T} W(\beta) \hat{\eta}(\beta),
\]

Thus \( \hat{\beta} \) may be seen as the least-squares solution of the linear model

\[
W^{1/2} \hat{\eta}(\beta) = W^{1/2} X \hat{\beta} + \hat{\epsilon},
\]

where in \( W = W(\beta) \) the dependence on \( \beta \) is suppressed. The corresponding hat matrix has the form

\[
H = W^{1/2} X (X^{T} W X)^{-1} X^{T} W^{1/2}.
\]
3.11 Quasi-Likelihood Modeling

In generalized linear models it is assumed that the true density of the responses follows an exponential family. However, in applications it is not too infrequently found that the implicitly specified variation of the responses is not consistent with the variation of the data. Approaches that use only the first two moments of the response distribution are based on so-called quasi-likelihood estimates (Wedderburn, 1974; McCullagh and Nelder, 1989). When using quasi-likelihood estimates, the exponential family assumption is dropped, and the mean and variance structures are separated. No full distributional assumptions are necessary. Under appropriate conditions, parameters can still be estimated consistently, and asymptotic inference is possible under appropriate modifications.

Quasi-likelihood approaches assume, like GLMs, that the mean and variance structures are correctly specified by

\[ E(y_i | x_i) = \mu_i = \mu(x_i ^T \beta) \]

\[ \text{var}(y_i | x_i) = \sigma^2(\mu_i) = \phi(\mu_i) \]

(3.17)

where \( \mu(\phi) \) is a variance function and \( \phi \) is a dispersion parameter. The main difference from GLMs is that the mean and variance do not have to be specified by an exponential family. The usual maximum likelihood estimates for GLMs are obtained by setting the score function equal to zero:

\[ \mathbf{s}(\beta) = X^T \mathbf{D}(\beta) \Sigma^{-1}(\beta) (y - \mu) = 0. \]

(3.18)

The ML estimation equation uses the link (in \( \mathbf{D}(\beta) \)) and the variance function (in \( \Sigma(\beta) \)), but no higher moments. The solution of (3.18) yields the ML estimates when the mean and variance correspond to an exponential family. The quasi-likelihood (QL) estimates are obtained when the specification of the mean and the variance is given by (3.17) without reference to an exponential family. One may understand \( \sigma^2_Q(\beta) = X^T \mathbf{D}(\beta) \Sigma^{-1}(\beta) (y - \mu) \)

with the specifications (3.17) as a quasi-score function with the corresponding estimation equation \( \sigma^2_Q(\beta) = 0 \). It is also possible to construct a quasi-likelihood function \( Q(\beta, \phi) \) that has the derivative \( \sigma^2_Q(\beta) = \partial Q(\beta, \phi) / \partial \beta \) (Nelder and Pregibon, 1987; McCullagh and Nelder, 1989).

It can be shown that the asymptotic properties are similar to those for GLMs. In particular, one obtains asymptotically a normal distribution with the covariance given in the form of a pseudo-Fisher matrix:

\[ F_Q(\beta) = \sum_{i=1}^n \left( \frac{\partial \mu_i}{\partial \mu} \right)^2 \left( \frac{\partial^2 \mu_i}{\partial \mu \partial \mu} \right) X^T \mathbf{D}(\beta) \Sigma^{-1}(\beta) (y - \mu) X \]

(see also Wedderburn, 1974; McCullagh and Nelder, 1989).

It should be noted that quasi-likelihood models weaken the distributional assumptions considerably. One obtains estimates of parameters without assuming a specific distribution. One just has to specify the mean and the variance and is free to select a variance function \( \phi(\mu) \) that is not determined by a fixed distribution.

A major area of application is the modeling of overdispersion. For example, in count data the assumption of the Poisson distribution means that the variance depends on the mean in the form \( \text{var}(y_i) = \mu_i = \exp(\mu_i)^2 \). In quasi-likelihood approaches one might assume that the variance is \( \text{var}(y_i) = \phi(\mu_i) = \phi(\exp(\mu_i)) \), with an additional unknown dispersion parameter \( \phi \). Since the Poisson distribution holds for \( \phi = 1 \) only, one does not assume the Poisson model to hold (see Sections 5.3 and 7.5).

3.12 Further Reading


R packages. GLMs can be fitted by use of the model fitting functions glm from the MASS package. Many tools for diagnostics and inference are available.

3.13 Exercises

3.3. Let independent observations \( y_1, \ldots, y_n \) have a fixed covariate vector \( x_1, \ldots, x_n \), and let \( y_i \) denote the sample size at covariate value \( x_i \). The model to be examined has the form \( y_i = n_i + \text{var}(\mu_i) = \text{var}(\exp(\mu_i)) \).

