# Chapter 1

# Introduction

Categorical data play an important role in many statistical analyses. They appear whenever the outcomes of one or more categorical variables are observed. A categorical variable can be seen as a variable for which the possible values form a set of categories, which can be finite or, in the case of count data, infinite. These categories can be records of answers (yes/no) in a questionnaire, diagnoses like normal/abnormal resulting from a medical examination, or choices of brands in consumer behavior. Data of this type are common in all sciences that use quantitative research tools, for example, social sciences, economics, biology, genetics, and medicine, but also engineering and agriculture.

In some applications all of the observed variables are categorical and the resulting data can be summarized in contingency tables that contain the counts for combinations of possible outcomes. In other applications categorical data are collected together with continuous variables and one may want to investigate the dependence of one or more categorical variables on continuous and/or categorical variables.

The focus of this book is on regression modeling for categorical data. This distinguishes between explanatory variables or predictors and dependent variables. The main objectives are to find a parsimonious model for the dependence, quantify the effects, and potentially predict the outcome when explanatory variables are given. Therefore, the basic problems are the same as for normally distributed response variables. However, due to the nature of categorical data, the solutions differ. For example, it is highly advisable to use a transformation function to link the linear or non-linear predictor to the mean response, to ensure that the mean is from an admissible range. Whenever possible we will embed the modeling approaches into the framework of generalized linear models. Generalized linear models serve as a background model for a major part of the text. They are considered separately in Chapter 3.

In the following we first give some examples to illustrate the regression approach to categorical data analysis. Then we give an overview on the content of this book, followed by an overview on the constituents of structured regression.

## 1.1 Categorical Data: Examples and Basic Concepts

### 1.1.1 Some Examples

The mother of categorical data analysis is the $(2 \times 2)$-contingency table. In the following example data may be given in that simple form.

**Example 1.1: Duration of Unemployment**

The contingency table in Table 2.3 shows data from a study on the duration of employment. Duration

of unemployment is given in two categories, short-term unemployment (less than 6 months) and long-term employment (more than 6 months). Subjects are classified with respect to gender and duration of unemployment. It is quite natural to consider gender as the explanatory variable and duration as the response variable.

TABLE 1.1: Cross-classification of gender and duration of unemployment.

| Gender | Duration | | Total |
|---|---|---|---|
| | $\leq$ 6 months | > 6 months | |
| male | 403 | 167 | 570 |
| female | 238 | 175 | 413 |

□

A simple example with two influential variables, one continuous and the other categorical, is the following.

### Example 1.2: Car in Household

In a sample of $n = 6071$ German households (German socio-economic household panel) various characteristics of households have been collected. Here the response of interest is if a household has at least one car $(y = 1)$ or not $(y = 01)$. Covariates that may be considered influential are income of household in Euros and type of household: (1) one person in household, (2) more than one person with children, (3) more than one person without children). In Figure 1.1 the relative frequencies for having a car are shown for households within intervals of length 50. The picture shows that the link between the probability of owning a car and income is certainly non-linear. □

In many applications the response variable has more than two outcomes, for example, when a customer has to choose between different brands or when the transport mode is chosen. In some applications the response may take ordered response categories.

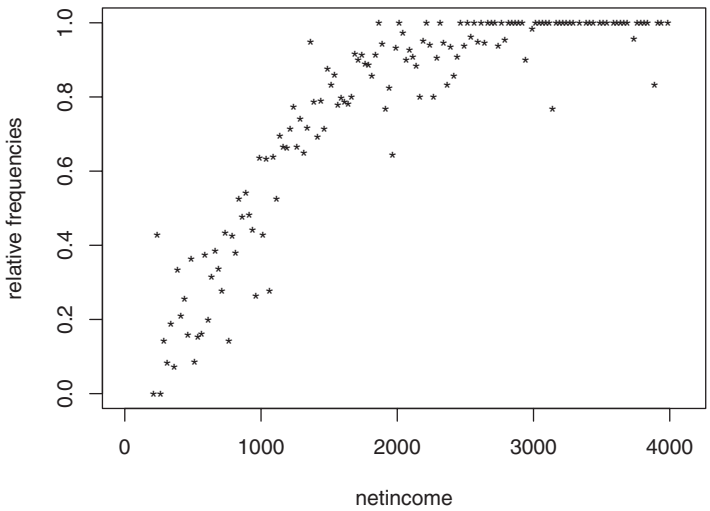


FIGURE 1.1: Car data, relative frequencies within intervals of length 50, plotted against net income in Euros.

### Example 1.3: Travel Mode

Greene (2003) investigated the choice of travel mode of $n = 840$ passengers in Australia. The available travel modes were air, train, bus, and car. Econometricians want to know what determines the choice and study the influence of potential predictor variables as, for example, travel time in vehicle, cost, or household income. □

### Example 1.4: Knee Injuries

In a clinical study focusing on the healing of sports-related injuries of the knee, $n = 127$ patients were treated. By random design, one of two therapies was chosen. In the treatment group an anti-inflammatory spray was used, while in the placebo group a spray without active ingredients was used. After 3, 7, and 10 days of treatment with the spray, the mobility of the knee was investigated in a standardized experiment during which the knee was actively moved by the patient. The pain $Y$ occurring during the movement was assessed on a five-point scale ranging from 1 for no pain to 5 for severe pain. In addition to treatment, the covariate age was measured. A summary of the outcomes for the measurements after 10 days of treatment is given in Table 1.2. The data were provided by Kurt Ulm (IMSE Munich, Germany). □

TABLE 1.2: Cross-classification of pain and treatment for knee data.

| | no pain | | | | severe pain | |
|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | |
| Placebo | 17 | 8 | 14 | 20 | 4 | 63 |
| Treatment | 19 | 26 | 11 | 6 | 2 | 64 |

A specific form of categorical data occurs when the response is given in the form of counts, as in the following examples.

### Example 1.5: Insolvent Companies in Berlin

The number of insolvent firms is an indicator of the economic climate; in particular, the dependence on time is of special interest. Table 1.3 shows the number of insolvent companies in Berlin from 1994 to 1996. □

TABLE 1.3: Number of insolvent companies in Berlin.

| | Month | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Jan. | Feb. | March | April | May | June | July | Aug. | Sep. | Oct. | Nov. | Dec. |
| 1994 | 69 | 70 | 93 | 55 | 73 | 68 | 49 | 97 | 97 | 67 | 72 | 77 |
| 1995 | 80 | 80 | 108 | 70 | 81 | 89 | 80 | 88 | 93 | 80 | 78 | 83 |
| 1996 | 88 | 123 | 108 | 92 | 84 | 89 | 116 | 97 | 102 | 108 | 84 | 73 |

### Example 1.6: Number of Children

There is ongoing research on the birthrates in Western countries. By use of microdata one can try to find the determinants that are responsible for the number of children a woman has during her lifetime. Here we will consider data from the German General Social Survey Allbus, which contains data on all aspects of life in Germany. Interesting predictors, among others, are age, level, and duration of education. □

In some applications the focus is not on the identification and interpretation of the dependence of a response variable on explanatory variables, but on prediction. For categorical responses prediction is also known as classification or pattern recognition. One wants to allocate a new observation into the class it stems from with high accuracy.

Example 1.7: Credit Risk

The aim of credit scoring systems is to identify risk clients. Based on a set of predictors, one wants to distinguish between risk and non-risk clients. A sample of 1000 consumers credit scores collected at a German bank contains 20 predictors, among them duration of credit in months, amount of credit, and payment performance in previous credits. The dataset was published in Fahrmeir and Hamerle (1984), and it is also available from the UCI Machine Learning Repository. □

## 1.1.2 Classification of Variables

The examples illustrate that variables in categorical data analysis come in different types. In the following some classifications of variables are given.

### Scale Levels: Nominal and Ordinal Variables

Variables for which the response categories are qualitative without ordering are called *nominal*. Examples are gender (male/female), choice of brand (brand $A$, ..., brand $K$), color of hair, and nationality. When numbers $1, \ldots, k$ are assigned to the categories, they have to be understood as mere labels. Any one-to-one mapping will do. Statistical analysis should not depend on the ordering, or, more technically, it should be *permutation invariant*.

Frequently the categories of a categorical variable are ordered. Examples are severeness of symptoms (none, mild, moderate, marked) and degree of agreement in questionnaires (strongly disagree, mildly disagree,. . . ,strongly agree). Variables of this type are measured on an ordinal scale level and are often simply called *ordinal*. With reference to the finite number of categories, they are also called *ordered categorical* variables. Statistical analysis may or may not use the ordering. Typically methods that use the ordering of categories allow for more parsimonious modeling, and, since they are using more of the information content in the data, they should be preferred. It should be noted that for ordinal variables there is no distance between categories available. Therefore, when numbers $1, \ldots, k$ are assigned to the categories, only the ordering of these labels may be used, but not the number itself, because it cannot be assumed that the distances are equally spaced.

Variables that are measured on *metric* scale levels (*interval* or *ratio* scale variables) represent measurements for which distances are also meaningful. Examples are duration (seconds, minutes, hours), weight, length, and also number of automobiles in household $(0, 1, 2, \ldots)$. Frequently metric variables are also called *quantitative*, in contrast to nominal variables, which are called *qualitative*. Ordinal variables are somewhat in between. Ordered categorical variables with few categories are sometimes considered as qualitative, although the ordering has some quantitative aspect.

A careful definition and reflection of scale levels is found in particular in the psychology literature. Measuring intelligence is no easy task, so psychologists needed to develop some foundation for their measurements and developed an elaborated mathematical theory of measurement (see, in particular, Krantz et al., 1971).

### Discrete and Continuous Variables

The distinction between discrete and continuous variables is completely unrelated to the concept of scale levels. It refers only to the number of values a variable can take. A *discrete* variable has a finite number of possible values or values that can at least be listed. Thus count data like the number of accidents with possible values from $0, 1, \ldots$ are considered discrete. The possible values of a *continuous* variable form an interval, although, in practice, due to the limitations of measuring instruments, not all of the possible values are observed.

Within the scope of this book discrete data like counts are considered as categorical. In particular, when the mean of a discrete response variable is small it is essential to recognize the discrete nature of the data.

## 1.2 Organization of This Book

The chapters may be grouped into five different units. After a brief review of basic issues in structured regression and classical normal distribution regression within this chapter, in the first unit, consisting of Chapters 2 through 7, the *parametric modeling* of univariate categorical response variables is discussed. In Chapter 2 the basic regression model for binary response, the logit or logistic regression model, is described. Chapter 3 introduces the class of generalized linear models (GLMs) into which the logit model as well as many other models in this book may be embedded. In Chapters 4 and 5 the modeling of binary response data is investigated more closely, including inferential issues but also the structuring of ordered categorical predictors, alternative link functions, and the modeling of overdispersion. Chapter 6 extends the approaches to high-dimensional predictors. The focus is on appropriate regularization methods that allow one to select predictor variables in cases where simple fitting methods fail. Chapter 7 deals with count data as a special case of discrete response.

Chapters 8 and 9 constitute the second unit of the book. They deal with parametric *multinomial response models*. Chapter 8 focuses on unordered multinomial responses, and Chapter 9 discusses models that make use of the order information of the response variable.

The third unit is devoted to *flexible non-linear regression*, also called *non-parametric regression*. Here the data determine the shape of the functional form with much weaker assumptions on the underlying structure. Non-linear smooth regression is the subject of Chapter 10. The modeling approaches are presented as extensions of generalized linear models. One section is devoted to functional data, which are characterized by high-dimensional but structured regressors that often have the form of a continuous signal. Tree-based modeling approaches, which provide an alternative to additive and smooth models, are discussed in Chapter 11. The method is strictly non-parametric and conceptually very simple. By binary recursive partitioning the feature space is partitioned into a set of rectangles, and on each rectangle a simple model is fitted. Instead of obtaining parameter estimates, one obtains a binary tree that visualizes the partitioning of the feature space.

Chapter 12 is devoted to the more traditional topic of *contingency analysis*. The main instrument is the log-linear model, which assumes a Poisson distribution, a multinomial distribution, or a product-multinomial distribution. For Poisson-distributed response there is a strong connection to count data as discussed in Chapter 7, but now all predictors are categorical. When the underlying distribution is multinomial, log-linear models and in particular graphical models are used to investigate the association structure between the categorical variables.

In the fifth unit *multivariate regression models* are examined. Multivariate responses occur if several responses together with explanatory variables are measured on one unit. In particular, repeated measurements that occur in longitudinal studies are an important case. The challenge is to link the responses to the explanatory variables and to account for the correlation between

responses. In Chapter 13, after a brief overview, conditional and marginal models are outlined. Subject-specific modeling in the form of random effects models is considered in Chapter 14.

The last unit, Chapter 15, examines *prediction issues*. For categorical data the problem is strongly related to the common classification problem, where one wants to find the true class from which a new observation stems. Classification problems are basically diagnostic problems with applications in medicine when one wants to identify the type of the disease, in pattern recognition when one aims at recognition of handwritten characters, or in economics when one wants to identify risk clients in credit scoring. In the last decade, in particular, the analysis of genetic data has become an interesting field of application for classification techniques.

## 1.3 Basic Components of Structured Regression

In the following the structuring components of regression are considered from a general point of view but with special emphasis on categorical responses. This section deals with the various assumptions made for the structuring of the independent and the dependent variables.

### 1.3.1 Structured Univariate Regression

Regression methods are concerned with two types of variables, the explanatory (or independent) variables $\boldsymbol{x}$ and the dependent variables $y$. The collection of methods that are referred to as regression methods have several objectives:

- Modeling of the response $y$ given $\boldsymbol{x}$ such that the underlying structure of the influence of $\boldsymbol{x}$ on $y$ is found.
- Quantification of the influence of $\boldsymbol{x}$ on $y$.
- Prediction of $y$ given an observation $\boldsymbol{x}$.

In regression the response variable $y$ is also called the *regressand*, the *dependent variable*, and the *endogeneous variable*. Alternative names for the independent variables $\boldsymbol{x}$ are *regressors, explanatory variables, exogeneous variables, predictor variables*, and *covariates*.

Regression modeling uses several structural components. In particular, it is useful to distinguish between the random component, which usually is specified by some distributional assumption, and the components, which specify the structuring of the covariates $\boldsymbol{x}$. More specifically, in a structured regression the mean $\mu$ (or any other parameter) of the dependent variable $y$ is modeled as a function in $\boldsymbol{x}$ in the form

$$\mu = h(\eta(\boldsymbol{x})),$$

where $h$ is a transformation and $\eta(\boldsymbol{x})$ is a structured term. A very simple form is used in classical linear regression, where one assumes

$$\mu = \beta_0 + x_1\beta_1 + \cdots + x_p\beta_p = \beta_0 + \boldsymbol{x}^T\boldsymbol{\beta}$$

with the parameter vector $\boldsymbol{\beta}^T = (\beta_1, \ldots, \beta_p)$ and the vector of covariates $\boldsymbol{x}^T = (x_1, \ldots, x_p)$. Thus, classical linear regression assumes that the mean $\mu$ is directly linked to a linear predictor $\eta(\boldsymbol{x}) = \beta_0 + \boldsymbol{x}^T\boldsymbol{\beta}$. Covariates determine the mean response by a linear term, and the link $h$ is the identity function. The distributional part in classical linear regression follows from assuming a normal distribution for $y|\boldsymbol{x}$.

In binary regression, when the response takes a value of 0 or 1, the mean corresponds to the probability $P(y = 1|\boldsymbol{x})$. Then the identity link $h$ is a questionable choice since the probabilities

are between 0 and 1. A transformation $h$ that maps $\eta(\boldsymbol{x})$ into the interval $[0, 1]$ typically yields more appropriate models.

In the following, we consider ways of structuring the dependence between the mean and the covariates, with the focus on discrete response data. To keep the structuring parts separated, we will begin with the structural assumption on the response, which usually corresponds to assuming a specific distributional form, and then consider the structuring of the influential term and finish by considering the link between these two components.

## Structuring the Dependent Variable

A common way of modeling the variability of the dependent variable $y$ is to assume a distribution that is appropriate for the data. For binary data with $y \in \{0, 1\}$, the distribution is determined by $\pi = P(y = 1)$. As special case of the binomial distribution it is abbreviated by $B(1, \pi)$. For count data $y \in \{0, 1, 2, \dots\}$, the Poisson distribution $P(\lambda)$ with mass function $f(x) = \lambda^x e^{-\lambda}/x!$, $x = 0, 1, \dots$ is often a good choice. An alternative is the negative binomial distribution, which is more flexible than the Poisson distribution. If $y$ is continuous, a common assumption is the normal distribution. However, it is less appropriate if the response is some duration for which $y \geq 0$ has to hold. Then, for example, a Gamma-distribution $\Gamma(\nu, \alpha)$ that has positive support might be more appropriate. In summary, the choice of the distributional model mainly depends on the kind of response that is to be modeled. Figures 1.2 and 1.3 show several discrete and continuous distributions, which may be assumed. Each panel shows two distributions that can be thought of as referring to two distinct values of covariates. For the normal distribution model where only the mean depends on covariates, the distributions referring to different values of covariates are simply shifted versions of each other. This is quite different for response distributions like the Poisson or the Bernoulli distribution. Here the change of the mean, caused by different values of covariates, also changes the shape of the distribution. This phenomenon is not restricted to discrete distributions but is typically found when responses are discrete.

Sometimes the assumption of a specific distribution, even if it reflects the type of data collected, is too strong to explain the variability in responses satisfactorily. In practice, one often finds that count data and relative frequencies are more variable than is to be expected under the Poisson and the binomial distributions. The data show *overdispersion*. Consequently, the structuring of the responses should be weakened by taking overdispersion into account.

One step further, one may even drop the assumption of a specific distribution. Instead of assuming a binomial or a Poisson distribution, one only postulates that the link between the mean and a structured term, which contains the explanatory variables, is correctly specified. In addition, one can specify how the variance of the response depends on explanatory variables. The essential point is that the assumptions on the response are very weak, within quasi-likelihood approaches structuring of the response in the form of distributional assumptions is not necessary.

## Structuring the Influential Term

It is tempting to postulate no structure at all by allowing $\eta(\boldsymbol{x})$ to be any function. What works in the unidimensional case has severe drawbacks if $\boldsymbol{x}^T = (x_1, \dots, x_p)$ contains many variables. It is hard to explain how a covariate $x_j$ determines the response if no structure is assumed. Moreover, estimation becomes difficult and less robust. Thus often it is necessary to assume some structure to obtain an approximation to the underlying functional form that works in practice. Structural assumptions on the predictor can be strict or more flexible, with the degree of flexibility depending on the scaling of the predictor.

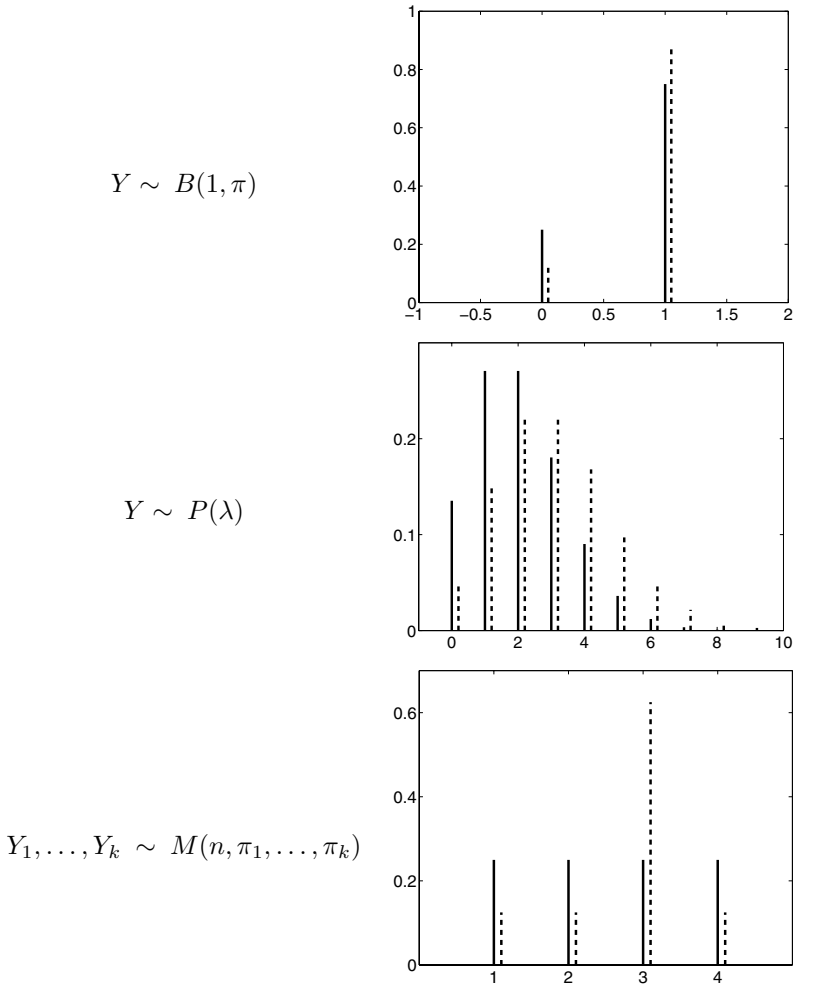


FIGURE 1.2: Binomial, Poisson, and multinomial distributions. Each panel shows two different distributions.

### Linear Predictor

The most common form is the linear structure

$$\eta(\boldsymbol{x}) = \beta_0 + \boldsymbol{x}^T\boldsymbol{\beta},$$

which is very robust and allows simple interpretation of the parameters. Often it is necessary to include some interaction terms, for example, by assuming

$$\begin{aligned}\eta(\boldsymbol{x}) &= \beta_0 + x_1\beta_1 + \cdots + x_p\beta_p + x_1x_2\beta_{12} + x_1x_3\beta_{13} + \cdots + x_1x_2x_3\beta_{123} \\ &= \boldsymbol{z}^T\boldsymbol{\beta}.\end{aligned}$$

By considering $\boldsymbol{z}^T = (1, x_1, \ldots, x_p, x_1x_2, \ldots, x_1x_2x_3, \ldots)$ as variables, one retains the linear structure. For estimating and testing (not for interpreting) it is only essential that the structure is linear in the parameters. When explanatory variables are quantitative, interpreting the parameters is straightforward, especially in the linear model without interaction terms.

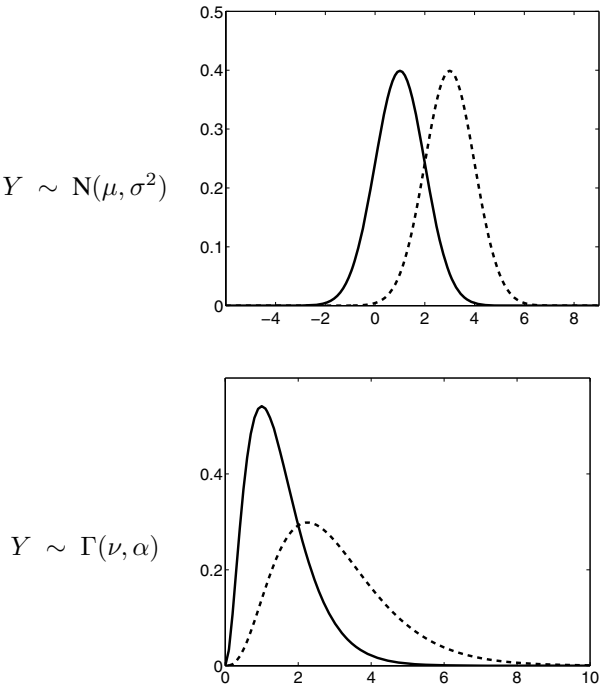


FIGURE 1.3: Normal and Gamma-distributions.

### Categorical Explanatory Variables

Categorical explanatory variables, also called factors, take values from a finite set $1, \ldots, k$, with the numbers representing the factor levels. They cannot be used directly within the linear predictor because one would falsely assume fixed ordering of the categories with the distances between categories being meaningful. That is not the case for nominal variables, not even for ordered categorical variables. Therefore, specific structuring is needed for factors. Common structuring uses dummy variables and again yields a linear predictor. The coding scheme depends on the intended use and on the scaling of the variable. Several coding schemes and corresponding interpretations of effects are given in detail in Section 1.4.1. The handling of ordered categorical predictors is also considered in Section 4.4.3.

When a categorical variable has many categories, the question arises of which categories can be distinguished with respect to the response. Should categories be collapsed, and if so, which ones? The answer depends on the scale level. While for nominal variables, for which categories have no ordering, any fusion categories seems sensible, for ordinal predictors collapsing means fusing adjacent categories. Figure 1.4 shows a simple application. It shows the effect of the urban district and the year of construction on the rent per square meter in Munich. Urban district is a nominal variable that has 25 categories, year of construction is an ordered predictor, where categories are defined by decades. The coefficient paths in Figure 1.4 show how, depending on a tuning parameter, urban districts and decades are combined. It turns out that only 10 districts are really different, and the year of construction can be combined into 8 distinct categories (see also Section 6.5).

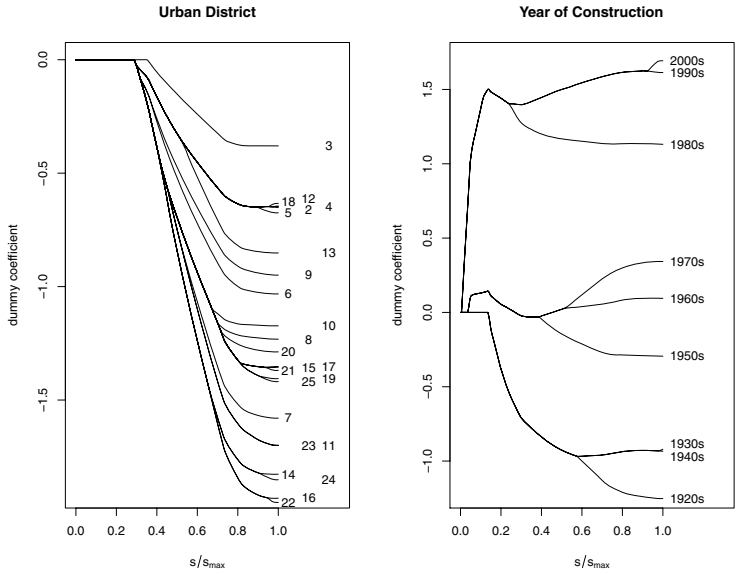


FIGURE 1.4: Effects of urban district and year of construction (in decades) on rent per square meter.

### Additive Predictor

For quantitative explanatory variables, a less restrictive assumption is

$$\eta(\boldsymbol{x}) = f_{(1)}(x_1) + \cdots + f_{(p)}(x_p),$$

where $f_{(j)}(x_j)$ are unspecified functions. Thus one retains the additive form, which still allows simple interpretation of the functions $f_{(j)}$ by plotting estimates but the approach is much less restrictive than in the linear predictor. An extension is the inclusion of unspecified interactions, for example, by allowing

$$\eta(\boldsymbol{x}) = f_{(1)}(x_1) + \cdots + f_{(p)(x_p)} + f_{(13)}(x_1, x_3),$$

where $f_{(13)}(x_1, x_3)$ is a function depending on $x_1$ and $x_3$.

For categorical variables no function is needed because only discrete values occur. Thus, when, in addition to quantitative variables, $x_1, \ldots, x_p$, categorical covariates are available, they are included in an additional linear term, $\boldsymbol{z}^T\boldsymbol{\gamma}$, which is built from dummy variables. Then one uses the *partial linear predictor*

$$\eta(\boldsymbol{x}) = f_{(1)}(x_1) + \cdots + f_{(p)}(x_p) + \boldsymbol{\gamma}.$$

### Additive Structure with Effect Modifiers

If the effect of a covariate, say gender $(x_1)$, depends on age $(x_2)$ instead of postulating an interaction model of the form $\eta = \beta_0 + x_1\beta_1 + x_2\beta_2 + x_1x_2\beta_{12}$, a more flexible model is given by

$$\eta = \beta_2(x_2) + x_1\beta_{12}(x_2),$$

where $\beta_2(x_2)$ is the smooth effect of age and $\beta_{12}(x_2)$ is the effect of gender $(x_1)$, which is allowed to vary over age. Both functions $\beta_2(.)$ and $\beta_{12}(.)$ are unspecified and the data determine their actual form.

### Tree-Based Methods

An alternative way to model interactions of covariates is the recursive partitioning of the predictor space into sets of rectangles. The most popular method, called CART for classification and regression trees, constructs for metric predictors partitions of the form $\{x_1 \leq c_1\} \cap \cdots \cap \{x_m \leq c_m\}$, where $c_1, \ldots, c_m$ are split-points from the regions of the variables $x_1, \ldots, x_m$. The splits are constructed successively beginning with a first split, producing, for example, $\{x_1 \leq c_1\}$, $\{x_1 > c_1\}$. Then these regions are split further. The recursive construction scheme allows us to present the resulting partition in a tree. A simple example is the tree given in Figure 1.5, where two variables are successively split by using split-points $c_1, \ldots, c_4$. The first split means that the dichotomization into $\{x_1 \leq c_1\}$ and $\{x_1 > c_1\}$ is of major importance for the prediction of the outcome. Finer prediction rules are obtained by using additional splits, for example, the split of the region $\{x_1 \leq c_1\}$ into $\{x_2 \leq c_2\}$ and $\{x_2 > c_2\}$. The big advantage of trees is that they are easy to interpret and the visualization makes it easy to communicate the underlying structure to practitioners.

## The Link between Covariates and Response

Classical linear regression assumes $\mu = \eta(\boldsymbol{x})$ with $\eta(\boldsymbol{x}) = \boldsymbol{x}^T\boldsymbol{\beta}$. For binary regression models, the more general form $\mu = h(\eta(\boldsymbol{x}))$ is usually more appropriate, since $h$ may be chosen such that $\mu$ takes values in the unit interval $[0, 1]$. Typically $h$ is chosen as a distribution function, for example, the logistic distribution function $h(\eta) = \exp(\eta)/(1 + \exp(\eta))$ or the normal distribution function. The corresponding models are the so-called logit and probit models. However, any distribution function that is strictly monotone may be used as a response function (see Section 5.1).



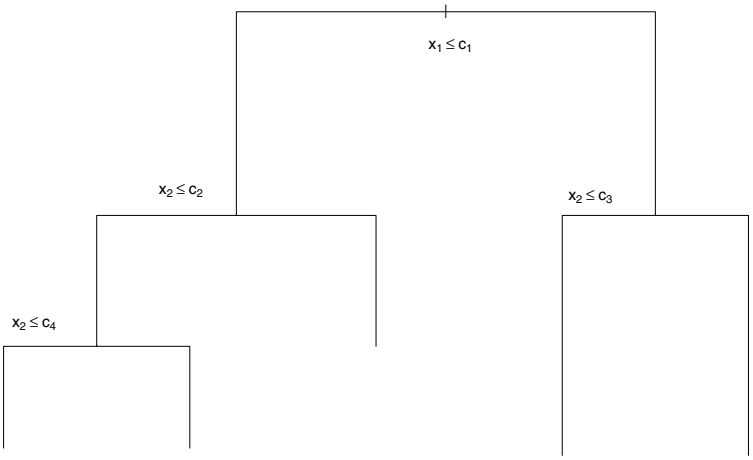


FIGURE 1.5: Example tree for two variables.

In many applications $h$ is considered as known. Then, of course, there is the danger of misspecification. It is often more appropriate to consider alternative transformation functions and choose the one that yields the best fit. Alternatively, one can estimate the transformation itself, and therefore let the data determine the form of the transformation function (Section 5.2).

### 1.3.2 Structured Multicategorical Regression

When the response is restricted to a fixed set of possible values, the so-called response categories, the typical assumption for the distribution is the multinomial distribution. When $1, \ldots, k$ denote the response categories of variable $Y$, the multinomial distribution specifies the probabilities $\pi_1(\boldsymbol{x}), \ldots, \pi_k(\boldsymbol{x})$, where $\pi_r(\boldsymbol{x}) = P(Y = r|\boldsymbol{x})$.

The simple structure of univariate regression models is no longer appropriate because the response is multivariate. One has to model the dependence of all the probabilities $\pi_1(\boldsymbol{x}), \ldots, \pi_k(\boldsymbol{x})$ on the explanatory variables. This may be accomplished by a multivariate model that has the basic structure

$$\pi_r(\boldsymbol{x}) = h_r(\eta_1(\boldsymbol{x}), \ldots, \eta_k(\boldsymbol{x})), r = 1, \ldots, k-1,$$

where $h_r, r = 1, \ldots, k-1$, are transformation functions that are specific for the category. Since probabilities sum up to one, it is sufficient to specify $k-1$ of the $k$ components. By using the $(k-1)$-dimensional vectors $\boldsymbol{\pi}(\boldsymbol{x})^T = (\pi_1(\boldsymbol{x}), \ldots, \pi_{k-1}(\boldsymbol{x}))$, $\boldsymbol{\eta}(\boldsymbol{x}))^T = (\eta_1(\boldsymbol{x}), \ldots, \eta_{k-1}(\boldsymbol{x}))$, models have the closed form

$$\boldsymbol{\pi}(\boldsymbol{x}) = h(\boldsymbol{\eta}(\boldsymbol{x})).$$

The choice of the transformation function depends on the scale level of the response. If the response is nominal, for example, when modeling the choice of different brands or the choice of transport mode, other response functions are more appropriate than in the ordinal case, when the response is given on a rating scale with categories like very good, good, fair, poor, and very poor (see Chapter 8 for nominal and Chapter 9 for ordinal responses).

The structuring of the predictor functions is in analogy to univariate responses. Strict linear structures assume

$$\eta_r(\boldsymbol{x}) = \boldsymbol{x}^T\boldsymbol{\beta}_r,$$

where the parameter vector depends on the category $r$. By defining an appropriate design matrix, one obtains a multivariate generalized linear model form $\boldsymbol{\pi}(\boldsymbol{x}) = h(\boldsymbol{X}\boldsymbol{\beta})$. More flexible predictors use additive or partial additive structures of the form

$$\eta_r(\boldsymbol{x}) = f_{(r1)}(x_1) + \cdots + f_{(rp)}(x_p),$$

with functions depending on the category.

### 1.3.3 Multivariate Regression

In many studies several response variables are observed for each unit. The responses may refer to different variables or to the same measurements that are observed repeatedly. The latter case is found in particular in longitudinal studies where measurements on an individual are observed at several times under possibly varying conditions. In both cases the response is a multivariate represented by a vector $\boldsymbol{y}_i^T = (y_{i1}, \ldots, y_{im})$, which collects the measurements on unit $i$. Since measurements taken on one unit or cluster tend to be more similar, one has to assume that in general the measurements are correlated.

## Structuring the Dependent Variables

The structuring of the vector of dependent variables by assuming an appropriate distribution is not as straightforward as in classical multivariate regression, where a multivariate normal distribution is assumed. Multivariate, normally distributed responses have been extensively investigated for a long time. They are simply structured with a clear separation of the mean and correlation structures, which are sufficient to define the distribution.

When the marginals, that is, single responses $y_{it}$, are discrete it is harder to find a sparse representation of the total response vector. Although the form of the distribution may have a simple form, the number of parameters can be extremely high. If, for example, the marginals are binary with $y_{it} \in \{0, 1\}$, the total distribution is multinomial but determined by $2^m$ probabilities for the various combinations of outcomes. With $m = 10$ measurements, the number of parameters is 1024. Simple measures like the mean of the marginals and the correlations between components do not describe the distribution sufficiently.

One strategy that is used in marginal modeling is to model the mean structure of the marginals, which uses univariate regression models, and in addition specify an association structure between components that does not have to be the correct association structure. An alternative approach uses random effects. One assumes that the components are uncorrelated given a fixed but unobserved latent variable, which is shared by the measurements within one unit or cluster. In both cases one basically uses parameterizations of the discrete marginal distributions.

## Structuring the Influential Term

The structuring of the influential term is more complex than in univariate response models because covariates may vary across measurements within one cluster. For example, in longitudinal studies, where the components refer to repeated measurements across time, the covariates that are to be included may also vary across time. Although the specification is more complex, in principle, the same form of predictors as in univariate regression applies. One can use linear terms or more flexible additive terms. However, new structuring elements are useful, and two of them are the following.

In random effects models one models explicitly the heterogeneity of clustered responses by assuming cluster-specific random effects. For example, the binary response of observation $t$ on cluster $i$ with covariate vector $\boldsymbol{x}_{it}$ at measurement $t$ may be modeled by

$$P(y_{it} = 1|\boldsymbol{x}_{it}, b_i) = h(\eta_{it}), \quad \eta_{it} = b_i + \boldsymbol{x}_{it}^T\boldsymbol{\beta}.$$

While $\boldsymbol{\beta}$ is a fixed effect that is common to all clusters, each cluster has its own cluster-specific random effect $b_i$, which models the heterogeneity across clusters. More generally, one can assume that not only the intercept but also the slopes of the variables are cluster-specific. Then one has to specify which components have category-specific effects. Moreover, one has to specify a distribution for these category-specific effects. Thus, some additional structuring and therefore decision making by the modeler is needed.

Another effect structure that can be useful in repeated measurements is the variation of effect strength across time, which can be modeled by letting the parameter depend on time:

$$\eta_{it} = b_i + \boldsymbol{x}_{it}^T\boldsymbol{\beta}_t.$$

Marginal modeling approaches are given in detail in Chapter 13, and random effects models are found in Chapter 14.

### 1.3.4 Statistical Modeling

Statistical modelling refers to the process of using models to extract information from data. In statistics that means, in particular, to separate the systematic effects or underlying patterns from the random effects. A model, together with its estimation method, can be seen as a measuring instrument. Like magnifying glasses and telescopes serve as instruments to uncover structures not seen to the unarmed eye, a model allows one to detect patterns that are in the data but not seen without the instrument. What is seen depends on the instrument. Only those patterns are found for which the instrument is sensitive. For example, linear models allow one to detect linear structures. If the underlying pattern is non-linear, they fail and the results can be very misleading. Or, effect modifiers are detected only if the model allows for them. In that sense the model determines what is found. More flexible models allow one to see more complex structures, at least if reliable estimation methods are found. In the same way as a telescope depends on basic conditions like the available amount of light, statistical models depend on basic conditions like the sample size and the strength of the underlying effects. Weak patterns typically can be detected only if much information is available.

The use and choice of models is guided by different and partly contradictory objectives. Models should be simple but should account for the complexity of the underlying structure. Simplicity is strongly connected to interpretability. Users of statistical models mostly prefer interpretable models over black boxes. In addition, the detection of interpretable and simple patterns and therefore understanding is the essential task of science. In science, models often serve to understand and test subject-matter hypotheses about underlying processes.

The use of models, parametric or more flexible, is based on the assumption that a stochastic data model has generated the data. The statistician is expected to use models that closely approximate the data driving the model, quantify the effects within the model, and account for the estimation error. Ideally, the analysis also accounts for the closeness of the model to the data-generating model, for example, in the form of goodness-of-fit tests.

A quite different objective that may determine the choice of the model is the exactness of the prediction. One wants to use that model that will give the best results in terms of prediction error when used on future data. Then black box machines, which do not try to uncover latent structures, also apply and may show excellent prediction results. Ensemble methods like random forests or neural networks work in that way. Breiman (2001b) calls them *algorithmic models* in contrast to *data models*, which assume that a stochastic data-generating model is behind the data. Algorithmic models are less models than an approach to find good prediction rules by designing algorithms that link the predictors to the responses. Especially in the machine learning community, where the handling of data is guided by a more pragmatic view, algorithms with excellent prediction properties in the form of black boxes are abundant.

What type of model is to be preferred depends mainly on the objective of the scientific question. If prediction is the focus, intelligently designed algorithms may serve the purpose well. If the focus is on understanding and interpretation, data models are to be preferred. The choice of the model depends on the structures that are of interest to the user and circumstances like sample size and strength of the parameters. When effects are weak and the sample size is small, simple parametric models will often be more stable than models that allow for complex patterns. As a model does not fit all datasets, for a single dataset different models that uncover different structures may be appropriate. Typically there is not a single best model for a set of data.

In the following chapters we will predominantly consider tools for data models; that is alternative models, estimation procedures, and diagnostic tools will be discussed. In the last chapter, where prediction is the main issue, algorithmic models/methods will also be included. Specification of models is treated throughout the book in various forms, ranging from classical

test procedures to examine parameters to regularization techniques that allow one to select variables or link functions.

There is a rich literature on model selection. Burnham and Anderson (2002) give an extensive account of the information-theoretic approach to model selection; see also Claeskens and Hjort (2008) for a survey on the research in the field. The alternative modeling cultures, data modeling versus algorithmic modeling, was treated in a stimulating article by Breiman (2001b). His strong opinion on stochastic data models is worth reading, in particular together with the included and critical discussion.

## 1.4 Classical Linear Regression

Since the linear regression model is helpful as a background model, in this section a brief overview of classical linear regression is given. The section may be skipped if one feels familiar with the model. It is by no means a substitute for a thorough introduction to Gaussian response models, but a reminder of the basic concepts. Parametric regression models including linear models are discussed in detail in many statistics books, for example, Cook and Weisberg (1982), Ryan (1997), Harrell (2001), and Fahrmeir et al. (2011).

The basic multiple linear regression model is often given in the form

$$y = \beta_0 + x_1\beta_1 + \cdots + x_p\beta_p + \varepsilon,$$

where $\varepsilon$ is a noise variable that fulfills $\mathrm{E}(\varepsilon) = 0$. Thus, for given $x_1, \ldots, x_p$ the expectation of the response variable $\mu = \mathrm{E}(y|x_1, \ldots, x_p)$ is specified as a linear combination of the explanatory variables $x_1, \ldots, x_p$:

$$\mu = \beta_0 + x_1\beta_1 + \cdots + x_p\beta_p = \beta_0 + \boldsymbol{x}^T\boldsymbol{\beta},$$

with the parameter vector $\boldsymbol{\beta}^T = (\beta_1, \ldots, \beta_p)$ and vector of covariates $\boldsymbol{x}^T = (x_1, \ldots, x_p)$.

### 1.4.1 Interpretation and Coding of Covariates

When interpreting the parameters of linear models it is useful to distinguish between quantitative (metrically scaled) covariates and categorical covariates, also called factors, which are measured on a nominal or ordinal scale level.

#### Quantitative Explanatory Variables

For a quantitative covariate like age, the linear model has the form

$$\mathrm{E}(y|x) = \beta_0 + x\beta_A = \beta_0 + \text{Age} * \beta_A.$$

From $\mathrm{E}(y|x+1) - \mathrm{E}(y|x) = \beta_A$ it is immediately seen that $\beta_A$ reflects the change of the mean response if the covariate is increased by one unit. If the response is income in dollars and the covariate age is given in years, the units of $\beta_A$ are dollars per year and $\beta_A$ is the change of expected income resulting from increasing the age by one year.

#### Binary Explanatory Variables

A binary covariate like gender (G) has to be coded, for example, as a (0-1)-variable in the form

$$x_G = \begin{cases} 1 & \text{male} \\ 0 & \text{female.} \end{cases}$$

Then the interpretation is the same as for quantitative variables. If the response is income in dollars, the parameter $\beta_G$ in

$$\mathrm{E}(y|x) = \beta_0 + x_G\beta_G$$

represents the change in mean income if $x_G$ is increased by one unit. Since $x_G$ has only two values, $\beta_G$ is equivalent to the increase or decrease of the mean response resulting from the transition from $x_G = 0$ to $x_G = 1$. Thus $\beta_G$ is the difference in mean income between men and women.

### Multicategorical Explanatory Variables or Factors

Let a covariate have $k$ possible categories and categories just reflect labels. This means that the covariate is measured on a nominal scale level. Often covariates of this type are called factors or factor variables. A simple example is in the medical profession $P$, where the possible categories gynaecologist, dermatologist, and so on, are coded as numbers $1, \ldots, k$. When modeling a response like income, a linear term $P * \beta_P$ will produce nonsense since it assumes a linear relationship between the response and the values of $P \in \{1, \ldots, k\}$ , but the coding for profession by numbers is arbitrary. To obtain parameters that are meaningful and have simple interpretations one defines dummy variables. One possibility is dummy or (0-1)-coding.

**(0-1)-Coding of $P \in \{1, \ldots, k\}$**

$$x_{P(j)} = \begin{cases} 1 & \text{if } P = j \\ 0 & \text{otherwise} \end{cases}$$

When an intercept is in the model only $k-1$ variables can be used. Otherwise, one would have too many parameters that would not be identifiable. Therefore, one dummy variable is omitted and the corresponding category is considered the *reference category*. When one chooses $k$ as the reference category the linear predictor is determined by the first $k-1$ dummy variables:

$$\mathrm{E}(y|P) = \beta_0 + x_{P(1)}\beta_{P(1)} + \ldots + x_{P(k-1)}\beta_{P(k-1)}. \tag{1.1}$$

Interpretation of the parameters follows directly from considering the response for different values of $P$:

$$\mathrm{E}(y|P = i) = \beta_0 + \beta_{P(i)}, \quad i = 1, \ldots k-1, \quad \mathrm{E}(y|P = k) = \beta_0.$$

$\beta_0$ is the mean for the reference category $k$ and $\beta_{P(i)}$ is the increase or decrease of the mean response in comparison to the reference category $k$. Of course any category can be used as the reference. Thus, for a categorial variable, $k-1$ functionally independent dummy variables are introduced. A particular choice of the reference category determines the set of variables, or in the terminology of analysis of variance, the set of contrasts. The use of all $k$ dummy variables results in overparameterization because they are not functionally independent. The sum over all $k$ dummy variables yields 1, and therefore $\beta_0$ would not be identifiable.

An alternative coding scheme is effect coding, where categories are treated in a symmetric way.

**Effect Coding of $P \in \{1, \ldots, k\}$**

$$x_{P(j)} = \begin{cases} 1 & \text{if } P = j \\ -1 & \text{if } P = k, \quad j = 1, \ldots, k-1 \\ 0 & \text{otherwise} \end{cases}$$

The linear predictor (1.1) now yields

$$\begin{aligned} \mathrm{E}(y|P = i) &= \beta_0 + \beta_{P(i)} \ , \ i = 1, \ldots, k-1, \\ \mathrm{E}(y|P = k) &= \beta_0 - \beta_{P(1)} - \ldots - \beta_{P(k-1)}. \end{aligned}$$

It is easily seen that

$$\beta_0 = \frac{1}{k} \sum_{j=1}^{k} \mathrm{E}(y|P = j)$$

is the average response across the categories and

$$\beta_{P(j)} = \mathrm{E}(y|P = j) - \beta_0$$

is the deviation of category $j$ from the average response level given by $\beta_0$. Although only $k-1$ dummy variables are used, interpretation does not refer to a particular category. There is no reference category as in the case of (0-1)-coding. For the simple example of four categories one obtains

$$\begin{array}{cc|ccc} & & x_{P(1)} & x_{P(2)} & x_{P(3)} \\ & 1 & 1 & 0 & 0 \\ P & 2 & 0 & 1 & 0 \\ & 3 & 0 & 0 & 1 \\ & 4 & -1 & -1 & -1 \end{array} .$$

Alternative coding schemes are helpful, for example, if the categories are ordered. Then a split that distinguishes between categories below and above a certain level often reflects the ordering in a better way.

**Split-Coding of $P \in \{1, \ldots, k\}$**

$$x_{P(j)} = \begin{cases} 1 & \text{if } P > j \\ 0 & \text{otherwise} \end{cases} \qquad j = 1, \ldots, k-1$$

With split-coding, the model $\mathrm{E}(y|P) = \beta_0 + x_{P(1)}\beta_{P(1)} + \ldots + x_{P(k-1)}\beta_{P(k-1)}$ yields

$$\begin{aligned} \mathrm{E}(y|P = 1) &= \beta_0, \\ \mathrm{E}(y|P = i) &= \beta_0 + \beta_{P(1)} + \cdots + \beta_{P(i-1)} \ , \ i = 2, \ldots, k \end{aligned}$$

and therefore the coefficients

$$\begin{aligned}\beta_0 &= \mathrm{E}(y|P=1),\\ \beta_{P(i)} &= \mathrm{E}(y|P=i+1) - \mathrm{E}(y|P=i)\ ,\ i=1,\ldots,k-1.\end{aligned}$$

Thus the coefficients $\beta_{P(i)}, i=1,\ldots,k-1$ represent the difference in expectation when the factor level increases from $i$ to $i+1$. They may be seen as the stepwise change over categories given in the fixed order $1,\ldots,k$, with 1 serving as the reference category where the process starts. In contrast to (0-1)-coding and effect coding, split-coding uses the ordering of categories: $\beta_{P(1)} + \cdots + \beta_{P(i-1)}$ can be interpreted as the change in expectation for the transition from category 1 to category $i$ with intermediate categories $2,3,\ldots,i-1$.

For further coding schemes see, for example, Chambers and Hastie (1992). The structuring of the linear predictor is examined in more detail in Chapter 4, where models with binary responses are considered.

### 1.4.2 Linear Regression in Matrix Notation

Let the observations be given by $(y_i, x_{i1}, \ldots, x_{ip})$ for $i=1,\ldots,n$, where $y_i$ is the response and $x_{i1},\ldots,x_{ip}$ are the given covariates. The model takes the form

$$y_i = \beta_0 + x_{i1}\beta_1 + \cdots + x_{ip}\beta_p + \varepsilon_i,\ i=1,\ldots n.$$

With $\boldsymbol{x}_i^T = (1, x_{i1}, \ldots x_{ip}),\ \boldsymbol{\beta}^T = (\beta_0, \beta_1, \ldots \beta_p)$ it may be written more compact as

$$y_i = \boldsymbol{x}_i^T\boldsymbol{\beta} + \varepsilon_i,$$

where $\boldsymbol{x}_i$ has length $\tilde{p} = p+1$. In matrix notation one obtains

$$\begin{bmatrix} y_1 \\ \vdots \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & \ldots & x_{1p} \\ 1 & x_{21} & \ldots & x_{2p} \\ \vdots & & & \\ 1 & x_{n1} & \ldots & x_{np} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

or simply

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

where

$\boldsymbol{y}^T = (y_1, \ldots, y_n)$ is the vector of responses;

$\boldsymbol{X}$ is the design matrix, which is composed from the explanatory variables;

$\boldsymbol{\beta}^T = (\beta_0, \ldots, \beta_p)$ is the $(p+1)$-dimensional parameter vector;

$\boldsymbol{\varepsilon}^T = (\varepsilon_1, \ldots, \varepsilon_n)$ is the vector of errors.