(a) Let \( \text{var}(\mu_i) = (b_i)^2 \), with \( b_i > 0 \). Show that the mean \( y_i = n_i \) of \( y_i \) has the form of a simple exponential family. Compare the canonical parameter \( \eta_i \) with the dispersion parameter \( \phi \).

(b) Let \( \text{var}(\mu_i) = \text{var}(\exp(\mu_i)) \). Show that the mean \( y_i \) always has the form of a simple exponential family if the mean \( y_i \) has the form of a simple exponential family.

3.2. Let observations \( y_i, x_i \), \( i = 1, \ldots, n \), with binary response \( y_i \in \{0, 1\} \) be given. The used models fit the logistic model \( \text{var}(\mu_i) = \text{var}(\exp(\mu_i)) \), with \( \mu_i = 0 \) the standard normal distribution function. Give the logistic function and derive the score function \( s(\theta) \), the matrix of derivatives \( D(\beta) \), and the covariance matrix \( E(\beta) \) for both models. In addition, give the observed and expected information matrices.

3.3. Consider a GLM with Poisson-distributed responses.
(a) Derive the Fisher matrix for the canonical link function.
(b) Show that the asymptotic distribution for grouped data in \( N \) groups has approximate covariance 
\[
cov(\hat{\beta}) \approx (X' \text{diag}(\mu) X)^{-1},
\]
where \( \text{diag}(\mu) = \mu' \).

3.4 Let the independent observations \( y_1, \ldots, y_n \), observed at predictor value \( x_n \), follow a Poisson distribution, \( y_i \sim P(\lambda_i) \). Then one obtains for the sum \( \bar{y} = \sum_{i=1}^{n} y_i \sim P(\sum_{i=1}^{n} \lambda_i) \), where \( \lambda_i = \sum_{j} \lambda_{ij} \).
Discuss modeling strategies for the observations. Consider in particular models for single variables \( y_{ij} \), for accumulated counts \( \bar{y}_i \), and for average counts \( \bar{y}_i/n_i \).

3.5 Let \( (x_i, y_i) \) denote independent observations. A linear model with log-transformed responses is given by \( \log(y_i) = \beta' \vec{x}_i + \epsilon_i \sim N(0, \sigma^2) \). Compare the model to the GLM 
\[
y_i | x_i \sim N(\mu_i, \sigma^2) \quad \text{and} \quad \mu_i = \exp(\beta' \vec{x}_i),
\]
and explain the difference between the two models.

3.6 The R Package caerst provides the data set \texttt{caerst}.

(a) Find a GLM with response \texttt{rent} (net rent in lira) and the explanatory variables \( \texttt{size} \) (size in square meters) and \( \texttt{roomy} \) (number of rooms) that fits the data well. Try several distribution functions like Gaussian and Gamma and try alternative links.
(b) Discuss strategies to select a model from the models fitted in (a).

Chapter 4

Modeling of Binary Data

In Chapter 2 in particular, the logit model is considered as one specific binomial regression model. In this section we will discuss modeling issues for the more general binary regression model
\[
P(y_i = 1 | x_i) = \pi(x_i) = h(a_i^T \beta),
\]
where the response function \( h \) is a fully specified function, which in the case of the logit model is the logistic distribution function \( h(q) = \exp(q)/(1 + \exp(q)) \). A general parametric binary regression model is determined by the link function (the inverse of the response function) and the linear predictor. While the link function determines the functional form of the response probabilities, the linear predictor determines which variables are included and in what form they determine the response. In particular, when categorical and metric variables are present, the linear predictor can, in addition to simple linear terms, contain polynomial versions of continuous variables, dummy variables, and interaction effects. Therefore some care should be taken when specifying constituents of the model like the linear predictor. Statistical reorganization modeling always entails a series of decisions concerning the structuring of the dependence of the response on the predictor. Various aspects are important when making these decisions, among them are the following:

• Discrepancy between data and model. Does the fit of the model support the inferences drawn from the model?
• Relevance of variables and form of the linear predictor. Which variables should be included and how?
• Explanatory power of the covariates.
• Prognostic power of the model.
• Choice of link function. Which link function fits the data well and has a simple interpretation.

Figure 4.1 illustrates aspects of regression modeling and the corresponding evaluation instruments. Since the evaluation of a model starts with parameter estimation, it is at the top of the panel. When estimates have been obtained one can deal with problems concerning the appropriateness of the model, the specification of the predictor, and the obtained explanatory value. Some of the tools that can be used to cope with these problems are given at the bottom of the panel. It should be noted that these aspects are not independent. A model should represent an appropriate approximation of the data when one investigates if the linear predictor may be