Common assumptions are that the errors have expectation zero, $\mathrm{E}(\varepsilon_i) = 0,\ i=1,\ldots,n$; the variance of each error component is given by $\mathrm{var}(\varepsilon_i) = \sigma^2,\ i=1,\ldots,n$, called *homoscedasticity* or *homogeneity*; and the error components from different observations are uncorrelated, $\mathrm{cov}(\varepsilon_i, \varepsilon_j) = 0,\ i \neq j$.

Multiple Linear Regression

$$y_i = \boldsymbol{x}_i^T\boldsymbol{\beta} + \varepsilon_i \quad \text{or} \quad \boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

Assumptions:

$$\mathrm{E}(\varepsilon_i) = 0, \quad \mathrm{var}(\varepsilon_i) = \sigma^2, \quad \mathrm{cov}(\varepsilon_i, \varepsilon_j) = 0, i \neq j$$

In matrix notation the assumption of homogeneous variances may be condensed into $\mathrm{E}(\boldsymbol{\varepsilon}) = \mathbf{0}$, $\mathrm{cov}(\boldsymbol{\varepsilon}) = \sigma^2\boldsymbol{I}$. If one assumes in addition that responses are normally distributed, one postulates $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2\boldsymbol{I})$. It is easy to show that the assumptions carry over to the observable variables $y_i$ in the form

$$\mathrm{E}(\boldsymbol{y}|\boldsymbol{X}) = \boldsymbol{X}\boldsymbol{\beta}, \quad \boldsymbol{y} \sim N(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{I}).$$

The last representation may be seen as a structured representation of the classical linear model, which gives the distributional and systematic components separately without using the noise variable $\varepsilon$. With $\boldsymbol{\mu}$ denoting the mean response vector with components $\mu_i = \boldsymbol{x}_i^T\boldsymbol{\beta}$ one has the following form.

Multiple Linear Regression with Normally Distributed Errors

$$\boldsymbol{\mu} = \boldsymbol{X}\boldsymbol{\beta}, \quad \boldsymbol{y} \sim N(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{I})$$

The separation of the structural and distributional components is an essential feature that forms the basis for the extension to more general models considered in Chapter 3.

### 1.4.3 Estimation

#### Least-Squares Estimation

A simple criterion to obtain estimates of the unknown parameter $\boldsymbol{\beta}$ is the least-squares criterion, which minimizes

$$Q(\boldsymbol{\beta}) = \sum_{i=1}^{n}(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta})^2.$$

Thus the parameter $\hat{\boldsymbol{\beta}}$, which minimizes $Q(\boldsymbol{\beta})$, is the parameter that minimizes the squared distance between the actual observation $y_i$ and the predicted value $\hat{y}_i = \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}$. The choice of the squared distance has the advantage that an explicit form of the estimate is available. It means that large discrepancies between $y_i$ and $\hat{y}_i$ are taken more seriously than for the Euclidean distance $|y_i - \hat{y}_i|$, which would be an alternative criterion to minimize. Simple calculation shows that the derivative of $Q(\boldsymbol{\beta})$ has the form

$$\frac{\partial Q(\boldsymbol{\beta})}{\partial \beta_s} = \sum_{i=1}^{n} 2(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta})x_{is},$$

$s = 0, \ldots, p$, where $x_{i0} = 1$. A minimum can be expected if the derivative equals zero. Thus the least-squares estimate has to fulfill the equation $\sum_i x_{is}(y_i - \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}) = 0$. In vector notation

one obtains the form

$$\sum_{i=1}^{n} \boldsymbol{x}_i y_i = \sum_i \boldsymbol{x}_i \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}},$$

which may be written in the form of the normal equation $\boldsymbol{X}^T \boldsymbol{y} = \boldsymbol{X}^T \boldsymbol{X} \hat{\boldsymbol{\beta}}$. Assuming that the inverse of $\boldsymbol{X}^T \boldsymbol{X}$ exists, an explicit solution is given by

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

## Maximum Likelihood Estimation

The least-squares estimate is strongly connected to the maximum likelihood (ML) estimate if one assumes that the error is normally distributed. The normal distribution contains a quadratic term. Thus it is not surprising that the ML estimate is equivalent to minimizing squared distances. If one assumes $\varepsilon_i \sim N(0, \sigma^2)$ or, equivalently, $y_i \sim N(\boldsymbol{x}_i^T \boldsymbol{\beta}, \sigma^2)$, the conditional likelihood (given $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$) is given by

$$L(\boldsymbol{\beta}, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-(y_i - \boldsymbol{x}_i^T \boldsymbol{\beta})^2/(2\sigma^2)).$$

The corresponding log-likelihood has the form

$$\begin{aligned} l(\boldsymbol{\beta}, \sigma^2) &= -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \boldsymbol{x}_i^T \boldsymbol{\beta})^2 - \frac{n}{2} \log(2\pi) - \log(\sigma^2) \\ &= -\frac{1}{2\sigma^2} Q(\boldsymbol{\beta}) - \frac{n}{2} \log(2\pi) - n \log(\sigma^2). \end{aligned}$$

As far as $\boldsymbol{\beta}$ is concerned, maximization of the log-likelihood is equivalent to minimizing the squared distances $Q(\boldsymbol{\beta})$. Simple derivation shows that maximization of $l(\boldsymbol{\beta}, \sigma^2)$ with respect to $\sigma^2$ yields

$$\hat{\sigma}^2_{ML} = \frac{1}{n} \sum_{i=1}^{n} (y_i - \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}})^2.$$

It is noteworthy that the maximum likelihood estimate $\hat{\boldsymbol{\beta}}$ (which is equivalent to the least-squares estimate) does not depend on $\sigma^2$. Thus the parameter $\boldsymbol{\beta}$ is estimated without reference to the variability of the response.

## Properties of Estimates

A disadvantage of the ML estimate $\hat{\sigma}^2_{ML}$ is that it underestimates the variance $\sigma^2$. An unbiased estimate is given by

$$\hat{\sigma}^2 = \frac{1}{n - (p+1)} \sum_{i=1}^{n} (y_i - \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}})^2,$$

where the correction in the denominator reflects the number of estimated parameters in $\hat{\boldsymbol{\beta}}$, which is $p + 1$ since an intercept is included. The essential properties of estimates are given in the so-called Gauss-Markov theorem. Assuming for all observations $\mathrm{E}(\varepsilon_i) = 0$, $\mathrm{var}(\varepsilon_i) = \sigma^2$, $\mathrm{cov}(\varepsilon_i, \varepsilon_j) = 0, i \neq j$, one obtains

(1) $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}^2$ are unbiased, that is, $\mathrm{E}(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}, \mathrm{E}(\hat{\sigma}^2) = \sigma^2$.

(2) $\operatorname{cov}(\hat{\boldsymbol{\beta}}) = \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}$.

(3) $\hat{\boldsymbol{\beta}}$ is the best linear unbiased estimate of $\boldsymbol{\beta}$. This means that, for any vector, $\boldsymbol{c}$ $\operatorname{var}(\boldsymbol{c}^T\hat{\boldsymbol{\beta}}) \leq \operatorname{var}(\boldsymbol{c}^T\tilde{\boldsymbol{\beta}})$ holds where $\tilde{\boldsymbol{\beta}}$ is an unbiased estimator of $\boldsymbol{\beta}$, which has the form $\tilde{\boldsymbol{\beta}} = \boldsymbol{A}\boldsymbol{y} + \boldsymbol{d}$ for some matrix $\boldsymbol{A}$ and vector $\boldsymbol{d}$.

**Estimators in Linear Multiple Regression**

Least-squares estimate

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

Unbiased estimate of $\sigma^2$

$$\hat{\sigma}^2 = \frac{1}{n-p-1}\sum_{i=1}^{n}(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta})^2$$

### 1.4.4 Residuals and Hat Matrix

For single observations the discrepancy between the actual observation and the fitted value $\hat{y}_i = \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}$ is given by the simple residual

$$r_i = y_i - \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}.$$

It is a preliminary indicator for ill-fitting observations, that is, observations that have large residuals. Since the classical linear model assumes that the variance is the same for all observations, one might suspect that the residuals also have the same variance. However, because $\hat{\boldsymbol{\beta}}$ depends on all of the observations, they do not. Thus, for the diagnosis of an ill-fitting value, one has to take the variability of the estimate into account. For the derivation of the variance a helpful tool is the hat matrix. Consider the vector of residuals given by

$$\boldsymbol{r} = \boldsymbol{y} - \hat{\boldsymbol{y}} = \boldsymbol{y} - \boldsymbol{H}\boldsymbol{y} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{y},$$

where $\boldsymbol{H}$ is the projection matrix $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$. The matrix $\boldsymbol{H}$ is called the hat matrix because one has $\hat{\boldsymbol{y}} = \boldsymbol{H}\boldsymbol{y}$; thus $\boldsymbol{H}$ maps $\hat{\boldsymbol{y}}$ into $\boldsymbol{y}$. $\boldsymbol{H}$ is a projection matrix because it is symmetric and idempotent, that is, $\boldsymbol{H}^2 = \boldsymbol{H}$. It represents the projection of the observed values into the space spanned by $\boldsymbol{H}$. The decomposition

$$\boldsymbol{y} = \hat{\boldsymbol{y}} + \boldsymbol{y} - \hat{\boldsymbol{y}} = \boldsymbol{H}\boldsymbol{y} + (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{y}$$

is orthogonal because $\hat{\boldsymbol{y}}^T(\boldsymbol{y} - \hat{\boldsymbol{y}}) = \boldsymbol{y}^T\boldsymbol{H}(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{y} = \boldsymbol{y}^T(\boldsymbol{H} - \boldsymbol{H})\boldsymbol{y} = 0$. The covariance of $\boldsymbol{r}$ is easily derived by

$$\operatorname{cov}(\boldsymbol{r}) = (\boldsymbol{I} - \boldsymbol{H})\operatorname{cov}(\boldsymbol{y})(\boldsymbol{I} - \boldsymbol{H}) = \sigma^2(\boldsymbol{I} - \boldsymbol{H} - \boldsymbol{H} + \boldsymbol{H}^2) = \sigma^2(\boldsymbol{I} - \boldsymbol{H}).$$

Therefore one obtains with the diagonal elements from $\boldsymbol{H} = (h_{ij})$ the variance $\operatorname{var}(r_i) = \sigma^2(1 - h_{ii})$. Scaling to the same variance produces the form

$$\tilde{r}_i = \frac{r_i}{\sqrt{1 - h_{ii}}},$$

with $\text{var}(\tilde{r}_i) = \sigma^2$. If, in addition, one divides by the estimated variance $\hat{\sigma}^2 = (\boldsymbol{r}^T\boldsymbol{r})/(n-p-1)$, where $p+1$ is the length of $\boldsymbol{x}_i$, one obtains the *studentized residual*

$$r_i^* = \frac{\tilde{r}_i}{\hat{\sigma}} = \frac{y_i - \hat{\mu}_i}{\hat{\sigma}\sqrt{1-h_{ii}}},$$

which behaves much like a Student's $t$ random variable except for the fact that the numerator and denominator are not independent.

The hat matrix itself is a helpful tool in diagnosis. From $\hat{\boldsymbol{y}} = \boldsymbol{H}\boldsymbol{y}$ it is seen that the element $h_{ij}$ of the hat matrix $\boldsymbol{H} = (h_{ij})$ shows the amount of leverage or influence exerted on $\hat{y}_i$ by $y_j$. Since $\boldsymbol{H}$ depends only on $\boldsymbol{x}$, this influence is due to the "design" and not to the dependent variable. The most interesting influence is that of $y_i$ on the fitted value $\hat{y}_i$, which is reflected by the diagonal element $h_{ii}$. For the projection matrix $\boldsymbol{H}$ one has

$$\text{rank}(\boldsymbol{H}) = \sum_{i=1}^{n} h_{ii} = p+1$$

and $0 \le h_{ii} \le 1$. Therefore, $(p+1)/n$ is the average size of a diagonal element. As a rule of thumb, an $\boldsymbol{x}$-point for which $h_{ii} > 2(p+1)/n$ holds is considered a high-leverage point (e.g., Hoaglin and Welsch, 1978).

## Case Deletion as Diagnostic Tool

In case deletion let the deletion of observation $i$ be denoted by subscript $(i)$. Thus $\boldsymbol{X}_{(i)}$ denotes the matrix that is obtained from $\boldsymbol{X}$ by omitting the $i$th row; in $\boldsymbol{\mu}_{(i)}, \boldsymbol{y}_{(i)}$ the $i$th observation component is also omitted. Let $\hat{\boldsymbol{\beta}}_{(i)}$ denote the least-squared estimate resulting from the reduced dataset. The essential connection between the full dataset and the reduced set is given by

$$(\boldsymbol{X}_{(i)}^T\boldsymbol{X}_{(i)})^{-1} = (\boldsymbol{X}^T\boldsymbol{X})^{-1} + (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_i\boldsymbol{x}_i^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}/(1-h_{ii}), \tag{1.2}$$

where $h_{ii} = \boldsymbol{x}_i^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_i$ is the diagonal element of $\boldsymbol{H}$. One obtains after some computation

$$\hat{\boldsymbol{\beta}}_{(i)} = \hat{\boldsymbol{\beta}} - (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_i r_i/(1-h_{ii}).$$

Thus the change in $\boldsymbol{\beta}$ that results if the $i$th observation is omitted may be measured by

$$\Delta_i\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_i r_i/(1-h_{ii}).$$

Again, the diagonal element of the hat matrix plays an important role. Large values of $h_{ii}$ yield values $\hat{\boldsymbol{\beta}}_{(i)}$, which are distinctly different from $\hat{\boldsymbol{\beta}}$.

The simple *deletion residual* is given by

$$r_{(i)} = y_i - \hat{\mu}_{(i)},$$

where $\hat{\mu}_{(i)} = \boldsymbol{x}_i^T\boldsymbol{\beta}_{(i)}$. It measures the deviation of $y_i$ from the value predicted by the model fitted to the remaining points and therefore reflects the accuracy of the prediction. From $\hat{\mu}_{(i)} = \boldsymbol{x}_i^T\boldsymbol{\beta}_{(i)}$ one obtains $\text{var}(r_{(i)}) = \sigma^2 + \sigma^2\boldsymbol{x}_i^T(\boldsymbol{X}_{(i)}^T\boldsymbol{X}_{(i)})^{-1}\boldsymbol{x}_i = \sigma^2(1+h_{(i)})$, where $h_{(i)} = \boldsymbol{x}_i^T(\boldsymbol{X}_{(i)}^T\boldsymbol{X}_{(i)})^{-1}\boldsymbol{x}_i$. It follows easily from equation (1.2) that $h_{(i)}$ is given by $h_{(i)} = h_{ii}/(1-h_{ii})$. One obtains the standardized value

$$\tilde{r}_{(i)} = \frac{r_{(i)}}{\sqrt{1+h_{(i)}}},$$

which has variance $\sigma^2$. With $\hat{\sigma}^2_{(i)} = \boldsymbol{r}^T_{(i)}\boldsymbol{r}_{(i)}/(n-p-1)$ one obtains the *studentized* version

$$r^*_{(i)} = \frac{\tilde{r}_{(i)}}{\hat{\sigma}_{(i)}} = \frac{y_i - \hat{\mu}_{(i)}}{\hat{\sigma}_{(i)}\sqrt{1+h_{(i)}}} = \frac{y_i - \hat{\mu}_i}{\hat{\sigma}_{(i)}\sqrt{1-h_{ii}}}.$$

The last transformation follows from $\hat{\mu}_{(i)} = \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}_{(i)} = \boldsymbol{x}_i^T(\hat{\boldsymbol{\beta}} - (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_i r_i/(1-h_{ii})) = \hat{\mu}_i - \boldsymbol{x}_i^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_i r_i/(1-h_{ii}) = \hat{\mu}_i - h_{ii}r_i/(1-h_{ii}) = \hat{\mu}_i - r_i h_{(i)}$. Therefore one has $y_i - \hat{\mu}_{(i)} = (y_i - \hat{\mu}_i)(1+h_{(i)})$.

The studentized deletion residual is related to the studentized residual by $r^*_{(i)} = r_i\hat{\sigma}/\hat{\sigma}_{(i)}$. It represents a standardization of the scaled residual $(y_i - \hat{\mu}_i)/\sqrt{1-h_i}$, which has variance $\hat{\sigma}^2$ by an estimate of $\sigma^2$, which does not depend on the $i$th observation. Therefore, when normality holds, the standardized case deletion residual is distributed as Student's $t$ with $(n-p)$ degrees of freedom. Cook and Weisberg (1982) refer to $r^*_i$ as the studentized residuals with internal studentization, in contrast to external studentization for $r^*_{(i)}$. The $r^*_i$'s are also called cross-validatory or jackknife residuals. Rawlings et al. (1998) used the term studentized residuals for $r^*_i$. For more details on residuals see Cook and Weisberg (1982).

**Residuals**

Simple residual

$$r_i = y_i - \hat{\mu}_i = \boldsymbol{y}_i - \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}$$

Studentized residual

$$r^*_i = \frac{y_i - \hat{\mu}_i}{\hat{\sigma}\sqrt{1-h_{ii}}}$$

Case deletion residual

$$r_{(i)} = y_i - \hat{\mu}_{(i)} = y_i - \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}_{(i)}$$

Studentized case deletion residual

$$r^*_{(i)} = \frac{y_i - \hat{\mu}_i}{\hat{\sigma}_{(i)}\sqrt{1-h_{ii}}}$$

### 1.4.5 Decomposition of Variance and Coefficient of Determination

The sum of squared deviations from the mean may be partitioned in the following way:

$$\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (\hat{\mu}_i - \bar{y})^2 + \sum_{i=1}^n (\hat{\mu}_i - y_i)^2, \tag{1.3}$$

where $\bar{y} = \sum_i y_i/n$ is the mean over the responses. The partitioning has the form $\mathrm{SST} = \mathrm{SSR} + \mathrm{SSE}$, where

$\mathrm{SST} = \sum_{i=1}^n (y_i - \bar{y})^2$ is the *total sum of squares*, which represents the total variation in $y$ that is to be explained by $x$-variables;

SSR = $\sum_{i=1}^{n}(\hat{\mu}_i - \bar{y})^2$ ($R$ for regression) is the *regression sum of squares* built from the squared deviations of the fitted values around the mean;

SSE = $\sum_{i=1}^{n}(\hat{\mu}_i - y_i)^2$ ($E$ for error) is the sum of the squared residuals, also called the *error sum of squares*.

The partitioning (C.3) may also be seen from a geometric view. The fitted model based on the least-squares estimate $\hat{\boldsymbol{\beta}}$ is given by

$$\hat{\boldsymbol{\mu}} = \boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} = \boldsymbol{H}\boldsymbol{y},$$

which represents a projection of $\boldsymbol{y}$ into the space span($\boldsymbol{X}$), which is spanned by the columns of $\boldsymbol{X}$. Since span($\boldsymbol{X}$) contains the vector $\mathbf{1}$ and projections are linear operators, one obtains with $P_{\boldsymbol{X}}$ denoting the projection into span($\boldsymbol{X}$) and $\bar{\boldsymbol{y}}^T = (\bar{y}, \ldots, \bar{y})$ the orthogonal decomposition

$$\boldsymbol{y} - \bar{\boldsymbol{y}} = \hat{\boldsymbol{\mu}} - \bar{\boldsymbol{y}} + \boldsymbol{y} - \hat{\boldsymbol{\mu}},$$

where $\hat{\boldsymbol{\mu}} - \bar{\boldsymbol{y}} = \boldsymbol{H}\boldsymbol{y} - \bar{\boldsymbol{y}}$ is the projection of $\boldsymbol{y} - \bar{\boldsymbol{y}}$ into span($\boldsymbol{X}$) and $\boldsymbol{y} - \hat{\boldsymbol{\mu}} = \boldsymbol{y} - \boldsymbol{H}\boldsymbol{y}$ is from the orthogonal complement of span($\boldsymbol{X}$) such that $(\boldsymbol{y} - \hat{\boldsymbol{\mu}})^T(\hat{\boldsymbol{\mu}} - \bar{\boldsymbol{y}}) = 0$.

The *coefficient of determination* is defined by

$$R^2 = \frac{\text{SSR}}{\text{SST}} = \frac{\sum_i(\hat{\mu}_i - \bar{y})^2}{\sum_i(y_i - \bar{y})^2} = 1 - \frac{\sum_i(\hat{\mu}_i - y_i)^2}{\sum_i(y_i - \bar{y}_i)^2}.$$

Thus $R^2$ gives the proportion of variation explained by the regression model and therefore is a measure for the adequacy of the linear regression model.

From the definition it is seen that $R^2$ is not defined for the trivial case where $y_i = \bar{y},\ i = 1, \ldots, n$, which is excluded here. Extreme values of $R^2$ are

$R^2 = 0 \Leftrightarrow \hat{\mu}_i = \bar{y}$, that is, a horizontal line is fitted;

$R^2 = 1 \Leftrightarrow \hat{\mu}_i = y_i$, that is, all observations are on a line with slope unequal to 0.

Although $R^2$ is often considered as a measure of goodness-of-fit, it hardly reflects goodness-of-fit in the sense that a high value of $R^2$ tells us that the underlying model is a linear regression model. Though built from residuals, $R^2$ compares the residuals of the model that specifies a linear effect of variables $\boldsymbol{x}$ and the residuals of the simple intercept model (the null model). Thus it measures the additional explanatory values of variable vector $\boldsymbol{x}$ within the linear model. It cannot be used to decide whether a model shows appropriate fit. Rather, $R^2$ and its generalizations measure the strength of association between covariates and the response variable. If $R^2$ is large, the model should be useful since some aspect of the association between the response and covariates is captured by the linear model. On the other hand, if $R^2$ is close to zero, that does not mean that the model has a bad fit. On the contrary, if a horizontal line is fitted ($R^2 = 0$), the data may be very close to the fitted data, since any positive value $\sum_i(\hat{\mu}_i - y_i)^2$ is possible. $R^2 = 0$ just means that there is no linear association between the response and the linear predictor beyond the horizontal line. $R^2$ tells how much of the variation is explained by the included variables within the linear approach. It is a *relative measure* that reflects the improvement by the inclusion of predictors as compared to the simple model, where only the constant term is included.

Now we make some additional remarks to avoid misrepresentation. That $R^2$ is not a tool to decide if the linear model is true or not may be easily seen from considering an underlying

linear model. Let a finite number of observations be drawn from the range of $\boldsymbol{x}$-values. Now, in addition to the sample of size $n$, let $n_0$ observations be drawn at a fixed design point $\boldsymbol{x}_0$. Then, for $n_0 \to \infty$, it follows that $\bar{y} \to \mu_0 = \mathrm{E}(y|x_0)$ and $\hat{\mu}_i \to \mu_0$ for $i \gg n$ such that $R^2 \to 0$. This means that although the linear model is true, $R^2$ approaches zero and therefore cannot be a measure for the truth of the model. On the other hand, if a non-linear model is the underlying model and observations are only drawn at two distinct design points, $R^2$ will approach 1 since two points may always be fitted by a line. The use of $R^2$ *is restricted to linear models*. There are examples where $R^2$ can be larger than 1 if a non-linear function is fitted by least squares (see Exercise 1.4).

### 1.4.6 Testing in Multiple Linear Regression

The most important tests are for the hypotheses

$$H_0 : \beta_1 = \cdots = \beta_p = 0 \qquad H_1 : \beta_i \neq 0 \text{ for at least one variable} \tag{1.4}$$

and

$$H_0 : \boldsymbol{\beta}_i = 0 \qquad H_1 : \boldsymbol{\beta}_i \neq 0. \tag{1.5}$$

The first null hypothesis asks if there is any explanatory value of the covariates, whereas the latter concerns the question if one specific variable may be omitted – given that all other variables are included. The test statistics may be easily derived as special cases of linear hypotheses (see next section). For normally distributed responses one obtains for $H_0 : \beta_j = 0$

$$t = \frac{\hat{\beta}_j}{\hat{cov}(\hat{\beta}_j)} \sim t(n-p-1),$$

where $\hat{cov}(\hat{\beta}_j) = \hat{\sigma}\sqrt{a_{jj}}$ with $a_{jj}$ denoting the $j$th diagonal element of $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ and $H_0$ is rejected if $|t| > t_{1-\alpha/2}(n-p-1)$. For $H_0 : \beta_1 = \ldots = \beta_p = 0$ and normally distributed responses one obtains

$$F = \frac{n-p-1}{p} \frac{R^2}{1-R^2} = \frac{(\mathrm{SST} - \mathrm{SSE})/p}{\mathrm{SSE}/(n-p-1)} \overset{H_0}{\sim} F(p, n-p-1)$$

and $H_0$ is rejected if $F > F_{1-\alpha}(p, n-p-1)$. The $F$-test for the global hypothesis $H_0 : \beta_1 = \cdots = \beta_p = 0$ is often given within an analysis-of-variance (ANOVA) framework. Consider again the partitioning of the total sum of squares:

$$\mathrm{SST} = \mathrm{SSR} + \mathrm{SSE},$$

$$(\boldsymbol{y} - \bar{\boldsymbol{y}})^T(\boldsymbol{y} - \bar{\boldsymbol{y}}) = (\hat{\boldsymbol{\mu}} - \bar{\boldsymbol{y}})^T(\hat{\boldsymbol{\mu}} - \bar{\boldsymbol{y}}) + (\boldsymbol{y} - \hat{\boldsymbol{\mu}})^T(\boldsymbol{y} - \hat{\boldsymbol{\mu}}).$$

If the regression model holds, the error sum of squares has a scaled $\chi^2$-distribution, $\mathrm{SSE} \sim \sigma^2\chi^2(n-p-1)$. The degrees of freedom follow from considering the residuals $\boldsymbol{r} = \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{y} - \boldsymbol{H}\boldsymbol{y} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{y}$, which represent an orthogonal projection by use of the projection matrix $\boldsymbol{I} - \boldsymbol{H}$. Since SSE is equivalent to the squared residuals, $\mathrm{SSE} = \boldsymbol{r}^T\boldsymbol{r}$ , and $\boldsymbol{I} - \boldsymbol{H}$ has rank $n - p - 1$, one obtains $\sigma^2\chi^2(n-p-1)$. If the regression model holds and in addition $\beta_1 = \cdots = \beta_p = 0$, then one obtains for SSE and SSR the $\chi^2$-distributions

$$\mathrm{SST} \sim \sigma^2\chi^2(n-1),\ \mathrm{SSR} \sim \sigma^2\chi^2(p).$$

In addition, in this case SSR and SSE are independent. The corresponding means squares are given by

$$\mathrm{MSE} = \mathrm{SSE}/(n-p-1), \mathrm{MSR} = \mathrm{SSR}/p.$$

It should be noted that while SSE and SSR sum up to SST, the sum of MSE and MSR does not give the average over all terms.

TABLE 1.4: ANOVA table for multiple linear regression.

| Source of variation | $SS$ | $df$ | $MS$ |
|---|---|---|---|
| Regression | $\text{SSR} = \sum_{i=1}^{n}(\hat{y}_i - \bar{y}_i)^2$ | $p$ | $MSR = \frac{\text{SSR}}{p}$ |
| Error | $\text{SSE} = \sum_{i=1}^{n}(\hat{\mu}_i - y_i)^2$ | $n-(p+1)$ | $MSE = \frac{\text{SSE}}{n-p-1}$ |

## Submodels and the Testing of Linear Hypotheses

A general framework for testing all kinds of interesting hypotheses is the testing of linear hypotheses given by

$$H_0 : \boldsymbol{C\beta} = \boldsymbol{\xi} \qquad H_1 : \boldsymbol{C\beta} \neq \boldsymbol{\xi}.$$

The simple null hypothesis $H_0 : \beta_1 = \beta_2 = \cdots = \beta_p, = 0$ turns into

$$H_0 : \begin{pmatrix} 0 & 1 & & & \\ \vdots & & 1 & & \\ \vdots & & & \ddots & \\ 0 & & & & 1 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}.$$

The hypothesis $H_0 : \ \beta_i = 0$ is given by $H_0 : (0 \ldots 1 \ldots 0)\boldsymbol{\beta} = 0$. Comparisons of covariates of the form $H_0 : \beta_i = \beta_j$ are given by

$$H_0 : \ (0, \ldots, 1, \ldots, -1, \ldots 0)\boldsymbol{\beta} = 0,$$

where 1 corresponds to $\beta_i$ and $-1$ corresponds to $\beta_j$. Hypotheses like $H_0 : \beta_1 = \cdots = \beta_p = 0$ or $H_0 : \beta_i = 0$ are linear hypotheses, and the corresponding models may be seen as submodels of the multiple regression model. They are submodels because the parameter space is more restricted than in the original multiple regression model.

Let the more general $\tilde{M}$ be a submodel of $M$ ($\tilde{M} \subset M$), where $M$ is the unrestricted multiple regression model and $\tilde{M}$ is restricted to a linear subspace of dimension $(p+1) - s$, that is, $\text{rank}(\boldsymbol{C}) = s$. For example, if the restricted model contains only the intercept, one has $\text{rank}(\boldsymbol{C}) = 1$ and the restricted model specifies a subspace of dimension one. Let $\hat{\boldsymbol{\beta}}$ denote the usual least-squares estimate for the multiple regression model and $\tilde{\boldsymbol{\beta}}$ be the restricted estimate that minimizes

$$(\boldsymbol{y} - \boldsymbol{X\beta})^T(\boldsymbol{y} - \boldsymbol{X\beta})$$

under the restriction $\boldsymbol{C\beta} = \boldsymbol{\xi}$. Using Lagrange multipliers yields

$$\tilde{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}} - (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{C}^T[\boldsymbol{C}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{C}^T]^{-1}[\boldsymbol{C}\hat{\boldsymbol{\beta}} - \boldsymbol{\xi}]. \tag{1.6}$$

One obtains two discrepancies, namely, the discrepancy between $M$ and the data and the discrepancy between $\tilde{M}$ and the data. As a discrepancy measure one may use a residual or error sums of squares:

$$\text{SSE}(M) = \sum_{i=1}^{n}(\boldsymbol{y}_i - \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}})^2 = (\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^T(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}),$$

$$\text{SSE}(\tilde{M}) = \sum_{i=1}^{n}(\boldsymbol{y}_i - \boldsymbol{x}_i^T\tilde{\boldsymbol{\beta}})^2 = (\boldsymbol{y} - \boldsymbol{X}\tilde{\boldsymbol{\beta}})^T(\boldsymbol{y} - \boldsymbol{X}\tilde{\boldsymbol{\beta}}).$$

Since $\tilde{M}$ is a more restricted model, $\mathrm{SSE}(\tilde{M})$ tends to be greater than $\mathrm{SSE}(M)$. One may decompose the discrepancy $\mathrm{SSE}(\tilde{M})$ by considering

$$\mathrm{SSE}(\tilde{M}) = \mathrm{SSE}(M) + \mathrm{SSE}(\tilde{M}|M), \tag{1.7}$$

where $\mathrm{SSE}(\tilde{M}|M) = \mathrm{SSE}(\tilde{M}) - \mathrm{SSE}(M)$ is the increase of residuals that results from using the more restrictive model $\tilde{M}$ instead of $M$. It may also be seen as the amount of variation explained by $M$ but not by $\tilde{M}$. The notation refers to the interpretation as a conditional discrepancy; $\mathrm{SSE}(\tilde{M}|M)$ is the discrepancy of $\tilde{M}$ within model $M$, that is, the additional discrepancy between data and model. This results from fitting $\tilde{M}$ instead of the less restrictive model $M$. The decomposition (1.7) may be used for testing the fit of $\tilde{M}$ given that $M$ is an accepted model. This corresponds to testing $H_0 : \boldsymbol{C\beta} = \boldsymbol{\xi}$ (corresponding to $\tilde{M}$) within the multiple regression model (corresponding to $M$).

An important property of the decomposition (1.7) is that it is based on orthogonal components. Behind (1.7) is the trivial decomposition

$$\boldsymbol{y} - \boldsymbol{X}\tilde{\boldsymbol{\beta}} = (\boldsymbol{X}\hat{\boldsymbol{\beta}} - \boldsymbol{X}\tilde{\boldsymbol{\beta}}) + (\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}), \tag{1.8}$$

where $\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}$ is orthogonal to $\boldsymbol{X}\hat{\boldsymbol{\beta}} - \boldsymbol{X}\tilde{\boldsymbol{\beta}}$, i.e. $(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^T(\boldsymbol{X}\hat{\boldsymbol{\beta}} - \boldsymbol{X}\tilde{\boldsymbol{\beta}}) = 0$. Decomposition (1.7) follows from (1.8) by considering $\mathrm{SSE}(\tilde{M}) = (\boldsymbol{y} - \boldsymbol{X}\tilde{\boldsymbol{\beta}})^T(\boldsymbol{y} - \boldsymbol{X}\tilde{\boldsymbol{\beta}})$, $\mathrm{SSE}(M) = (\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^T(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})$. From (1.8) and (1.6) the explicit form of $\mathrm{SSE}(\tilde{M}|M)$ follows as

$$\mathrm{SSE}(\tilde{M}|M) = (\boldsymbol{C}\hat{\boldsymbol{\beta}} - \boldsymbol{\xi})^T[\boldsymbol{C}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{C}^T]^{-1}(\boldsymbol{C}\hat{\boldsymbol{\beta}} - \boldsymbol{\xi}).$$

If $M$ holds, $\mathrm{SSE}(M)$ is $\chi^2$-distributed with $\mathrm{SSE}(M)/\sigma^2 \sim \chi^2(n-p-1)$; if $\tilde{M}$ holds ($H_0$ is true), $\mathrm{SSE}(\tilde{M}|M)/\sigma^2 \sim \chi^2(s)$ and $\mathrm{SSE}(M)$ and $\mathrm{SSE}(\tilde{M}|M)$ are independent. One obtains

| | | | | |
|---|---|---|---|---|
| $\mathrm{SSE}(\tilde{M})$ | $=$ | $\mathrm{SSE}(\tilde{M}\|M)$ | $+$ | $\mathrm{SSE}(M)$ |
| $\sigma^2\chi^2(n-p-1+s)$ | | $\sigma^2\chi^2(s)$ | | $\sigma^2\chi(n-p-1)$ |
| if $\tilde{M}$ holds | | if $\tilde{M}$ holds | | if $M$ holds |

Thus, if $\tilde{M}$ holds,

$$F = \frac{\left(\mathrm{SSE}(\tilde{M}) - \mathrm{SSE}(M)\right)/s}{\mathrm{SSE}(M)/(n-p-1)} \sim F(s,\, n-p-1),$$

which may be used as test statistics for $H_0 : \boldsymbol{C\beta} = \boldsymbol{\xi}$. $H_0$ is rejected if $F$ is larger than the $(1-\alpha)$-quantile $F_{1-\alpha}(s,\, n-p-1)$.

## 1.5 Exercises

1.1 Consider a linear model that specifies the rent to pay as a function of the size of the flat and the city (with data for 10 cities available). Let the model be given as

$$\mathrm{E}(y|\text{size}, C = i) = \beta_0 + \text{size} * \beta_s + \beta_{C(i)}, \quad i = 1, \ldots, 10,$$

where $\beta_{C(i)}$ represents the effect of the city. Since the parameters $\beta_{C(1)}, \ldots, \beta_{C(10)}$ are not identifiable, one has to specify an additional constraint.

(a) Give the model with dummy variables by using the symmetric side constraint $\sum_i \beta_{C(i)} = 0$.

(b) Give the model with dummy variables by specifying a reference category.

(d) Specify $\boldsymbol{C}$ and $\boldsymbol{\xi}$ of the linear hypothesis $H_0 : \boldsymbol{C\beta} = \boldsymbol{\xi}$ if you want to test if rent does not vary over cities.

(e) What is the meaning if the hypothesis $H_0 : \beta_{C(j)} = 0$ for fixed $j$ holds?

(f) Find the transformation that transforms parameters with a reference category into parameters with a symmetric side constraint, and vice versa for a general number of cities $k$.

1.2 The R Package *catdata* provides the dataset *rent*.

(a) Use descriptive tools to learn about the data.

(b) Fit a linear regression model with response *rent* (net rent in Euro) and explanatory variables *size* (size in square meters) and *rooms* (number of rooms). Discuss the results.

(c) Fit a linear regression model with response *rent* and the single explanatory variable *rooms*. Compare with the results from (b) and explain why the coefficients differ even in sign.

1.3 The dataset *rent* from R Package *catdata* contains various explanatory variables.

(a) Use the available explanatory variables when fitting a linear regression model with the response *rent*. Include polynomial terms and dummy variables if necessary. Evaluate if explanatory variables can be excluded.

(b) Fit a linear model with the response *rentm* (rent per square meter) by using the available explanatory variables. Discuss the effects and compare to the results from (a).

1.4 Kockelkorn (2000) considers the model $y_i = \mu(x_i) + \varepsilon_i$ with $\mu(x) = x^\beta$ if $x \geq 0$ and $\mu(x) = -(-x)^\beta$ if $x < 0$. For some $z > 0$ let observations $(y_i, x_i)$ be given by $\{(0,1), (0,-1), (-z^3, -z), (z^3, z)\}$.

(a) Compute the value $\beta$ that minimizes the least-squares criterion.

(b) Compute $R^2$ as a function of $z$ and investigate what values $R^2$ takes.

# Chapter 2

# Binary Regression: The Logit Model

Categorical regression has the same objectives as metric regression. It aims at an economic representation of the link between covariables considered as the independent variables and the response as the dependent variable. Moreover, one wants to evaluate the influence of the independent variables regarding their strength and the way they exert their influence. Predicting new observations can be based on adequate modeling of the response pattern.

Categorical regression modeling differs from classical normal regression in several ways. The most crucial difference is that the dependent variable $y$ follows a quite different distribution. A categorical response variable can take only a limited number of values, in contrast to normally distributed variables, in which any value might be observed. In the simplest case of binary regression the response takes only two values, usually coded as $y = 0$ and $y = 1$. One consequence is that the scatterplots look different. Figure 2.1 shows data from the household panel described in Example 1.2. The outcomes "car in household" $(y = 1)$ and "no car in household" $(y = 0)$ are plotted against net income (in Euros). It is seen that for low income the responses $y = 0$ occur more often, whereas for higher income $y = 1$ is observed more often. However, the structural connection between the response and the covariate is hardly seen from this representation. Therefore, in Figure 2.1 the relative frequencies for owning a car are shown for households within intervals of length $50$. The picture shows that a linear connection is certainly not the best choice. This leads to the second difference, which concerns the link between the covariates and the mean of the response. Although the covariates might enter the model as a linear term in the same way as in classical regression models, the link between response and linearly structured explanatory variables usually has to be modified to avoid improper models. Thus at least two structuring elements have to be modified: the distribution of the response and the link between the response and the explanatory variables.

In the following, first distributions of $y$ are considered and then methods for structuring the link between the response and the independent variables are outlined. The chapter introduces the concept of models for binary responses. Estimation and inference are considered in Chapter 4, and extensions are given in Chapter 5.

## 2.1 Distribution Models for Binary Responses and Basic Concepts

### 2.1.1 Single Binary Variables

Let the binary response $y$ be coded by $y = 1$ and $y = 0$. In Example 1.2, the two outcomes refer to “car in household” and “no car in household”. Often $y = 1$ is considered as success

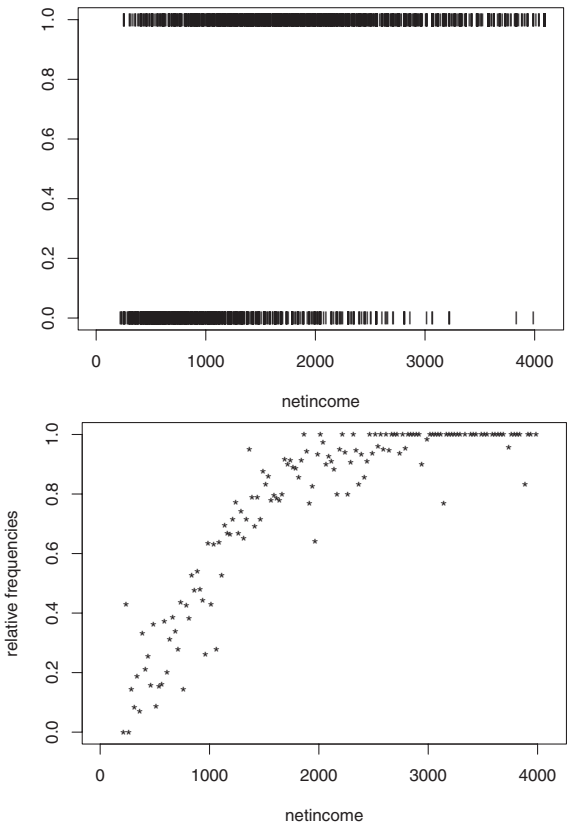


FIGURE 2.1: Car data: Upper panel shows the raw data with $y = 1$ for "car in household" and $y = 0$ for "no car in household" plotted against net income in Euros. Lower panel shows the relative frequencies within intervals of length 50, plotted against net income.

and $y = 0$ as failure, a convention that will be used in the following. The distribution of the simple binary random variable $y \in \{0, 1\}$ is completely characterized by the probability

$$\pi = P(y = 1).$$

The probability for $y = 0$ is then given by $P(y = 0) = 1 - \pi$. The mean of $y$ is simply computed by

$$\mathrm{E}(y) = 1 \times \pi + 0 \times (1 - \pi) = \pi.$$

Therefore, the response probability $\pi$ represents the mean of the binary distribution. The variance is computed equally simple by

$$\mathrm{var}(y) = \mathrm{E}(y - \mathrm{E}(y))^2 = (1 - \pi)^2\pi + (0 - \pi)^2(1 - \pi) = \pi(1 - \pi).$$

It is seen that the variance is completely determined by $\pi$ and depends on $\pi$ with minimal value zero at $\pi = 0$ and $\pi = 1$ and maximal value at $\pi = 1/2$. This is in accordance with intuition; if $\pi = 0$, only $y = 0$ can be observed; consequently, the variance is zero since there is no variability in the responses. The same holds for $\pi = 1$, where only $y = 1$ can be observed.

### 2.1.2 The Binomial Distribution

In many applications a binary variable is observed repeatedly and the focus is on the number of successes (occurrence of $y = 1$). The classical example is the flipping of a coin $n$ times and then counting the number of trials where heads came up. More interestingly, the trials may refer to a standardized treatment of $n$ persons and the outcome is the number of persons for whom treatment was successful. The same data structure is found if in Example 1.2 income is measured in categories, where categories refer to intervals of length 50. Considering the households within one interval as having the same response distribution, one has repeated trials with $n$ denoting the number of households within a specific interval.

The basic assumptions underlying the binomial distribution are that the random variables $y_1, \ldots, y_n$ with fixed $n$ are binary, $y_i \in \{0, 1\}$ with the same response probability $\pi = P(y_i = 1), i = 1, \ldots, n$, and are independent. Then the number of successes $y = y_1 + \cdots + y_n$ in $n$ trials is called a *binomial random variable*, $y \sim \mathrm{B}(n, \pi)$, and has the distribution function

$$P(y = r) = \binom{n}{r} \pi^r (1 - \pi)^{n-r}, \quad r = 1, \ldots, n.$$

The mean and variances of a binomial variable are easily calculated. One obtains

$$\mathrm{E}(y) = n\pi, \ \mathrm{var}(y) = n\pi(1 - \pi).$$

The random variable $y$ counts the number of successes in $n$ trials. Often it is useful to look at the relative frequencies or proportions $y/n$ rather than at the number of successes $y$. The form of the distribution remains the same; only the support changes since for $n$ trials $y$ takes values from $\{0, 1, \ldots, n\}$, whereas $y/n$ takes values from $\{0, 1/n, \ldots, 1\}$. The probability function of $y/n$ is given by

$$P(y/n = z) = \binom{n}{nz} \pi^{nz} (1 - \pi)^{n-nz},$$

where $z \in \{0, 1/n, \ldots (n - 1)/n, 1\}$. The distribution of $y/n$ is called a *scaled binomial distribution*, frequently abbreviated by $y/n \sim \mathrm{B}(n, \pi)/n$. One obtains

$$\mathrm{E}(y/n) = \pi, \ \mathrm{var}(y/n) = \pi(1 - \pi)/n.$$

It is seen that $y/n$ has mean $\pi$. Therefore, the relative frequency is a natural and unbiased estimate for the underlying $\pi$. The variance of the estimate $y/n$ depends on $\pi$ and $n$ with large values for $\pi$ close to 0.5 and small values if $\pi$ approaches 0 or 1. Generally the variance decreases with increasing $n$. It is noteworthy that in Figure 2.1 the estimates vary in their degree of trustworthiness since the sample sizes and underlying probabilities vary across intervals.

### Odds, Logits, and Odds Ratios

A binary response variable is distinctly different from a continuous response variable. While the essential characteristics of a continuous response variable are the mean $\mu = \mathrm{E}(y)$, the variance $\sigma^2 = \mathrm{var}(y)$, skewness, and other measures, which frequently may vary independently, in a binary response variable all these characteristics are determined by only one value, which often is chosen as the probability of $y = 1$. Instead of using $\pi$ as an indicator of the random behavior of the response, it is often useful to consider some transformation of $\pi$. Of particular importance are *odds* and *log-odds* (also called *logits*), which are functions of $\pi$.

| Odds | Log Odds or Logits |
|---|---|
| $\gamma(\pi) = \frac{\pi}{1-\pi}$ | $\text{logit}\,(\pi) = \log(\gamma(\pi)) = \log(\frac{\pi}{1-\pi})$ |

The odds $\gamma(\pi) = \pi/(1-\pi)$ are a directed measure that compares the probability of the occurrence of $y = 1$ and the probability of the occurrence of $y = 0$. If $y = 1$ is considered a "success" and $y = 0$ a "failure", a value $\gamma = 1/4$ means that failure is four times as likely as a success. While $\gamma$ compares $y = 1$ to $y = 0$, the inverse compares $y = 0$ to $y = 1$, yielding

$$\gamma = \frac{P(y=1)}{P(y=0)}, \qquad \frac{1}{\gamma} = \frac{P(y=0)}{P(y=1)}.$$

Therefore, if odds are considered as functions of $\pi$, one obtains

$$\gamma(1-\pi) = \frac{1-\pi}{\pi} = \frac{1}{\gamma(\pi)}$$

because $\gamma(1-\pi)$ corresponds to comparing $y = 0$ to $y = 1$. Odds fulfill

$$\gamma(\pi)\gamma(1-\pi) = 1.$$

For the log-odds the relation is additive:

$$\text{logit}(1-\pi) = \log\left(\frac{1-\pi}{\pi}\right) = -\,\text{logit}(\pi),$$

and therefore

$$\text{logit}(\pi) + \text{logit}(1-\pi) = 0.$$

## Comparing Two Groups

The concept of odds and log-odds may be illustrated by the simple case of comparing two groups on a binary response. The groups can be two treatment groups – a treatment group and a control group – in a clinical trial or correspond to two populations, for example, men and women. The data are usually collected in a (2×2)-contingency table with the underlying probabilities given in the following table.

| | y | | |
|---|---|---|---|
| | 1 | 0 | |
| 1 | $\pi_1$ | $1-\pi_1$ | 1 |
| 2 | $\pi_2$ | $1-\pi_2$ | 1 |

In the table $\pi_t = P(y = 1|T = t)$, $t = 1, 2$, denotes the probability of response $y = 1$, corresponding to success. The probability for failure is then determined by $P(y = 0|T = t) = 1 - P(y = 1|T = t)$. A comparison of the two groups may be based on various measures. One may use the *difference of success probabilities*:

$$d_{12} = \pi_1 - \pi_2,$$

which implies the difference of failures, since $P(y = 0|T = 1) - P(y = 0|T = 2) = \pi_2 - \pi_1$. Groups are equivalent with respect to the response if the difference is zero. An alternative measure is the proportion, frequently called the *relative risk*:

$$r_{12} = \pi_1/\pi_2,$$

which has an easy interpretation. A relative risk of 2 means that the probability of success in group 1 is twice the probability of success in group 2. There is no difference between groups when the relative risk is 1.

Another quite attractive measure compares the odds rather than the probabilities. The *odds ratio* between groups 1 and 2 is

$$\gamma_{12} = \frac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)} = \frac{\gamma(\pi_1)}{\gamma(\pi_2)}.$$

The ratio of the odds $\gamma(\pi_1)$ and $\gamma(\pi_2)$ compares odds instead of ratios of probabilities of success. The odds ratio $\gamma_{21}$ between groups 2 and 1, $\gamma_{21} = \gamma(\pi_2)/\gamma(\pi_1)$, is simply the inverse, $\gamma_{21} = 1/\gamma_{12}$. Groups are equivalent as far as the response is concerned if $\gamma_{21} = \gamma_{12} = 1$. The odds ratio also measures the association between rows and columns, which is the association between the grouping variable and the response.

Rather than measuring association by odds ratios, one can use the log-transformed odds ratios

$$\log(\gamma_{12}) = \log\left(\frac{\gamma(\pi_1)}{\gamma(\pi_2)}\right).$$

While odds ratios can equal any non-negative number, log-odds ratios have the advantage that they are not restricted at all. A value of 0 means that the groups are equivalent.

Regression modeling in the case of two groups means modeling the response $y$ as a function of a binary predictor. The models for $\pi_t$ have to distinguish between $T = 1$ and $T = 2$. A model that uses the probabilities itself has the form

$$\pi_1 = \beta_0 + \beta, \qquad \pi_2 = \beta_0,$$

and the effect $\beta$ is equivalent to the difference in probabilities. This means in particular that the effect is restricted to the interval $[-1, 1]$. A more attractive model is the linear model for the log-odds which has the form

$$\log\left(\pi_1/(1-\pi_1)\right) = \beta_0 + \beta, \qquad \log(\pi_2/(1-\pi_2)) = \beta_0, \tag{2.1}$$

and one easily derives that $\beta$ is equal to the log-odds ratio and $\exp(\beta)$ is equal to the odds ratio:

$$\beta = \log(\gamma_{12}) = \log\left(\frac{\gamma(\pi_1)}{\gamma(\pi_2)}\right), \quad \exp(\beta) = \gamma_{21} = \frac{\gamma(\pi_2)}{\gamma(\pi_1)}.$$

Therefore, parameters correspond to common measures of association. The value $\beta = 0$, which means that no effect is present, corresponds to the case of no association, where the odds ratio is 1 and the log odds ratio is 0. Since model (2.1) parameterizes the logits, it is called the *logit model*, a model that will be considered extensively in the following. Model (2.1) is just the simplest version of a logit model where only one binary predictor is included.

## 2.2 Linking Response and Explanatory Variables

### 2.2.1 Deficiencies of Linear Models

Let $(y_i, \boldsymbol{x}_i), i = 1, \ldots, n$, be observations with binary response $y_i \in \{0, 1\}$ on a covariate vector $\boldsymbol{x}_i$. One could try to link $y_i$ and $\boldsymbol{x}_i$ by using the classical regression model

$$y_i = \beta_0 + \boldsymbol{x}_i^T \boldsymbol{\beta} + \varepsilon_i, \tag{2.2}$$

where $y_i$ is the response given $\boldsymbol{x}_i$ and $\varepsilon_i$ represents the noise with $\mathrm{E}(\varepsilon_i) = 0$. Several problems arise from this approach. First, the structural component $\mathrm{E}(y_i) = \beta_0 + \boldsymbol{x}_i^T\boldsymbol{\beta}$, which specifies the assumed dependence of the mean of $y$ on the covariates $\boldsymbol{x}_i$, is given by

$$\pi_i = \beta_0 + \boldsymbol{x}_i^T\boldsymbol{\beta},$$

where $\pi_i = P(y_i = 1|\boldsymbol{x}_i)$. Thus the range where a model of this type may hold is restricted severely. Since $\pi_i \in [0, 1]$ even for the simplest model with univariate predictor $\pi_i = \beta_0 + x_i\beta$, it is easy to find regressor values $x_i$ such that $\pi_i > 1$ or $\pi_i < 0$ if $\beta \neq 0$. The second problem concerns the random component of the model. From $\pi_i = \beta_0 + \boldsymbol{x}_i^T\boldsymbol{\beta}$ it follows that $\varepsilon_i$ takes only two values, $\varepsilon_i \in \{1 - \pi_i, -\pi_i\}$. Moreover, the variance of $\varepsilon_i$ is given by $\mathrm{var}(\varepsilon_i) = \mathrm{var}(y_i) = \pi_i(1 - \pi_i)$, which is the usual variance of a binary variable. Therefore the model is heteroscedastic, with the variance being directly connected to the mean. In summary, the usual linear regression model that assumes homogenous variances, continuous noise, and linear influence for all values of covariates is unsatisfactory in several ways. Nevertheless, as is shown in the next section, one may obtain a binary regression model by using the continuous regression model as a background model.

### 2.2.2 Modeling Binary Responses

In the following we show two ways to obtain more appropriate models for binary responses. The first gives some motivation for a model by considering underlying continuous responses that yield a rather general family of models. The second is based on conditional normal distribution models that yield a more specific model.

#### Binary Responses as Dichotomized Latent Variables

Binary regression models can be motivated by assuming that a linear regression model holds for a continuous variable that underlies the binary response. In many applications one can imagine that an underlying continuous variable steers the decision process that results in a categorical outcome. For example, the decision to buy a car certainly depends on the wealth of the household. If the wealth exceeds a certain threshold, a car is bought $(y_i = 1)$; if not, one observes $y_i = 0$. In a bioassay, where $y_i = 1$ stands for death of an animal depending on the dosage of some poison, the underlying variable that determines the response may represent the damage on vital bodily functions.

To formalize the concept, let us assume that the model $\tilde{y}_i = \gamma_0 + \boldsymbol{x}_i^T\boldsymbol{\beta} + \varepsilon_i$ holds for a latent response $\tilde{y}_i$. Moreover, let $-\varepsilon_i$ have the distribution function $F$, which does not depend on $\boldsymbol{x}_i$. The essential concept is to consider $y_i$ as a dichotomized version of the latent variable $\tilde{y}_i$ with the link between the observable variable $y_i$ and the latent variable $\tilde{y}_i$ given by

$$y_i = 1 \quad \text{if} \quad \tilde{y}_i \geq \theta, \tag{2.3}$$

where $\theta$ is some unknown threshold. One obtains

$$\begin{aligned}\pi_i = \pi(\boldsymbol{x}_i) &= P(y_i = 1|\boldsymbol{x}_i) = P(\tilde{y}_i \geq \theta) = P(\gamma_0 + \boldsymbol{x}_i^T\boldsymbol{\beta} + \varepsilon_i \geq \theta) \\ &= P(-\varepsilon_i \leq \gamma_0 - \theta + \boldsymbol{x}_i^T\boldsymbol{\beta}) = F(\beta_0 + \boldsymbol{x}_i^T\boldsymbol{\beta}),\end{aligned}$$

where $\beta_0 = \gamma_0 - \theta$. The resulting model has the simple form

$$\pi(\boldsymbol{x}_i) = F(\beta_0 + \boldsymbol{x}_i^T\boldsymbol{\beta}). \tag{2.4}$$

In this model the mean of the response is still determined by a term that is linear in $\boldsymbol{x}_i$ but the connection between $\pi_i$ and the linear term $\eta_i = \beta_0 + \boldsymbol{x}_i^T\boldsymbol{\beta}$ involves some transformation that is determined by the distribution function $F$.

The basic assumption behind the derivation of model (2.4) is that the observable variable $y_i$ is a coarser, even binary, version of the latent variable $\tilde{y}_i$. It is important that the derivation from latent variables be seen as a mere motivation for the binary response models. The resulting model and parameters may be interpreted without reference to any latent variable. This is also seen from the handling of the threshold $\theta$. Since it is never observed and is not identifiable, it vanishes in the parameter $\beta_0 = \gamma_0 - \theta$. A point in favor of the derivation from the latent model is that $\pi_i = F(\beta_0 + \boldsymbol{x}_i^T\boldsymbol{\beta})$ is always from the admissible range $\pi_i \in [0, 1]$ because $F$ is a distribution function for which $F(\eta) \in [0, 1]$ always holds. A simple example is the *probit model*,

$$\pi(\boldsymbol{x}_i) = \Phi(\beta_0 + \boldsymbol{x}_i^T\boldsymbol{\beta}),$$

where $\Phi$ is the distribution function of the standardized normal distribution $\mathrm{N}(0, 1)$. The more widely used model is the *logit model*,

$$\pi(\boldsymbol{x}_i) = F(\beta_0 + \boldsymbol{x}_i^T\boldsymbol{\beta}),$$

where $F$ is the logistic distribution function $F(\eta) = \exp(\eta)/(1 + \exp(\eta))$. An alternative derivation of the logit model is given in the following.

## Modeling the Common Distribution of a Binary and a Continuous Distribution

The motivation by latent variables yields the very general class of models (2.4), where $F$ may be any continuous distribution function. Models of this type will be considered in Chapter 5. Here we want to motivate the choice of a special distribution by considering a model for the total distribution of $y, \boldsymbol{x}$.

Let us consider the bivariate random variable $(y, \boldsymbol{x})$, where $y \in \{0, 1\}$ is a binary variable and $\boldsymbol{x}$ is a vector of continuous random variables. The most common continuous distribution is the normal distribution. But, rather than assuming $\boldsymbol{x}$ to be normally distributed in the total population, let us assume that $\boldsymbol{x}$ is conditionally normally distributed within the groups that are determined by $y = 0$ and $y = 1$. One assumes

$$\boldsymbol{x}|y = r \sim \mathrm{N}(\boldsymbol{\mu}_r, \boldsymbol{\Sigma}),$$

where the covariance matrix $\boldsymbol{\Sigma}$ is the same for all conditions $y = r$. Simple derivations based on Bayes' theorem shows that $y|\boldsymbol{x}$ is determined by

$$\pi(\boldsymbol{x}) = P(y = 1|\boldsymbol{x}) = \frac{\exp(\beta_0 + \boldsymbol{x}^T\boldsymbol{\beta})}{1 + \exp(\beta_0 + \boldsymbol{x}^T\boldsymbol{\beta})}, \tag{2.5}$$

where

$$\beta_0 = -\frac{1}{2}\boldsymbol{\mu}_1'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1 + \frac{1}{2}\boldsymbol{\mu}_0\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_0 + \log\left(\frac{P(y = 1)}{P(y = 0)}\right),$$
$$\boldsymbol{\beta} = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0).$$

It is seen that model (2.5) is of the general form (2.4), with $F$ being the logistic distribution function $F(\eta) = \exp(\eta)/(1 + \exp(\eta))$. A more general form is considered in Exercise 2.1.

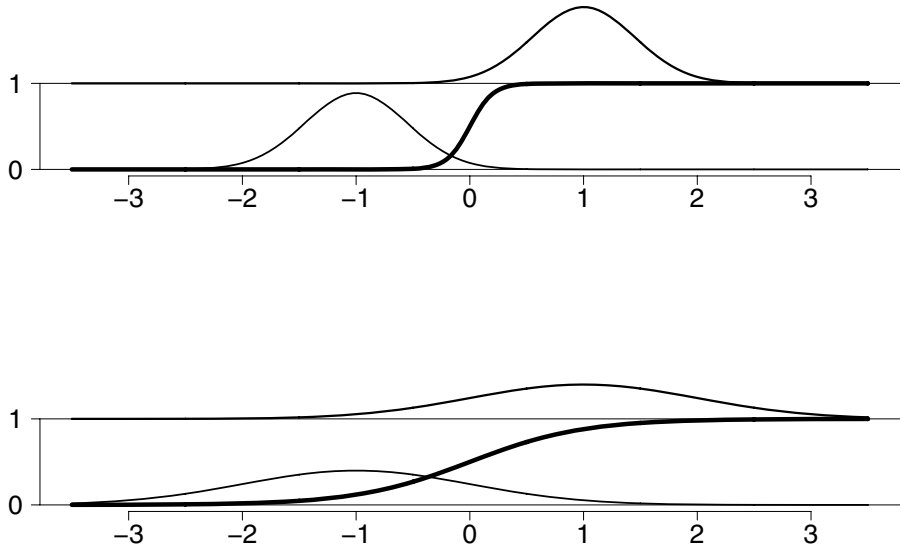


FIGURE 2.2: Logistic regression model resulting from conditionally normally distributed $x|y = i \sim \mathrm{N}(\pm 1, \sigma^2)$, $\sigma^2 = 0.16$ (upper panel), $\sigma^2 = 1$ (lower panel).

To illustrate the functional form that is specified by the logistic model, let us consider the simple case of a one-dimensional variable $x$. The derivation from the normal distribution is illustrated in Figure 2.2. Let $x$ be normally distributed with $x|y = 0 \sim \mathrm{N}(-1, \sigma^2), x|y = 1 \sim \mathrm{N}(1, \sigma^2)$; then one obtains $\beta = 2/\sigma^2$. It is seen from Figure 2.2 how the response is determined by the variance within the subpopulations. The resulting logistic response function is rather flat for large $\sigma(\sigma = 1)$ and distinctly steeper for smaller $\sigma(\sigma = 0.4)$. Thus, if the groups determined by $y = 0$ and $y = 1$ are well separated by the covariate $x$ (small $\sigma$), the response $\pi(x)$ varies strongly with $x$.

## Basic Form of Binary Regression Models

In structured regression models one often distinguishes between two components, the *structural component* and the *random component*. A general form of the structural component is $\mu = h(\eta(\boldsymbol{x}))$, where $\mu$ denotes the mean, $h$ is a transformation function, and $\eta(\boldsymbol{x})$ denotes a structured predictor. In the case of a binary variable, the structural component that specifies the mean response has exactly this form, given by

$$\pi_i = F(\beta_0 + \boldsymbol{x}_i^T \boldsymbol{\beta}), \tag{2.6}$$

with the linear predictor $\eta(\boldsymbol{x}_i) = \beta_0 + \boldsymbol{x}_i^T \boldsymbol{\beta}$ and the distribution function $F$. For a binary response $y_i$, the *random component* that describes the variation of the response is already specified by (2.6). The distribution of $y_i$ is determined uniquely if $\pi_i$ is fixed. In particular, the variance is a simple function of $\pi_i$, since $\mathrm{var}(y_i) = \pi_i(1 - \pi_i)$.

## 2.3 The Logit Model

In the following we consider the logit model, which is the most widely used model in binary regression. Alternative models will be considered in Chapter 5.

### 2.3.1 Model Representations

In the previous section the logit model was derived from the assumption that covariates are normally distributed given the response categories. Collecting the predictors in $\boldsymbol{x}$, where an intercept is included, the basic form is given by

$$\pi(\boldsymbol{x}) = F(\eta(\boldsymbol{x})),$$

with $F(\eta) = \exp(\eta)/(1+\exp(\eta))$ and the linear predictor $\eta(\boldsymbol{x}) = \boldsymbol{x}^T\boldsymbol{\beta}$. Simple derivation shows that the following equivalent representations of the model hold.

**Binary Logit Regression Model**

$$\pi(\boldsymbol{x}) = \frac{\exp(\boldsymbol{x}^T\boldsymbol{\beta})}{1+\exp(\boldsymbol{x}^T\boldsymbol{\beta})}$$

or

$$\frac{\pi(\boldsymbol{x})}{1-\pi(\boldsymbol{x})} = \exp(\boldsymbol{x}^T\boldsymbol{\beta})$$

or

$$\log\left(\frac{\pi(\boldsymbol{x})}{1-\pi(\boldsymbol{x})}\right) = \boldsymbol{x}^T\boldsymbol{\beta}$$

The different forms of the model focus on different aspects of the dependence of the response on the covariates. The first form shows how the response probability $\pi(\boldsymbol{x})$ is determined by the covariates $\boldsymbol{x}$. The second form shows the dependence of the odds $\pi(\boldsymbol{x})/(1-\pi(\boldsymbol{x}))$ on the covariates. The third form shows that in the case of the logit model, the logits $\log(\pi(\boldsymbol{x}))/(1-\pi(\boldsymbol{x}))$ depend linearly on the covariates.

### 2.3.2 Logit Model with Continuous Predictor

In the following we consider first properties of the model for the simple case of univariate predictors. The logistic regression model for a one-dimensional covariate $x$ postulates a monotone association between $\pi(x) = P(y=1|x)$ and a covariate $x$ of the form

$$\pi(x) = \frac{\exp(\beta_0 + x\beta)}{1+\exp(\beta_0+x\beta)}, \tag{2.7}$$

where $\beta_0, \beta$ are unknown parameters. The function that determines the form of the response is the logistic distribution function $F(\eta) = \exp(\eta)/(1+\exp(\eta))$. The functional form of $F$ is given in Figure 2.3, where it is seen that $F$ is strictly monotone with $F(0) = 0.5$. In addition, $F$ is symmetric, that is, $F(\eta) = 1 - F(-\eta)$ holds for all $\eta$. $F$ transforms the linear predictor $\beta_0 + x\beta$ such that $\pi(x)$ is between 0 and 1 and $\pi(x)$ is monotone in $x$.

From $\pi(x) = F(\beta_0 + x\beta)$ one obtains that $\pi(x) = 0.5$ if $\beta_0 + x\beta = 0$ or, equivalently, if $x = -\beta_0/\beta$. Thus $x = -\beta_0/\beta$ may be used as an anchor point on the $x$-scale where

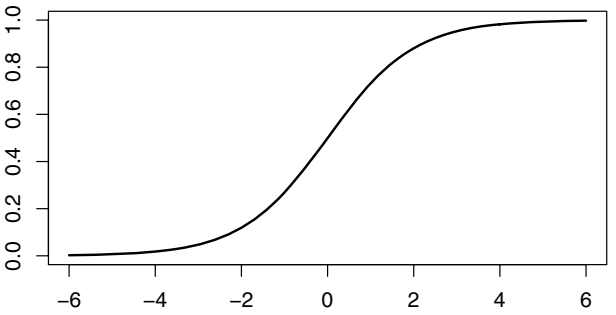


FIGURE 2.3: Logistic distribution function $F(\eta) = \exp(\eta)/(1 + \exp(\eta))$.

$\pi(x) = 0.5$. The slope of the function $\pi(x)$ is essentially determined by $\beta$. If $\beta$ is large and positive, the probability increases strongly for increasing $x$. If $\beta$ is negative, it decreases with the decrease being stronger if $\beta$ is very small. Figure 2.4 shows the function $\pi(x)$ for several values of $\beta$.

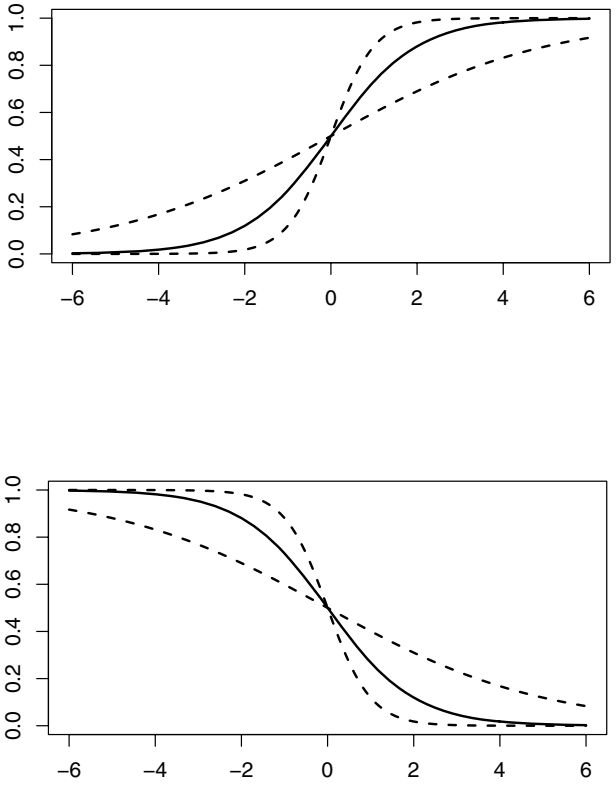


FIGURE 2.4: Response curve $\pi(x) = F(\beta_0 + \beta x)$ for $\beta_0 = 0$ and $\beta = 0.4$, $\beta = 1$, $\beta = 2$ (top) and for $\beta_0 = 0$ and $\beta = -0.4$, $\beta = -1$, $\beta = -2$ (bottom).

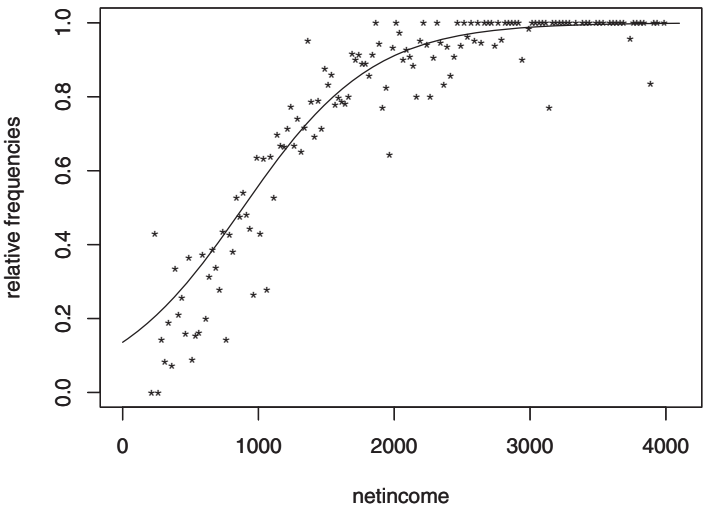


FIGURE 2.5: Car in household against income in Euros.

Example 2.1: Car in Household

In Figure 2.5, the logit model is fitted to the household data, where $\pi(x) = P(y = 1|x)$ is the probability that at least one car is owned. The estimates are $\hat{\beta}_0 = -1.851, \hat{\beta} = 0.00209$. The model suggests a distinct increase of the probability with increasing net income. Figure 2.5 also shows the relative frequencies in intervals of length 50. It is seen that the fit is quite good for higher income but less satisfactory for low income. □

It is often useful to work with the alternative representations of the model:

$$\frac{\pi(x)}{1-\pi(x)} = \exp(\beta_0 + x\beta) \tag{2.8}$$

and

$$\log(\frac{\pi(x)}{1-\pi(x)}) = \beta_0 + x\beta. \tag{2.9}$$

Representation (2.8) is based on the odds $\pi(x)/(1-\pi(x)) = P(y = 1|x)/P(y = 0|x)$, whereas the left-hand side of (2.9) represent the log-odds or logits for a given $x$. Thus (2.8) may be seen as the odds representation of the model while (2.9) gives the logit representation. Since it is easier to think in odds than in log-odds, most program packages give an estimate of $e^{\beta}$ in addition to an estimate of $\beta$.

When interpreting parameters for the logit model one has to take into account the logit transformation. For the linear model $\mathrm{E}(y|x) = \beta_0 + x\beta$, the parameter $\beta$ is simply the change in mean response if the $x$-value increases by one unit, that is, $\beta = \mathrm{E}(y|x+1) - \mathrm{E}(y|x)$. For the logit model, the corresponding change is measured in logits. Let

$$\mathrm{logit}(x) = \log(\gamma(x)) = \log\left(\frac{\pi(x)}{1-\pi(x)}\right)$$

denote the logit or log-odds at value $x$. Then, from $\mathrm{logit}(x) = \beta_0 + x\beta$, it follows that $\beta$ is given by

$$\beta = \mathrm{logit}(x+1) - \mathrm{logit}(x).$$

Thus $\beta$ measures the change in logits if $x$ is increased by one unit. Alternatively, one might use the form (2.8) of the logit model:

$$\pi(x)/(1-\pi(x)) = \exp(\beta_0 + x\beta) = \mathrm{e}^{\beta_o}(\mathrm{e}^{\beta})^x.$$

It is seen that $\mathrm{e}^{\beta_o}$ represents the odds at value $x = 0$ and the odds change by the factor $\mathrm{e}^{\beta}$ if $x$ increases by one unit. With $\gamma(x) = \pi(x)/(1-\pi(x))$ denoting the odds at a covariate value $x$, one obtains

$$e^{\beta} = \frac{\pi(x+1)/(1-\pi(x+1))}{\pi(x)/(1-\pi(x))} = \frac{\gamma(x+1)}{\gamma(x)}. \tag{2.10}$$

Thus $e^{\beta}$ is the factor by which the odds $\gamma(x)$ increase or decrease (depending on $\beta > 0$ or $\beta < 0$) if $x$ is increased by one unit. As a proportion between $\gamma(x+1)$ and $\gamma(x)$, the term $\mathrm{e}^{\beta}$ is an *odds ratio*. The odds ratio is a measure of the dependence of $y$ on $x$. It reflects how strong the odds change if $x$ increases by one unit. It is noteworthy that the logit model assumes that the odds ratios given by (2.10) do not depend on $x$. Therefore, interpretation of $\beta$ or $\mathrm{e}^{\beta}$ is very simple.

Let us consider the car data for several scalings of net income. The first model uses income in Euros as the predictor. Alternatively, one might center income around some fixed value, say $1500$, such that the variable $x$ corresponds to *net income* $- 1500$. As a third version we consider the covariate $(\textit{net income} - 1500)/1000$ such that one unit corresponds to $1000$ Euros. The estimates for all three models are given in Table 2.1. For the first model, $\mathrm{e}^{\hat{\beta}_0} = 0.157$ corresponds to the odds that there is a car in the household (instead of no car) for zero income. For income centered around $1500$, $\mathrm{e}^{\hat{\beta}_0} = 3.6$ refers to the odds at income $1500$ Euros. In the first two models the change in odds by increasing income by one Euro is determined by the factor $\mathrm{e}^{\hat{\beta}} = 1.002$, which does not help much in terms of interpretation. For a rescaled covariate, measured in $1000$ Euros, the factor is $\mathrm{e}^{\hat{\beta}} = 8.064$, which gives some intuition for the increase in odds when income is increased by one unit, which means $1000$ Euros.

TABLE 2.1: Parameter for logit model with predictor income.

| Income in Euros | |
|---|---|
| Parameter | Odds |
| $\hat{\beta}_0 = -1.851$ | $\mathrm{e}^{\hat{\beta}_0} = 0.157$ |
| $\hat{\beta} = 0.00209$ | $\mathrm{e}^{\hat{\beta}} = 1.002$ |
| **Income in Euros, centered at Euro 1500** | |
| Parameter | Odds |
| $\hat{\beta}_0 = 1.281$ | $\mathrm{e}^{\hat{\beta}_0} = 3.600$ |
| $\hat{\beta} = 0.00209$ | $\mathrm{e}^{\hat{\beta}} = 1.002$ |
| **Income in thousands of Euros, centered at Euro 1500** | |
| Parameter | Odds |
| $\hat{\beta}_0 = 1.281$ | $\mathrm{e}^{\hat{\beta}_0} = 3.600$ |
| $\hat{\beta} = 2.088$ | $\mathrm{e}^{\hat{\beta}} = 8.069$ |

An alternative way to get some understanding of the parameter values has been suggested by Cox and Snell (1989). It may be shown that $1/\beta$ is approximately the distance between the

$75\%$ point and the $50\%$ point of the estimated logistic curve. The distance between the $95\%$ point and the $50\%$ point is approximately $3/\beta$. From $1/\hat{\beta} = 478$ for the first model of the car data (and $1/\hat{\beta} = 0.478$ for the scaling in 1000 Euros) one gets some intuition for the increase in the logistic curve without having to plot it. As a general rule one might keep the following in mind:

Change in $x$ by $\pm 1/\beta$: Change in $\pi$ from $\pi = 0.5$ to $0.5 \pm 0.23$,
Change in $x$ by $\pm 3/\beta$: Change in $\pi$ from $\pi = 0.5$ to $0.5 \pm 0.45$.

The value $\pi = 0.5$ itself occurs for the $x$-value $-\beta_0/\beta$.

## Multivariate Predictor

When several continuous covariates are available the linear predictor may have the form

$$\eta(\boldsymbol{x}) = \beta_0 + x_1\beta_1 + \ldots + x_m\beta_m.$$

The interpretation is the same as for univariate predictors; however, one should be aware that the other variables are present. From

$$\text{logit}(\boldsymbol{x}) = \log \frac{\pi(\boldsymbol{x})}{1 - \pi(\boldsymbol{x})} = \beta_0 + x_1\beta_1 + \ldots + x_m\beta_m$$

and

$$\gamma(x) = \frac{\pi(\boldsymbol{x})}{1 - \pi(x)} = \mathrm{e}^{\beta_0}(\mathrm{e}^{\beta_1})^{x_1} \ldots (\mathrm{e}^{\beta_m})^{x_m}$$

one derives immediately that $\beta_j$ corresponds to the *additive change in logits* when variable $\beta_j$ is increased by one unit *while all other variables are kept fixed*,

$$\beta_j = \text{logit}(x_1, \ldots, x_j + 1, \ldots, x_m) - \text{logit}(x_1, \ldots, x_m),$$

and $\mathrm{e}^{\beta_j}$ corresponds to the multiplicative change in odds when $x_j$ is increased by one unit,

$$\mathrm{e}^{\beta_i} = \frac{\gamma(x_1, \ldots, x_j + 1, \ldots, x_m)}{\gamma(x_1, \ldots, x_m)}.$$

It is essential that the effects of predictors measured by $\beta_j$ or $\mathrm{e}^{\beta_j}$ assume that all other values are kept fixed. A strong assumption is implied here, namely that the effect of a covariate is the same, whatever values the other variables take. If that is not the case, one has to include interaction effects between predictors (see Chapter 4).

### Example 2.2: Vasoconstriction

A classical example in logistic regression is the vasoconstriction data that were used by Finney (1947). The data (Figure 2.2) were obtained in a carefully controlled study in human physiology where a reflex "vasoconstriction" may occur in the skin of the digits after taking a single deep breath. The response $y$ is the occurrence $(y = 1)$ or non-occurrence $(y = 0)$ of vasoconstriction in the skin of the digits of one subject after he or she inhaled a certain volume of air at a certain rate. The responses of three subjects are available. The first contributed 9 responses, the second contributed 8 responses, and the third contributed 22 responses.

Although the data represent repeated measurements, usually independent observations were assumed. The effect of the volume of air and inspiration rate on the occurrence of vasoconstriction may be based on the binary logit model

$$\text{logit}(\boldsymbol{x}) = \log(\frac{\pi(\boldsymbol{x})}{1 - \pi(\boldsymbol{x})}) = \beta_0 + \text{volume}\beta_1 + \text{rate}\beta_2.$$

Then interpretation of the parameters refers to the multiplicative change in odds when the variables are increased by one unit. Alternatively, one can apply the logit model with log-transformed variables,

$$\text{logit}(\boldsymbol{x}) = \log(\frac{\pi(\boldsymbol{x})}{1-\pi(\boldsymbol{x})}) = \beta_0 + \log(\text{volume})\beta_1 + \log(\text{rate})\beta_2,$$

which is equivalent to

$$\frac{\pi(\boldsymbol{x})}{1-\pi(\boldsymbol{x})} = \text{e}^{\beta_0}\,\text{volume}^{\beta_1}\,\text{rate}^{\beta_2}.$$

Then the effect of the covariates volume and rate is multiplicative on the odds. When rate changes by the factor $c$ the odds change by the factor $c^{\beta_1}$. The maximum likelihood estimates for the log-transformed covariates are $\beta_0 = -2.875$, $\beta_1 = 5.179$, $\beta_2 = 4.562$. The logit model for the original variables yields $\beta_0 = -9.5296$, $\beta_1 = 3.8822$, $\beta_2 = 2.6491$. In both models the covariates are highly significant. □

TABLE 2.2: Vasoconstriction data.

| Index | Volume | Rate | Y | Index | Volume | Rate | Y |
|---|---|---|---|---|---|---|---|
| 1 | 3.70 | 0.825 | 1 | 20 | 1.80 | 1.800 | 1 |
| 2 | 3.50 | 1.090 | 1 | 21 | 0.40 | 2.000 | 0 |
| 3 | 1.25 | 2.500 | 1 | 22 | 0.95 | 1.360 | 0 |
| 4 | 0.75 | 1.500 | 1 | 23 | 1.35 | 1.350 | 0 |
| 5 | 0.80 | 3.200 | 1 | 24 | 1.50 | 1.360 | 0 |
| 6 | 0.70 | 3.500 | 1 | 25 | 1.60 | 1.780 | 1 |
| 7 | 0.60 | 0.750 | 0 | 26 | 0.60 | 1.500 | 0 |
| 8 | 1.10 | 1.700 | 0 | 27 | 1.80 | 1.500 | 1 |
| 9 | 0.90 | 0.750 | 0 | 28 | 0.95 | 1.900 | 0 |
| 10 | 0.90 | 0.450 | 0 | 29 | 1.90 | 0.950 | 1 |
| 11 | 0.80 | 0.570 | 0 | 30 | 1.60 | 0.400 | 0 |
| 12 | 0.55 | 2.750 | 0 | 31 | 2.70 | 0.750 | 1 |
| 13 | 0.60 | 3.000 | 0 | 32 | 2.35 | 0.030 | 0 |
| 14 | 1.40 | 2.330 | 1 | 33 | 1.10 | 1.830 | 0 |
| 15 | 0.75 | 3.750 | 1 | 34 | 1.10 | 2.200 | 1 |
| 16 | 2.30 | 1.640 | 1 | 35 | 1.20 | 2.000 | 1 |
| 17 | 3.20 | 1.600 | 1 | 36 | 0.80 | 3.330 | 1 |
| 18 | 0.85 | 1.415 | 1 | 37 | 0.95 | 1.900 | 0 |
| 19 | 1.70 | 1.060 | 0 | 38 | 0.75 | 1.900 | 0 |
| | | | | 39 | 1.30 | 1.625 | 1 |

### 2.3.3 Logit Model with Binary Predictor

When both the predictor and the response variable are binary the usual representation of data is in a (2×2)-contingency table, as in the next example.

#### Example 2.3: Duration of Unemployment

A simple example of a dichotomous covariate is given by the contingency table in Table 2.3, which shows data from a study on the duration of unemployment. The duration of unemployment is given in two categories, short-term unemployment (less than 6 months) and long-term employment (more than 6 months). Subjects are classified with respect to gender and duration of unemployment. Gender is considered as the explanatory variable and duration as the response variable. □

When a binary variable is used as an explanatory variable it has to be coded. Thus, instead of $G = 1$ for males and $G = 2$ females, one uses a dummy variable. There are two forms in common use, (0-1)-coding and effect coding.

TABLE 2.3: Cross-classification of gender and duration of unemployment.

| | | Duration | | | | |
|---|---|---|---|---|---|---|
| | | ≤ 6 Months | > 6 Months | Marginals | Odds | Log-Odds |
| Gender | male | 403 | 167 | 570 | 2.413 | 0.881 |
| | female | 238 | 175 | 413 | 1.360 | 0.307 |

## Logit Model with (0-1)-Coding of Covariates

Let $x_G$ denote the dummy variable for $G$ in (0-1)-coding, given by $x_G = 1$ for males and $x_G = 0$ for females. With $y$ denoting the dichotomous response, specified by $y = 1$ for short-term unemployment and $y = 0$ for long-term unemployment and $\pi(x_G) = P(y = 1|x_G))$, the corresponding logit model has the form

$$\log\left(\frac{\pi(x_G)}{1-\pi(x_G)}\right) = \beta_0 + x_G\beta \quad \frac{\pi(x_G)}{1-\pi(x_G)} = \mathrm{e}^{\beta_0}\left(e^{\beta}\right)^{x_G}. \tag{2.11}$$

It is immediately seen that $\beta_0$ is given by

$$\beta_0 = \log\left(\frac{\pi(x_G = 0)}{1-\pi(x_G = 0)}\right) = \mathrm{logit}(x_G = 0),$$

which corresponds to the logits in the reference category, where $x_G = 0$. The effect of the covariate takes the form

$$\begin{aligned}\beta &= \log\left(\frac{\pi(x_G = 1)}{1-\pi(x_G = 1)}\right) - \log\left(\frac{\pi(x_G = 0)}{1-\pi(x_G = 0)}\right)\\ &= \mathrm{logit}(x_G = 1) - \mathrm{logit}(x_G = 0),\end{aligned}$$

which is the *additive* change in logits for the transition from $x_G = 0$ to $x_G = 1$. A simpler interpretation holds for the transformed parameters

$$\mathrm{e}^{\beta_0} = \frac{\pi(x_G = 0)}{1-\pi(x_G = 0)} = \gamma(x_G = 0),$$

which corresponds to the odds in the reference category $x_G = 0$, and

$$e^{\beta} = \frac{\pi(x_G = 1)/(1-\pi(x_G = 1)}{\pi(x_G = 0)/(1-\pi(x_G = 0))} = \frac{\gamma(x_G = 1)}{\gamma(x_G = 0)} = \gamma(1|0),$$

which corresponds to the odds ratio between $x_G = 1$ and $x_G = 0$. It may also be seen as the factor by which the logits change if $x_G = 0$ is replaced by $x_G = 1$.

**Logit Model with (0-1)-Coding**

$$\beta_0 = \mathrm{logit}(x_G = 0) \qquad \mathrm{e}^{\beta_0} = \gamma(x_G = 0)$$

$$\beta = \mathrm{logit}(x_G = 1) - \mathrm{logit}(x_G = 0) \qquad e^{\beta} = \frac{\gamma(x_G = 1)}{\gamma(x_G = 0)}$$

Example 2.4: Duration of Unemployment

For the contingency table in Example 2.3 one obtains for the logit model with gender given in (0-1)-coding

$$\hat{\beta}_0 = 0.307, \quad \mathrm{e}^{\hat{\beta}_0} = 1.360, \qquad \hat{\beta} = 0.574, \quad \mathrm{e}^{\hat{\beta}} = 1.774.$$

It is usually easier to interpret $\mathrm{e}^{\beta}$ rather than $\beta$ itself. While $\beta$ refers to the change in logits, $\mathrm{e}^{\beta}$ gives the odd ratio. Thus $\mathrm{e}^{\hat{\beta}} = 1.774$ means that the odds of short-time unemployment for men are almost twice the odds for women. □

## Logit Model with Effect Coding

An alternative form of the logit model is given by

$$\log\left(\frac{P(y=1|G=i)}{1-P(y=1|G=i)}\right) = \beta_0 + \beta_i.$$

Here $\beta_i$ is the effect of the factor G if $G = i$. But since only two logits are involved, namely, for $G = 1$ and $G = 2$, one needs a restriction to make the parameters $\beta_0, \beta_1, \beta_2$ identifiable. The symmetric constraint $\beta_1 + \beta_2 = 0$ (or $\beta_2 = -\beta_1$) is equivalent to the model

$$\log\left(\frac{\pi(x_G)}{1-\pi(x_G)}\right) = \beta_0 + x_G\beta,$$

where $\beta = \beta_1$ and $x_G$ is given in effect coding, that is, it is given by $x_G = 1$ for males and $x_G = -1$ for females. With $\gamma(x_G) = \pi(x_G)/(1-\pi(x_G))$ one obtains $\gamma(x_G = 1) = \mathrm{e}^{\beta_0}\,\mathrm{e}^{\beta}, \gamma(x_G = -1) = \mathrm{e}^{\beta_0}\,\mathrm{e}^{-\beta}$. Simple computation shows that

$$\beta_0 = \frac{1}{2}(\mathrm{logit}(x_G = 1) + \mathrm{logit}(x_G = -1))$$

is the *arithmetic* mean over the logits of categories $G = 1$ and $G = 2$. Therefore,

$$\mathrm{e}^{\beta_0} = (\gamma(x_G = 1)\gamma(x_G = -1))^{1/2}$$

is the *geometric* mean over the odds $\gamma(x_G = 1)$ and $\gamma(x_G = -1)$, representing some sort of "baseline" odds. For $\beta$ one obtains

$$\beta = \frac{1}{2}(\mathrm{logit}(x_G = 1) - \mathrm{logit}(x_G = -1)),$$

which is half the change in logits for the transition from $x_G = -1$ to $x_G = 1$. Thus

$$e^{\beta} = (\gamma(x_G = 1)/\gamma(x_G = -1))^{\frac{1}{2}} = \gamma(x_G = 1|x_G = -1)^{1/2}$$

is the square root of the odds ratio. Since $\gamma(x_G = 1) = \mathrm{e}^{\beta_0}\,\mathrm{e}^{\beta}$, the term $\mathrm{e}^{\beta}$ is the factor that modifies the "baseline" odds $\mathrm{e}^{\beta_0}$ to obtain $\gamma(x_G = 1)$. To obtain $\gamma(x_G = -1)$ one has to use the factor $\mathrm{e}^{-\beta}$.

**Logit Model with Effect Coding**

$$\beta_0 = \frac{1}{2}(\mathrm{logit}(x_G = 1) + \mathrm{logit}(x_G = -1)) \qquad \mathrm{e}^{\beta_0} = (\gamma(x_G = 1)\gamma(x_G = -1))^{1/2}$$

$$\beta = \frac{1}{2}(\mathrm{logit}(x_G = 1) - \mathrm{logit}(x_G = -1)) \qquad e^{\beta} = \gamma(x_G = 1|x_G = -1)^{\frac{1}{2}}$$

Example 2.5: Duration of Unemployment

For the data in Example 2.3 one obtains for the logit model with gender given in effect coding

$$\hat{\beta}_0 = 0.594, \quad \mathrm{e}^{\hat{\beta}_0} = 1.811, \qquad \hat{\beta} = 0.287, \quad \mathrm{e}^{\hat{\beta}} = 1.332.$$

As in (0-1)-coding it is easier to interpret $\mathrm{e}^{\beta}$ rather than $\beta$ itself. While $\hat{\beta}_0$ is the average (arithmetic mean) over logits, $\mathrm{e}^{\hat{\beta}_0} = 1.811$ is the geometric mean over the odds. One obtains the odds in the male population by applying the factor $\mathrm{e}^{\hat{\beta}} = 1.332$, which shows that the odds are better in the male population. Multiplication with $\mathrm{e}^{-\hat{\beta}} = 1/1.332$ yields the odds in the female population. □

### 2.3.4 Logit Model with Categorical Predictor

In the more general case, categorical covariates have more than just two outcomes. As an example let us again consider the duration of unemployment given in two categories, short-term unemployment (less than 6 months) and long-term employment (more than 6 months), but now depending on education level. In Table 2.4 subjects are classified with respect to education level and duration of unemployment. Education level is considered the explanatory variable and duration the response variable. Interpretation of the parameters is similar to the binary covariate case. In the following we will again distinguish between the the restrictions yielding (0-1)-coding and effect coding.

TABLE 2.4: Cross-classification of level of education and duration of unemployment.

| | | Duration | | |
|---|---|---|---|---|
| | | $\leq 6$ Months | $> 6$ Months | |
| No specific training | 1 | 202 | 96 | 298 |
| Low level training | 2 | 307 | 162 | 469 |
| High level training | 3 | 87 | 66 | 153 |
| University degree | 4 | 45 | 18 | 63 |

## Logit Model with (0-1)-Coding

Consider a categorical covariable or factor $A$ with categories $A \in \{1, \ldots, I\}$ and let $\pi(i) = P(y = 1|A = i)$ denote the response probability. Then a general form of the logit model is given by

$$\log(\frac{\pi(i)}{1-\pi(i)}) = \beta_0 + \beta_i \quad \text{or} \quad \frac{\pi(i)}{1-\pi(i)} = \mathrm{e}^{\beta_0}\ \mathrm{e}^{\beta_i}\,. \tag{2.12}$$

Since one has only $I$ logits, $\log(\pi(i)/(1-\pi(i))), i = 1, ..., I$, but $I+1$ parameters $\beta_0, \beta_1, ..., \beta_I$, a constraint for the parameters is necessary. One possibility is to set $\beta_I = 0$, thereby defining I as the reference category. Then the intercept $\beta_0 = \log(\pi(I)/(1-\pi(I)))$ is the logit for the reference category $I$ and

$$\beta_i = \log\left(\frac{\pi(i)}{1-\pi(i)}\right) - \log\left(\frac{\pi(I)}{1-\pi(I)}\right) = \log\,\gamma(i|I)$$

is the additive change in logits if $A = I$ is replaced by $A = i$, which is equivalent to the logarithm of the odds ratio of $A = i$ to $A = I$. Alternatively, one may consider the exponentials

$$\mathrm{e}^{\beta_0} = \frac{\pi(I)}{1-\pi(I),} \qquad\qquad \mathrm{e}_i^{\beta} = \gamma(i|I).$$

The latter is the odds ratio of $A = i$ to $A = I$. The model (2.12) may be given in the form of a regression model by using the dummy variables $x_{A(i)} = 1$ if $A = i$, and $x_{A(i)} = 0$ otherwise. Then one has a model with multiple predictor $x_{A(1)}, \ldots, x_{A(I-1)}$

$$\log \frac{\pi(i)}{1-\pi(i)} = \beta_0 + x_{A(1)}\beta_1 + \cdots + x_{A(I-1)}\beta_{I-1}.$$

Example 2.6: Duration of Unemployment with Predictor Education Level

For the data in Table 2.4 one obtains the odds and the parameters for the logit model in (0-1)-coding as given in Table 2.5. □

TABLE 2.5: Odds and parameters for logit model with predictor education level.

| Level | Short Term | Long Term | Odds $e^{\beta_i}$ | $\beta_i$ | Standard Error |
|---|---|---|---|---|---|
| 1 | 202 | 96 | $\gamma(1/4) = 0.843$ | $-0.170$ | 0.30 |
| 2 | 307 | 162 | $\gamma(2/4) = 0.761$ | $-0.273$ | 0.29 |
| 3 | 87 | 66 | $\gamma(3/4) = 0.529$ | $-0.637$ | 0.32 |
| 4 | 45 | 18 | $\gamma(4/4) = 1$ | 0 | 0 |

Parameters with (0-1)-Coding

$$\beta_0 = \log\left(\frac{\pi(I)}{1-\pi(I)}\right) \qquad e^{\beta_0} = \frac{\pi(I)}{1-\pi(I)}$$
$$\beta_i = \log(\gamma(i|I)) \qquad e^{\beta}_i = \gamma(i|I)$$

## Logit Model with Effect Coding

An alternative restriction on the parameters in model (2.12)) is the symmetric restriction $\beta_1 + \ldots + \beta_I = 0$. By considering $\beta_I = -\beta_1 - \ldots - \beta_{I-1}$ one obtains

$$\log\left(\frac{\pi(i)}{1-\pi(i)}\right) = \beta_0 + \beta_i\ , \quad i = 1, \ldots, I-1,$$
$$\log\left(\frac{\pi(I)}{1-\pi(I)}\right) = \beta_0 - \beta_1 - \ldots - \beta_{I-1}.$$

This is equivalent to the form

$$\log\left(\frac{\pi(i)}{1-\pi(i)}\right) = \beta_0 + x_{A(1)}\beta_1 + \ldots + x_{A(I-1)}\beta_{I-1},$$

where the $x_{A(i)}$'s are given in effect coding, that is, $x_{A(i)} = 1$ if $A = i$; $x_{A(i)} = -1$ if $A = k$; and $x_{A(i)} = 0$ otherwise. One obtains for the parameters

$$\beta_0 = \frac{1}{I}\sum_{i=1}^{I} \log\left(\frac{\pi(i)}{1-\pi(i)}\right) = \frac{1}{I}\sum_{i=1}^{I} \log(\gamma(i)),$$
$$\beta_i = \log\left(\frac{\pi(i)}{1-\pi(i)}\right) - \beta_0 = \log(\gamma(i)) - \beta_0.$$

Therefore $\beta_0$ corresponds to the arithmetic mean of logits across all categories of $A$ and is some sort of a baseline logit, whereas $\beta_i$ represents the (additive) deviation of category $i$ from this baseline. For $e^{\beta_0}$ one obtains

$$e^{\beta_0} = (\gamma(1) \cdot \ldots \cdot \gamma(I))^{1/I},$$

which is the geometric mean over the odds. Since $\gamma(i) = e^{\beta_0} e^{\beta_i}$, $e^{\beta_i}$ represents the factor that transforms the baseline odds into the odds of population $i$.

**Parameters with Effect Coding**

$$\begin{array}{ll} \beta_0 = \frac{1}{I}\sum_{i=1}^{I} \log(\gamma(i)) & e^{\beta_0} = (\gamma(1) \cdot \ldots \cdot \gamma(I))^{1/I} \\ \beta_i = \log(\gamma(i)) - \beta_0 & e^{\beta_i} = \gamma(i)\, e^{-\beta_0} \end{array}$$

## Logit Model with Several Categorical Predictors

If more than one predictor is included, the interpretation of parameters is pretty much the same. Let us consider two predictors, $A \in \{1, \ldots, I\}$ and $B \in \{1, \ldots, J\}$, and the *main effect model* that contains dummies but no interactions. With $\pi(A = i, B = j) = P(y = 1 | A = i, B = j)$ the model with the reference category $I$ for factor $A$ and $J$ for factor $B$ is given by

$$\begin{aligned} &\log\left(\frac{\pi(A = i, B = j)}{1 - \pi(A = i, B = j)}\right) \\ &= \beta_0 + x_{A(1)}\beta_{A(1)} + ... + x_{A(I-1)}\beta_{A(I-1)} + x_{B(1)}\beta_{B(1)} + ... + x_{B(J-1)}\beta_{B(J-1)}, \end{aligned}$$

where $x_{A(i)}$, $x_{B(j)}$ are 0-1 dummy variables. It is easily derived (Exercise 2.2) that the parameters have the form given in the following box.

**Parameters with (0-1)-Coding**

$$e^{\beta_0} = \frac{\pi(A = I, B = J)}{1 - \pi(A = I, B = J)} = \gamma(A = I, B = J),$$

odds for $A = I, B = J$

$$e^{\beta_{A(i)}} = \frac{\pi(A = i, B = j)/(1 - \pi(A = i, B = j))}{\pi(A = I, B = j)/(1 - \pi(A = I, B = j))} = \frac{\gamma(A = i, B = j)}{\gamma(A = I, B = j)},$$

odds ratio compares $A = i$ to $A = I$, any $B = j$

$$e^{\beta_{B(j)}} = \frac{\pi(A = i, B = j)/(1 - \pi(A = i, B = j))}{\pi(A = i, B = J)/(1 - \pi(A = i, B = J))} = \frac{\gamma(A = i, B = j)}{\gamma(A = i, B = J),}$$

odds ratio compares $B = j$ to $B = J$, any $A = i$

The exponential of $\beta_0$ corresponds to the *odds* for the reference category $(A = I, B = J)$, the exponential of $\beta_{A(i)}$ corresponds to the *odds ratio* between $A = i$, and the reference category $A = I$ for any category of $B$. For $\beta_{B(j)}$, the corresponding odds ratio does not depend on $A$. It should be noted that this simple interpretation no longer holds if interactions are included.

### 2.3.5 Logit Model with Linear Predictor

In the general case one has several covariates, some of which are continuous and some of which are categorical. Then one may specify a logit model $\text{logit}(\boldsymbol{x}) = \eta(\boldsymbol{x})$ with the linear predictor given by

$$\eta(\boldsymbol{x}) = \boldsymbol{x}^T\boldsymbol{\beta} = \beta_0 + x_1\beta_1 + \ldots + x_p\beta_p,$$

which yields

$$\log(\frac{\pi(\boldsymbol{x})}{1-\pi(\boldsymbol{x})}) = \beta_0 + x_1\beta_1 + \ldots + x_p\beta_p.$$

The $x$-values in the linear predictor do not have to be the original variables. When variables are categorical the $x$-values represent dummy variables; when variables are continuous they can represent transformations of the original variables. Components within the linear predictor can be

| | |
|---|---|
| $x_i, x_i^2, x_i^3, \ldots,$ | main effects or polynomial terms built from variable $x_i$; |
| $x_{A(i)}, \ldots, x_{A(I-1)},$ | main effects (dummy variables) as transformations of categorical variable $A \in \{1, \ldots, I\}$; |
| $x_{A(i)} \cdot x_{B(j)},$ $i = 1, \ldots, I-1,$ $j = 1, \ldots, J-1;$ | interactions (products of dummy variables) between two factor $A \in \{1, \ldots, I\}, B \in \{1, \ldots, J\}$; |
| $x_i x_{B(j)}, j = 1, \ldots, J,$ | interactions between metrical variable $x_i$ and categorical variable $B \in \{1, \ldots, J\}$. |

One may also include products of more than two variables (or dummies), which means interactions of a higher order. The structuring of the linear predictor is discussed more extensively in Section 4.4.

## 2.4 The Origins of the Logistic Function and the Logit Model

In this chapter and in some of the following the logistic function occurs quite often. Therefore some remarks on the origins of the function seem warranted. We will follow Cramer (2003), who gives a careful account of the historical development.

A simple model for the growth of populations assumes that the growth rate $dw(t)/dt$ is directly proportional to the size $w(t)$ at time point $t$. The corresponding differential equation $dw(t)/dt = \beta w(t)$ has the solution $w(t) = c\exp(\beta t)$, where $c$ is a constant. This exponential growth model model means unopposed growth and is certainly not realistic over a wide range of time.

An alternative model that has been considered by the Belgian astronomer Quetelet (1795–1874) and the mathematician Verhulst (1804–1849) may be derived from a differential equation with an extra term. It is assumed that the rate is given by

$$dw(t)/dt = \tilde{\beta} w(t)(S - w(t)),$$

where $S$ denotes the upper limit or saturation level. The growth rate is now determined by the size $w(t)$ but also by the term $(S - w(t))$, which has the opposite effect, that is, the large size decelerates growth. By transforming the equation to

$$dF(t)/dt = \beta F(t)(1 - F(t)),$$

where $F(t) = w(t)/S$ and $\beta = S\tilde{\beta}$, one obtains that the solution has the form of the logistic function

$$F(t) = \exp(\alpha + \beta t)/(1 + \exp(\alpha + \beta t))$$

with constant $\alpha$. Therefore, the logistic function results as the solution of a rather simple growth model. It is widely used in modeling population size but also in marketing when the objective is the modeling of market penetration of new products.

The logit model came up much later than the probit model. Cramer (2003) traces the roots of the probit model back to Fechner (1801–1887), who is still well known to psychologists because of Fechner's law, which relates the physical strength of a stimulus and its strength as perceived by humans. It seems to have been reinvented several times until Bliss (1934) introduced the term *probit* for probability unit. The probit model used to be the classical model of bioassays. Berkson (1994) introduced as an alternative the logit model, which was easier to estimate. The logit model was not well received, but after the ideological conflict had abated the logit model was widely adopted in the late 1950s.

## 2.5 Exercises

2.1 Consider the bivariate random variable $(y, \boldsymbol{x})$, where $y \in \{0, 1\}$ is a binary variable and $\boldsymbol{x}$ is a vector of continuous random variables. Assume that $\boldsymbol{x}$ is conditionally normally distributed within the groups determined by $y = 0$ and $y = 1$, that is, $\boldsymbol{x}|y = r \sim \mathrm{N}(\boldsymbol{\mu}_r, \boldsymbol{\Sigma}_r)$, where the covariance matrix $\boldsymbol{\Sigma}_r$ depends on the category $y = r$. Use Bayes' theorem to derive the conditional distribution $y|\boldsymbol{x}$ and show that one obtains a logit model that is determined by a linear predictor, but $x$-variables themselves are included in quadratic form.

2.2 Consider a binary logit model with two factors, $A \in \{1, \ldots, I\}$ and $B \in \{1, \ldots, J\}$, and linear predictor $\eta = \beta_0 + x_{A(1)}\beta_{A(1)} + ... + x_{A(I-1)}\beta_{A(I-1)} + x_{B(1)}\beta_{B(1)} + ... + x_{B(J-1)}\beta_{B(J-1)}$. Show that the parameters can be represented as log odds and log odds ratios.

2.3 A logit model is used to model the probability of a car in household depending on the factor "type of household" (1: household includes more than one person and children, 2: household includes more than one person without children, 3: one-person household). The logit model uses two (0-1)-dummy variables for type 1 and type 2.

(a) Write down the model and interpret the parameters, which were estimated as $\beta_0 = -0.35$, $\beta_1 = 2.37$, $\beta_2 = 1.72$.

(b) Compute the parameters if category 1 is chosen as the reference category.

(c) Let the model now be given in effect coding with parameters $\tilde{\beta}_1, \tilde{\beta}_1$. Find the parameters $\tilde{\beta}_1, \tilde{\beta}_1$ and interpret.

(d) Give a general form of how parameters for a factor $A \in \{1, \ldots, k\}$ in (0-1)-coding $\beta_j$ can be transformed into parameters in effect coding $\tilde{\beta}_j$.

2.4 A logit model for the data from the previous exercise was fitted, but now including a linear effect of income. One obtains parameter estimates $\beta_0 = -2.06$, $\beta_1 = 1.34$, $\beta_2 = 0.93$, and $\beta_{income} = 0.0016$.

(a) Write down the model and interpret the parameters.

(b) Why do these parameters differ from the parameters in the previous exercise?

2.5 In a treatment study one wants to investigate the effect of age ($x_1$) in years and dose ($x_2$) in milligrams (mg) on side effects. The response is headache ($y = 1$: headache; $y = 1$: none). One fits a logit model with the predictor $\eta = \beta_0 + x_1\beta_1 + x_2\beta_2$.

(a) One obtains $\beta_0 = -3.4$, $\beta_1 = 0.02$, $\beta_2 = 0.15$. Interpret the parameters.

(b) Plot the logits of headache as a function of dose for a patient of age 40 and a patient of age 60.

(c) Plot the (approximate) probability of headache as a function of dose for a patient of age 40 and a patient of age 60.

(d) Give the probability of headache for a patient of age 40 when the dose is 5.

Assume now that an interaction effect has to be included and one has the predictor $\eta = \beta_0 + x_1\beta_1 + x_2\beta_2 + x_1x_2\beta_{12}$ with values $\beta_0 = -3.8$, $\beta_1 = 0.02$, $\beta_2 = 0.15$, $\beta_{12} = 0.005$.

(a) Interpret the effect of dose if age is fixed at 40 (60).

(b) Plot the logits of headache as a function of dose for a patient of age 40 and a patient of age 60.

(c) Plot the (approximate) probability of headache as a function of dose for a patient of age 40 and a patient of age 60.

(d) How can the effect of dose be interpreted if age is fixed at 40 (60)?

# 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 modeling, 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 $y_i$ given $\boldsymbol{x}_i$, whereas the systematic component specifies the link between the expected response and the covariates.

(1) *Random component and distributional assumptions*

Given $\boldsymbol{x}_i$, the $y_i$'s are (conditionally) independent observations from a simple exponential family. This family has a probability density function or mass function of the form

$$f(y_i|\theta_i, \phi_i) = \exp\left\{\frac{y_i\theta_i - b(\theta_i)}{\phi_i} + c(y_i, \phi_i)\right\}, \tag{3.1}$$

where

| | |
|---|---|
| $\theta_i$ | is the natural parameter of the family, |
| $\phi_i$ | is a scale or dispersion parameter, and |
| $b(.)$ and $c(.)$ | 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 its name specifies that the variables $\boldsymbol{x}_i$ enter the model in linear form by forming the linear predictor

$$\eta_i = \boldsymbol{x}_i^T \boldsymbol{\beta},$$

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

$$\mu_i = h(\eta_i) = h(\boldsymbol{x}_i^T \boldsymbol{\beta}), \tag{3.2}$$

or, equivalently, by

$$g(\mu_i) = \eta_i = \boldsymbol{x}_i^T \boldsymbol{\beta}, \tag{3.3}$$

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

$$\pi_i = \frac{\exp(\boldsymbol{x}_i^T \boldsymbol{\beta})}{1 + \exp(\boldsymbol{x}_i^T \boldsymbol{\beta})},$$

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

$$\log(\frac{\pi_i}{1 - \pi_i}) = \boldsymbol{x}_i^T \boldsymbol{\beta}, \tag{3.4}$$

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 $y_i|\boldsymbol{x}_i$;
- 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. Parametrization of specific distributions most often uses 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

$$y_i = \boldsymbol{x}_i^T \boldsymbol{\beta} + \epsilon_i$$

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

$$y_i | \boldsymbol{x}_i \sim \mathrm{N}(\mu_i, \sigma^2) \qquad \text{and} \qquad \mu_i = \eta_i = \boldsymbol{x}_i^T \boldsymbol{\beta}.$$

The form separates the distribution from the systematic component. While $y_i | \boldsymbol{x}_i \sim \mathrm{N}(\mu_i, \sigma^2)$ 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 $\mu_i = \eta_i = \boldsymbol{x}_i^T \boldsymbol{\beta}$. 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

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

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

$$\mu = \exp(\eta) = \exp(\boldsymbol{x}^T \boldsymbol{\beta}).$$

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,

$$\mu = \boldsymbol{x}^T \boldsymbol{\beta} = x_1\beta_1 + \ldots + x_p\beta_p,$$

where the change of $x_j$ by one unit means an *additive* effect of $\beta_j$ on the expectation. The modified link

$$\mu = \exp(x_1\beta_1 + \ldots + x_p\beta_p) = e^{x_1\beta_1} \cdot \ldots \cdot e^{x_p\beta_p}$$

specifies that the change of $x_j$ by one unit has a *multiplicative* effect on $\mu$ by the factor $e^{\beta_j}$, since $e^{(x_j+1)\beta_j} = e^{x_j\beta_j} e^{\beta_j}$. 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

$$f(y) = \lambda e^{-\lambda y} = \exp(-\lambda y + \log(\lambda)), \quad y \geq 0.$$

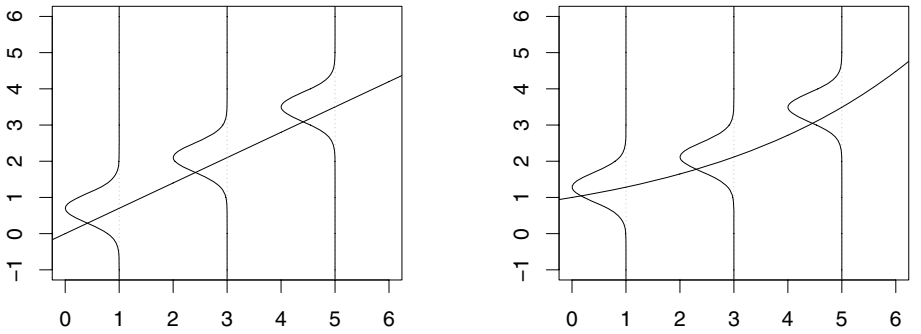


FIGURE 3.1: Normal regression with identity link (left) and with log-link (right).

With $\theta = -\lambda, \phi = 1$, and $b(\theta) = -\log(-\theta)$, the exponential distribution is of the simple exponential family type. Since the expectation and variance of the exponential distribution are given by $1/\lambda$ and $1/\lambda^2$, it is seen that in contrast to the normal distribution the variance increases with increasing expectation. Thus, although there is a fixed link between the expectation and the variance, the distribution model captures an essential property that is often found in real datasets. The so-called *canonical link*, which fulfills $\theta(\mu) = \eta$, is given by

$$g(\mu) = -\frac{1}{\mu} \quad \text{or} \quad h(\eta) = -\frac{1}{\eta}.$$

Since $\mu > 0$, the linear predictor is restricted to $\eta = \boldsymbol{x}^T\boldsymbol{\beta} < 0$, which implies severe restrictions on $\boldsymbol{\beta}$. Therefore, often a more adequate link function is given by the log-link

$$g(\mu) = \log(\mu) \quad \text{or} \quad h(\eta) = \exp(\eta),$$

yielding $\mu = \exp(\eta) = \exp(\boldsymbol{x}^T\boldsymbol{\beta})$.

### 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 $\mu > 0$ denoting the expectation and $\nu > 0$ the shape parameter, the Gamma-distribution has the form

$$\begin{aligned} f(y) &= \frac{1}{\Gamma(\nu)} \left(\frac{\nu}{\mu}\right)^{\nu} y^{\nu-1} \exp\left(-\frac{\nu}{\mu}\, y\right) \\ &= \exp\left(\frac{-(1/\mu)y - \log(\mu)}{1/\nu} + \nu\log(\nu) + (\nu-1)\log(y) - \log(\Gamma(\nu))\right). \end{aligned}$$

In exponential family parameterization one obtains the dispersion parameter $\phi = 1/\nu$ and $\theta(\mu) = -1/\mu, b(\theta) = -\log(-\theta)$. 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 3.2, 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 $y = \mu - \mu/\nu$ and is positively skewed. Usually the $\Gamma$-distribution is abbreviated by $\Gamma(\nu, \alpha)$,

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

The variance of the Gamma-distribution is given by $\text{var}(y) = \nu/\alpha^2 = \mu^2/\nu$. 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. The coefficient of variation, given by $c = \sigma/\mu$, is 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 $\sigma_i/\mu_i = \mu_i/(\sqrt{\nu}\mu_i) = 1/\sqrt{\nu}$. Since it does not depend on the observation, one may set $c = \sigma_i/\mu_i$. 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.

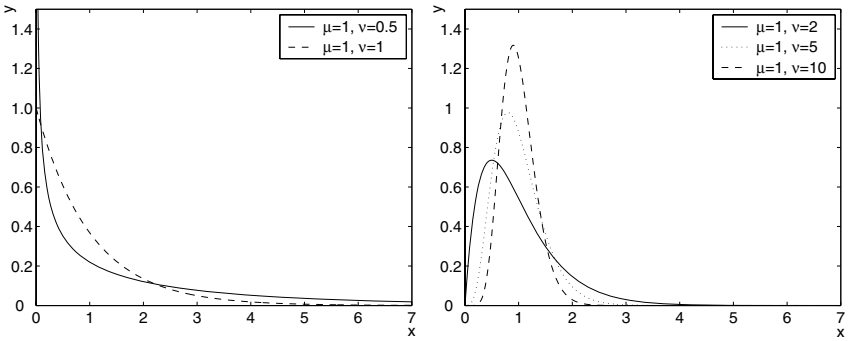


FIGURE 3.2: Gamma-distributions for several $\mu$,$\nu$.

The canonical link for the Gamma-disribution is the same as for the exponential distribution. Figure 3.3 shows the exponential and the Gamma regression model 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

$$f(y) = \left(\frac{\lambda}{2\pi y^3}\right)^{1/2} \exp\left\{-\frac{\lambda}{2\mu^2 y}(y-\mu)^2\right\}, \qquad y > 0,$$

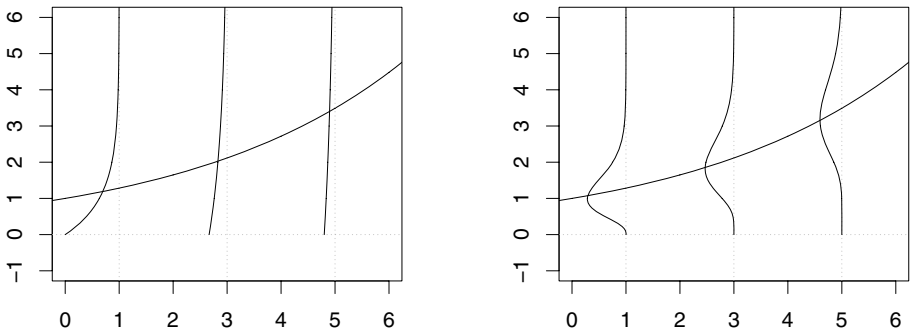


FIGURE 3.3: Exponential (left) and Gamma-distributed (right) regression model with log-link.

with abbreviation $IG(\mu, \lambda)$, where $\mu, \lambda > 0$ are the determining parameters. Straightforward derivation yields

$$f(y) = \exp\left\{\frac{y(-1/(2\mu^2)) + 1/\mu}{1/\lambda} - \frac{\lambda}{2y} - \frac{1}{2}\log(\lambda 2\pi) - \frac{3}{2}\log(y)\right\},$$

and therefore

$$\theta = -\frac{1}{2\mu^2}, \quad b(\theta) = -1/\mu = -\sqrt{-2\theta}, \quad \phi = 1/\lambda,$$
$$c(y, \phi) = -1/(2y\phi) - \frac{1}{2}\log(2\pi/\phi) - \frac{3}{2}\log(y).$$

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

$$g(\mu) = -\frac{1}{2\mu^2} \qquad \text{or} \qquad h(\eta) = -\frac{1}{\sqrt{2\eta}},$$

which implies the severe restriction $\eta = \boldsymbol{x}^T\boldsymbol{\beta} > 0$. A link function without these problems is the log-link function $g(\mu) = \log(\mu)$ and thus i$h(\eta) = \exp(\eta)$.

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

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} y_i, \quad \frac{1}{\hat{\lambda}} = \frac{1}{n}\sum(\frac{1}{y_i} - \frac{1}{\bar{y}}),$$

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 nested designs (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

$$f(y) = \pi^y(1-\pi)^{1-y} = \exp\left\{y\ \log\left(\frac{\pi}{1-\pi}\right) + \log(1-\pi)\right\},$$

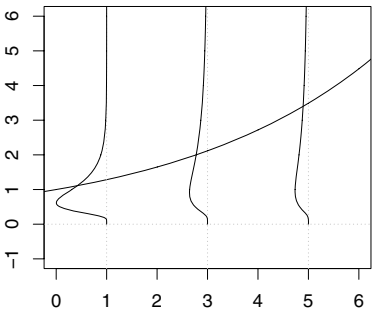


FIGURE 3.4: Inverse Gaussian-distributed model with $\lambda = 3$ and log-link.

where $\pi = P(y = 1)$ is the probability for "success." With $\mu = \pi$ it is an exponential family with $\theta(\pi) = \log(\pi/(1-\pi))$, $b(\theta) = \log(1 + \exp(\theta)) = -\log(1-\pi)$, $\phi = 1$.

The classical link that yields the logit model is

$$\pi = \frac{\exp(\eta)}{1 + \exp(\eta)}, \quad g(\pi) = \log\left(\frac{\pi}{1-\pi}\right)$$

(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 $\pi = F(\boldsymbol{x}_i^T\boldsymbol{\beta})$, with the response and link functions given by $h(\eta) = F(\eta)$, $g(\pi) = F^{-1}(\pi)$.

### 3.3.2 Models for Binomial Data

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

$$f(\tilde{y}) = \binom{m}{\tilde{y}}\pi^{\tilde{y}}(1-\pi)^{m-\tilde{y}} = \exp\left\{\frac{\frac{\tilde{y}}{m}\log\left(\frac{\pi}{1-\pi}\right) + \log(1-\pi)}{1/m} + \log\binom{m}{\tilde{y}}\right\}.$$

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

$$f(y) = \binom{m}{my}\pi^{my}(1-\pi)^{m-my} = \binom{m}{\tilde{y}}\pi^{\tilde{y}}(1-\pi)^{m-\tilde{y}},$$

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 $y_i = \tilde{y}_i/m_i$ 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

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

With expectation $\mu = \lambda$ the parameters of the exponential family are given by $\theta(\mu) = \log(\mu)$, $b(\theta) = \exp(\theta) = \mu, \phi = 1$. 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

$$\log(\lambda) = \boldsymbol{x}^T\boldsymbol{\beta} \quad \text{or} \quad \lambda = \exp(\boldsymbol{x}^T\boldsymbol{\beta}), \ \text{respectively.}$$

The distribution is shown for three distinct $x$-values in Figure 3.5. It is seen that differing means 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.

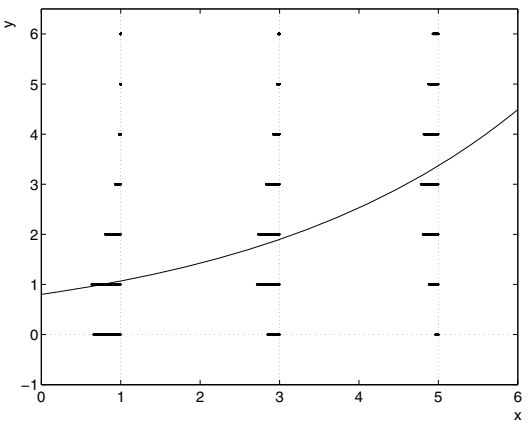


FIGURE 3.5: Poisson regression with log-link.

### 3.3.4 Negative Binomial Distribution

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

$$f(\tilde{y}) = \frac{\Gamma(\tilde{y}+\nu)}{\Gamma(\tilde{y}+1)\Gamma(\nu)} \left(\frac{\nu}{\tilde{\mu}+\nu}\right)^{\nu} \left(\frac{\tilde{\mu}}{\tilde{\mu}+\nu}\right)^{y}, \quad y = 0, 1, \ldots, \tag{3.5}$$

where $\nu, \tilde{\mu} > 0$ are parameters. We will use the abbreviation $\mathrm{NB}(\nu, \mu)$. 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 $\lambda \sim \Gamma(\nu, \frac{\nu}{\tilde{\mu}})$ with shape parameter $\nu$ and expectation $\tilde{\mu}$. Given $\lambda$, it is assumed that $\tilde{y}$ is Poisson-distributed, $\tilde{y}|\lambda \sim \mathrm{P}(\lambda)$. Then the marginal distribution of $\tilde{y}$ 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 $\tilde{\mu}^2/\nu$, 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

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

Thus, for $\nu \to \infty$, one obtains $\mathrm{E}(\tilde{y}) = \mathrm{var}(\tilde{y})$, 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

$$f(y) = \binom{\nu+y-1}{\nu-1} \quad \pi^{\nu}(1-\pi)^{y} \quad y = 0, 1, \ldots, \tag{3.6}$$

where $\pi = \nu/(\tilde{\mu}+\nu) \in (0,1)$ may be seen as an alternative parameter with a simple interpretation. If independent Bernoulli variables with probability of occurrence $\pi$ 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 $\pi = \nu/(\tilde{\mu}+\nu)$ from (3.5)

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

For fixed $\nu$ one has a simple exponential family for the scaled response $y = \tilde{y}/\nu$ and the dispersion $\phi = 1/\nu$. Since $\tilde{y}/\nu$ is considered as the response, one has expectation $\mu = \mathrm{E}(y) = \tilde{\mu}/\nu$ and therefore $\theta(\mu) = \log(1-\pi) = \log(\mu/(\mu+1))$ and $b(\theta) = -\log(1-\exp(\theta))$. 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 $\eta \to 0$, one has $\mu \to \infty$. For the log-link, $\log(\mu) = \eta \quad \text{or} \quad \mu = \exp(\eta)$; however, the predictor $\eta$ is not restricted.

The negative binomial response $y = \tilde{y}/\nu$ 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 $f(y_i|\theta_i, \phi_i) = \exp\{(y_i\theta_i - b(\theta_i))/\phi_i + c(y_i, \phi_i)\}$. In the previous sections examples have been given for the dependence of the natural parameters $\theta_i$ on $\mu_i$ and the parameters that characterize the distribution. For example, for the Bernoulli distribution one obtains $\theta_i = \theta(\mu_i)$ in the form $\theta_i = \log(\mu_i/(1-\mu_i))$, and since $\mu_i = \pi_i$, one has $\theta_i = \log(\pi_i/(1-\pi_i))$.

In general, in the exponential families the mean is directly related to the function $b(\theta_i)$ in the form

$$\mu_i = b'(\theta_i) = \partial b(\theta_i)/\partial\theta, \tag{3.7}$$

and for the variances one obtains

$$\sigma_i^2 = \text{var}(y_i) = \phi_i b''(\theta_i) = \phi_i \partial^2 b(\theta_i)/\partial\theta^2. \tag{3.8}$$

Thus the variances are composed from the dispersion parameter $\phi_i$ and the so-called *variance function* $b''(\theta_i)$. 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 $b(\theta)$. Because $\theta_i$ depends on the mean through the functional form $\theta_i = \theta(\mu_i)$, the variance function is a function of the mean, that is, $v(\mu_i) = \partial^2 b(\theta_i)/\partial\theta^2$, and the variance can be written as $\sigma_i^2 = \phi_i v(\mu_i)$. For the normal distribution one obtains $v(\mu_i) = 1$, and for the Poisson $v(\mu_i) = \mu_i$ (see Table 3.1).

The link between the mean and variance includes the dispersion parameter $\phi_i$. 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

$$\phi_i = \phi a_i,$$

where $a_i$ is known with $a_i = 1/m_i$ for the binomial distribution and $a_i = 1$ 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 $a_i \neq 1$ is the binomial distribution, which may be considered as replications of Bernoulli variables.

**Means and Variances**

$$\mu_i = b'(\theta_i)$$
$$\sigma_i^2 = \phi_i b''(\theta_i) = \phi_i v(\mu_i)$$

TABLE 3.1: Exponential family of distributions.

$$f(y_i|\theta_i,\phi_i) = \exp\left\{\frac{y_i\theta_i - b(\theta_i)}{\phi_i} + c(y_i,\phi_i)\right\},\ \phi_i = \phi a_i$$

(a) Components of the exponential family

| Distribution | Notation | $\mu_i$ | $\theta(\mu_i)$ | $b(\theta_i)$ | $\phi$ | $a_i$ |
|---|---|---|---|---|---|---|
| Normal | $\mathrm{N}(\mu_i, \sigma^2)$ | $\mu_i$ | $\mu_i$ | $\theta_i^2/2$ | $\sigma^2$ | 1 |
| Exponential | $\mathrm{E}(\lambda_i)$ | $1/\lambda_i$ | $-1/\mu_i$ | $-\log(-\theta_i)$ | 1 | 1 |
| Gamma | $\Gamma(\nu, \frac{\nu}{\mu_i})$ | $\mu_i$ | $-1/\mu_i$ | $-\log(-\theta_i)$ | $\frac{1}{\nu}$ | 1 |
| Inverse Gaussian | $\mathrm{IG}(\mu_i, \lambda)$ | | $-1/(2\mu_i^2)$ | $-(-2\theta_i)^{1/2}$ | $1/\lambda$ | 1 |
| Bernoulli | $\mathrm{B}(1, \pi_i)$ | $\pi_i$ | $\log(\frac{\mu_i}{1-\mu_i})$ | $\log(1+\exp(\theta_i))$ | 1 | 1 |
| Binomial (rescaled) | $\mathrm{B}(m_i, \pi_i)/m_i$ | $\pi_i$ | $\log(\frac{\mu_i}{1-\mu_i})$ | $\log(1+\exp(\theta_i))$ | 1 | $\frac{1}{m_i}$ |
| Poisson | $\mathrm{P}(\lambda_i)$ | $\lambda_i$ | $\log(\mu_i)$ | $\exp(\theta_i)$ | 1 | 1 |
| Negative binomial (rescaled) | $\mathrm{NB}(\nu, \frac{\nu(1-\pi_i)}{\pi_i})/\nu$ | $\frac{\nu(1-\pi_i)}{\pi_i}/\nu$ | $\log(\frac{\mu_i}{\mu_i+1})$ | $-\log(1-e^{\theta})$ | $\frac{1}{\nu}$ | 1 |

(b) Expectation and variance

| Distribution | $\mu_i = b'(\theta_i)$ | var. fct. $b''(\theta_i)$ | variance $\phi_i\ b''(\theta_i)$ |
|---|---|---|---|
| Normal | $\mu_i = \theta_i$ | 1 | $\sigma^2$ |
| Exponential | $\mu_i = -\frac{1}{\theta_i}$ | $\mu_i^2$ | $\mu_i^2$ |
| Gamma | $\mu_i = -\frac{1}{\theta_i}$ | $\mu_i^2$ | $\frac{\mu_i^2}{\nu}$ |
| Inverse Gaussian | $\mu_i = (-2\theta)^{-1/2}$ | $\mu_i^3$ | $\mu_i^3/\lambda$ |
| Bernoulli | $\mu_i = \frac{\exp(\theta_i)}{1+\exp(\theta_i)}$ | $\pi_i(1-\pi_i)$ | $\pi_i(1-\pi_i)$ |
| Binomial | $\mu_i = \frac{\exp(\theta_i)}{1+\exp(\theta_i)}$ | $\pi_i(1-\pi_i)$ | $\frac{1}{m_i}\pi_i(1-\pi_i)$ |
| Poisson | $\lambda_i = \exp(\theta_i)$ | $\lambda_i$ | $\lambda_i$ |
| Negative binomial | $\mu_i = \frac{\exp(\theta_i)}{1-\exp(\theta_i)}$ | $\mu(1+\mu)$ | $\frac{\mu(1+\mu)}{\nu}$ |

### 3.4.2 Canonical Link

The choice of the link function depends on the distribution of the response. For example, if $y$ is non-negative, a link function is appropriate that specifies non-negative means without restricting the parameters. For each distribution within the simple exponential family there is one link function that has some technical advantages, the so-called canonical link. It links the linear predictor directly to the canonical parameter in the form

$$\theta_i = \boldsymbol{x}_i^T\boldsymbol{\beta}.$$

Since $\theta_i$ is determined as a function $\theta(\mu_i)$, the canonical link $g$ may be derived from the general form $g(\mu_i) = \boldsymbol{x}_i^T\boldsymbol{\beta}$ as the transformation that transforms $\mu_i$ to $\theta_i$. In Table 3.1 the canonical

links are given, and one obtains, for example,

$$\begin{aligned} g(\mu) &= \mu && \text{for the normal distribution,} \\ g(\mu) &= \log(\pi/(1-\pi)) && \text{for the Bernoulli distribution,} \\ g(\mu) &= -1/\mu && \text{for the Gamma-distribution.} \end{aligned}$$

The last example shows that the canonical link might not always be the best choice because $-1/\mu_i = \boldsymbol{x}_i^T\boldsymbol{\beta}$ or $\mu_i = -1/\boldsymbol{x}_i^T\boldsymbol{\beta}$ implies severe restrictions on $\boldsymbol{\beta}$ arising from the restriction that $\mu_i$ 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 $i$ and therefore may vary across observations. With $\Delta_i$ denoting the underlying time interval for observation $y_i$, one may assume Poisson-distributed responses $y_i \sim \mathrm{P}(\Delta_i\lambda_i)$, 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 $\Delta_i\lambda_i$, but the intensity $\lambda_i$ in dependence on covariates and include the time intervals as known constants. The model

$$\lambda_i = \exp(\boldsymbol{x}_i^T\boldsymbol{\beta})$$

yields for the expectation $\mu_i = \mathrm{E}(y_i)$

$$\mu_i = \Delta_i\lambda_i = \exp(\log(\Delta_i) + \boldsymbol{x}_i^T\boldsymbol{\beta}).$$

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 $\log(\Delta_i)$ is omitted because it cannot be distinguished from the intercept within $\boldsymbol{x}_i^T\boldsymbol{\beta}$. 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 $(y_i, \boldsymbol{x}_i)$, $i = 1, \ldots, n$. Often, for example, if covariates are categorial or in experimental studies, several of the covariate values $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ 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

$$(y_{ij}, \boldsymbol{x}_i), \quad i = 1, \ldots, N, \quad j = 1, \ldots, n_i,$$

where observations $y_{i1}, \ldots, y_{in_i}$ have a fixed covariate vector $\boldsymbol{x}_i$ with $n_i$ denoting the sample size at covariate value $\boldsymbol{x}_i$, yielding the total sum of observations $n = n_1 + \ldots + n_N$. Since means depend on covariates, one has

$$\mu_i = \mu_{ij} = \mathrm{E}(y_{ij}) = h(\boldsymbol{x}_i^T\boldsymbol{\beta}), \quad j = 1, \ldots, n_i,$$

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

$$f(\bar{y}_i) = \exp\left\{\frac{(\bar{y}_i\theta_i - b(\theta_i))}{\phi/n_i} + \bar{c}(y_{i1}, \ldots, y_{in_i}, \phi)\right\}, \tag{3.9}$$

where $\bar{c}(\ldots)$ is a modified normalizing function (Exercise 3.3). Thus, the mean has an exponential family distribution with the same natural parameter $\theta_i$ and function $b(\theta_i)$ as for ungrouped data. However, the dispersion parameter has changed. With the value

$$\phi_i = \phi/n_i$$

it reflects the reduced dispersion due to considering the mean $\bar{y}_i$ across $n_i$ observations. For grouped data one has

$$\mu_i = \mathrm{E}(\bar{y}_i) = \mathrm{E}(y_{ij}) = b'(\theta_i)$$
$$\sigma_i^2 = \mathrm{var}(\bar{y}_i) = \frac{1}{n_i}\mathrm{var}(y_{ij}) = \frac{\phi}{n_i}b''(\theta_i).$$

The variance function $v(\mu_i) = b''(\theta(\mu_i))$ is the same as for ungrouped observations, but the variance is $\mathrm{var}(\bar{y}_i) = (\phi/n_i)v(\mu_i)$.

Consequently, one obtains for grouped data, considering $\bar{y}_1, \ldots, \bar{y}_N$ as responses, a GLM with the dispersion given by $\phi_i = \phi/n_i$. This is exactly what happens in the transition from the Bernoulli response to the scaled binomial response. The binomial response may be seen as grouped data from Bernoulli responses with local sample size given by $n_i$.

In the grouped case we will use $N$ as the number of grouped observations with local sample sizes $n_1, \ldots, n_N$. For ungrouped data the number of observations is $n$. In both cases the dispersion parameter is denoted by $\phi_i$.

In general, the mean $y_i$ over replications $y_{ij}, \ldots, y_{in_i}$ does not have the same type of distribution as the original observations $y_{ij}$. For example, in Poisson or binomially distributed responses the mean is not integer-valued and thus is not Poisson or binomially distributed, respectively. However, for these distributions the sum $\tilde{y}_i = \sum_{j=1}^{n_i} y_{ij}$ has the same type of distribution as the replications.

## 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 $f(y_i|\theta_i, \phi_i) = \exp\{(y_i\theta_i - b(\theta_i))/\phi_i + c(y_i, \phi_i)\}$ is that the mean and variance are given by

$$\mathrm{E}(y_i) = \partial b(\theta_i)/\partial\theta, \quad \mathrm{var}(y_i) = \phi_i \partial^2(\theta_i)/\partial\theta^2,$$

where the parameterization is in the canonical parameter $\theta_i$. 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 $y_1, \ldots, y_n$ the log-likelihood

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{n} l_i(\theta_i) = \sum_{i=1}^{n}(y_i\theta_i - b(\theta_i))/\phi_i,$$

where the term $c(y_i, \phi_i)$ is omitted because it does not depend on $\theta_i$ and therefore not on $\boldsymbol{\beta}$. For the maximization of the log-likelihood one computes the derivation $\boldsymbol{s}(\boldsymbol{\beta}) = \partial l(\boldsymbol{\beta})/\partial\boldsymbol{\beta}$,

which is called the score function. For the computation it is useful to consider the parameters as resulting from transformations in the form $\theta_i = \theta(\mu_i), \mu_i = h(\eta_i), \eta_i = \boldsymbol{x}_i^T\boldsymbol{\beta}$. One has the transformation structure

$$\eta_i \underset{g=h^{-1}}{\overset{h}{\rightleftarrows}} \mu_i \underset{\mu=\theta^{-1}}{\overset{\theta}{\rightleftarrows}} \theta_i$$

yielding $\theta_i = \theta(\mu_i) = \theta(h(\eta_i))$. Then the score function $\boldsymbol{s}(\boldsymbol{\beta}) = \partial l(\boldsymbol{\beta})/\partial\boldsymbol{\beta}$ is given by

$$\boldsymbol{s}(\boldsymbol{\beta}) = \frac{\partial l(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} \frac{\partial l_i(\theta_i)}{\partial \theta} \frac{\partial \theta(\mu_i)}{\partial \mu} \frac{\partial h(\eta_i)}{\partial \eta} \frac{\partial \eta_i}{\partial \boldsymbol{\beta}}.$$

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

$$\begin{aligned}
\frac{\partial l_i}{\partial \theta} &= (y_i - b'(\theta_i))/\phi_i = (y_i - \mu_i)/\phi_i, \\
\frac{\partial \theta(\mu_i)}{\partial \mu} &= \left(\frac{\partial \mu(\theta_i)}{\partial \theta}\right)^{-1} = \left(\frac{\partial^2 b(\theta_i)}{\partial \theta^2}\right)^{-1} = \phi_i / \operatorname{var}(y_i), \\
\frac{\partial \eta_i}{\partial \boldsymbol{\beta}} &= \boldsymbol{x}_i,
\end{aligned}$$

and therefore the score function

$$\boldsymbol{s}(\boldsymbol{\beta}) = \sum_{i=1}^{n} s_i(\boldsymbol{\beta}) = \sum_{i=1}^{n} \boldsymbol{x}_i \frac{\partial h(\eta_i)}{\partial \eta} \frac{(y_i - \mu_i)}{\operatorname{var}(y_i)}.$$

With $\sigma_i^2 = \phi_i v(\mu_i) = \operatorname{var}(y_i)$, the estimation equation $\boldsymbol{s}(\hat{\boldsymbol{\beta}}) = \boldsymbol{0}$ has the form

$$\sum_{i=1}^{n} \boldsymbol{x}_i \frac{\partial h(\eta_i)}{\partial \eta} \frac{(y_i - \mu_i)}{\phi_i v(\mu_i)} = \boldsymbol{0}. \tag{3.10}$$

In (3.10) the response (or link) function is found in the specification of the mean $\mu_i = h(\boldsymbol{x}_i^T\boldsymbol{\beta})$ and in the derivative $\partial h(\eta_i)/\partial\eta$, whereas from higher moments of the distribution of $y_i$ only the variance $\sigma_i^2 = \phi_i v(\mu_i)$ is needed. Since $\phi_i = \phi a_i$, the dispersion parameter $\phi$ may be canceled out and the estimate $\hat{\boldsymbol{\beta}}$ does not depend on $\phi$.

For the canonical link the estimation equation simplifies. Since $\theta_i = \eta_i = \boldsymbol{x}_i^T\boldsymbol{\beta}$, the score function reduces to $\boldsymbol{s}(\boldsymbol{\beta}) = \sum_i (\partial l_i/\partial\theta)(\partial\eta_i/\partial\boldsymbol{\beta})$ and one obtains

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

In particular, one has $\partial h(\eta_i)/\partial\eta = \operatorname{var}(y_i)/\phi_i$ if the canonical link is used.

In matrix notation the score function is given by

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

where $\boldsymbol{X}^T = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ is the design matrix, $\boldsymbol{D} = \text{ Diag } (\partial h(\eta_1)/\partial\eta, \ldots, \partial h(\eta_n)/\partial\eta)$ is the diagonal matrix of derivatives, $\boldsymbol{\Sigma} = \text{ Diag } (\sigma_1^2, \ldots, \sigma_n^2)$ is the covariance matrix, and $\boldsymbol{y}^T = (y_1, \ldots, y_n)$, $\boldsymbol{\mu}^T = (\mu_1, \ldots, \mu_n)$ are the vectors of observations and means. Sometimes it is useful to combine $\boldsymbol{D}$ and $\boldsymbol{\Sigma}$ into the weight matrix $\boldsymbol{W} = \boldsymbol{D}\ \boldsymbol{\Sigma}^{-1}\boldsymbol{D}^T$, yielding $s(\boldsymbol{\beta}) = \boldsymbol{X}^T\boldsymbol{W}\boldsymbol{D}^{-1}(\boldsymbol{y} - \boldsymbol{\mu})$ and $\boldsymbol{F}(\boldsymbol{\beta}) = \boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X}$.

## Information Matrix

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

$$\boldsymbol{F}_{obs}(\boldsymbol{\beta}) = -\frac{\partial^2 l(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} = \left( -\frac{\partial^2 l(\boldsymbol{\beta})}{\partial \beta_i \partial \beta_j} \right)_{i,j}.$$

Its 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

$$\boldsymbol{F}(\boldsymbol{\beta}) = \mathrm{E}(\boldsymbol{F}_{obs}(\boldsymbol{\beta})).$$

For the derivation it is essential that $\mathrm{E}(\boldsymbol{s}(\boldsymbol{\beta})) = \mathbf{0}$ and that $\mathrm{E}(-\partial^2 l_i / \partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T) = \mathrm{E}((\partial l_i / \partial \boldsymbol{\beta})(\partial l_i / \partial \boldsymbol{\beta}^T))$, which holds under general assumptions (see, for example, Cox and Hinkley, 1974). Thus one obtains

$$\begin{aligned} \boldsymbol{F}(\boldsymbol{\beta}) &= \mathrm{E}\left( \sum_{i=1}^{n} \boldsymbol{s}_i(\boldsymbol{\beta}) \boldsymbol{s}_i(\boldsymbol{\beta})^T \right) = \mathrm{E}\left( \sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i^T \left( \frac{\partial h(\eta_i)}{\partial \eta} \right)^2 \frac{(y_i - \mu_i)^2}{\mathrm{var}(y_i)^2} \right) \\ &= \sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i^T \left( \frac{\partial h(\eta_i)}{\partial \eta} \right)^2 / \sigma_i^2, \end{aligned}$$

where $\sigma_i^2 = \mathrm{var}(y_i)$. By using the design matrix $\boldsymbol{X}$ one obtains the information matrix $\boldsymbol{F}(\boldsymbol{\beta})$ in the form

$$\boldsymbol{F}(\boldsymbol{\beta}) = \boldsymbol{X}^T \boldsymbol{W} \boldsymbol{X},$$

where $\boldsymbol{W} = \mathrm{Diag}\left( \left( \frac{\partial h(\eta_1)}{\eta} \right)^2 / \sigma_1^2, \ldots, \left( \frac{\partial h(\eta_n)}{\partial \eta} \right)^2 / \sigma_n^2 \right)$ is a diagonal weight matrix that has the matrix form $\boldsymbol{W} = \boldsymbol{D}\, \boldsymbol{\Sigma}^{-1} \boldsymbol{D}^T$.

For the canonical link the corresponding simpler form is

$$\boldsymbol{F}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i^T \sigma_i^2 / \phi_i^2 = \boldsymbol{X}^T \boldsymbol{W} \boldsymbol{X},$$

with weight matrix $\boldsymbol{W} = (\sigma_1^2/\phi_1^2, \ldots, \sigma_n^2/\phi_n^2)$ and the observed information is identical to the information matrix, $\boldsymbol{F}_{obs}(\boldsymbol{\beta}) = \boldsymbol{F}(\boldsymbol{\beta})$. It is immediately seen that for the normal distribution model with a (canonical) identity link one has with $\phi_i = \sigma_i^2 = \sigma^2$ the familiar form

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

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

$$\hat{\mathrm{cov}}(\hat{\boldsymbol{\beta}}) \approx (\boldsymbol{X}' \hat{\boldsymbol{W}} \boldsymbol{X})^{-1},$$

where $\hat{\boldsymbol{W}}$ means the evaluation of $\boldsymbol{W}$ at $\hat{\boldsymbol{\beta}}$, that is, $\partial h(\eta_i)/\partial \eta$ is replaced by $\partial h(\hat{\eta}_i)/\partial \eta, \hat{\eta}_i = \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}}$, and $\sigma_i^2 = \phi_i v(\hat{\mu}_i), \hat{\mu}_i = h(\hat{\eta}_i)$.

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 $N$.

Log-Likelihood

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{n}(y_i\theta_i - b(\theta_i))/\phi_i$$

Score Function

$$\begin{aligned}\boldsymbol{s}(\boldsymbol{\beta}) = \frac{\partial l(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} &= \sum_{i=1}^{n} \boldsymbol{x}_i \frac{\partial h(\eta_i)}{\partial \eta}\frac{(y_i-\mu_i)}{\sigma_i^2} \\ &= \boldsymbol{X}^T\boldsymbol{D}\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{\mu}) = \boldsymbol{X}^T\boldsymbol{W}\boldsymbol{D}^{-1}(\boldsymbol{y}-\boldsymbol{\mu})\end{aligned}$$

Information Matrix

$$\begin{aligned}\boldsymbol{F}(\boldsymbol{\beta}) = \mathrm{E}\left(-\frac{\partial^2 l(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T}\right) &= \sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i^T\left(\frac{\partial h(\eta_i)}{\partial\eta}\right)^2/\sigma_i^2 \\ &= \boldsymbol{X}^T\boldsymbol{D}\boldsymbol{\Sigma}^{-1}\boldsymbol{D}\boldsymbol{X} = \boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X}\end{aligned}$$

If $\phi$ is an unknown (normal, Gamma-distribution), the moments estimate is

$$\hat{\phi} = \frac{1}{n-p}\sum_{i=1}^{n}\frac{(y_i-\hat{\mu}_i)^2}{v(\hat{\mu}_i)}.$$

For the normal model $\hat{\phi}$ reduces to the usual unbiased and consistent estimates $\hat{\phi} = \hat{\sigma}^2 = \sum_i (y_i-\hat{\mu}_i)^2/(n-p)$. For the Gamma-distribution one obtains

$$\hat{\phi} = \frac{1}{\hat{\nu}} = \frac{1}{n-p}\sum_{i=1}^{n}\left(\frac{y_i-\hat{\mu}_i}{\hat{\mu}_i}\right)^2.$$

In the approximation $\phi$ has to be replaced by $\hat{\phi}$ when computing $\boldsymbol{F}(\hat{\boldsymbol{\beta}})$. The likelihood, score function, and Fisher matrix are summarized in a box. In the matrix form derivations are collected in the matrix $\boldsymbol{D} = \text{ Diag }(\partial h(\eta_1)/\partial\eta, \ldots, \partial h(\eta_n)/\partial\eta)$. By using $\boldsymbol{W}$ and $\boldsymbol{D}$, the dependence on $\boldsymbol{\beta}$ is suppressed, and actually one has $\boldsymbol{W} = \boldsymbol{W}(\boldsymbol{\beta}), \boldsymbol{D} = \boldsymbol{D}(\boldsymbol{\beta})$.

Under regularity conditions the ML estimate $\hat{\boldsymbol{\beta}}$ exists and is unique asymptotically $(n \to \infty)$. It is consistent and the distribution may be approximated by a normal distribution with the covariance given by the inverse Fisher matrix. More precisely, under assumptions that ensure the convergence of $\boldsymbol{F}(\hat{\boldsymbol{\beta}})/n$ to a limit $\boldsymbol{F}_0(\hat{\boldsymbol{\beta}})$, one obtains for $\sqrt{n}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})$ asymptotically a normal distribution $\mathrm{N}(\mathbf{0}, \boldsymbol{F}_0(\hat{\boldsymbol{\beta}})^{-1})$. For finite $n$ one uses the approximation $\mathrm{cov}(\hat{\boldsymbol{\beta}}) \approx \boldsymbol{F}_0(\hat{\boldsymbol{\beta}})^{-1}/n$, which is approximated by $\boldsymbol{F}(\hat{\boldsymbol{\beta}})^{-1}$. For regularity conditions see Haberman (1977) and Fahrmeir and Kaufmann (1985). Bias correction by approximation of the first-order bias of ML estimates was investigated by Cordeiro and McCullagh (1991) and Firth (1993).

Approximation

$$\hat{\boldsymbol{\beta}} \overset{a}{\sim} \mathrm{N}\left(\boldsymbol{\beta}, F(\hat{\boldsymbol{\beta}})^{-1}\right),$$
$$\boldsymbol{F}(\hat{\boldsymbol{\beta}}) = \boldsymbol{X}^T\hat{\boldsymbol{W}}\boldsymbol{X} = \textstyle\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i^T\left(\frac{\partial h(\hat{\eta}_i)}{\partial\eta}\right)^2/(\hat{\phi}v(\hat{\mu}_i))$$

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 $\boldsymbol{D}$, which contains $\partial h(\eta_i)/\partial\eta$;
- the distribution, yielding the covariance matrix $\boldsymbol{\Sigma}$, which contains $\sigma_i^2 = \text{var}(y_i)$;
- the explanatory variables, which determine the design matrix $\boldsymbol{X}$.

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 $l(\boldsymbol{y}; \hat{\boldsymbol{\mu}}, \phi)$ denote the maximum of the log-likelihood of the model where $\boldsymbol{y}^T = (y_1, \ldots, y_n)$ represents the data, $\hat{\boldsymbol{\mu}}^T = (\hat{\mu}_1, \ldots, \hat{\mu}_n)$, $\hat{\mu}_i = h(\boldsymbol{x}_i^T\hat{\boldsymbol{\beta}})$ represent the fitted values based on the ML estimate $\hat{\boldsymbol{\beta}}$; and the dispersion of observations has the form $\phi_i = \phi a_i$ with known $a_i$. For the saturated model that matches the data exactly one has $\hat{\boldsymbol{\mu}} = \boldsymbol{y}$ and the log-likelihood is given by $l(\boldsymbol{y}; \boldsymbol{y}, \phi)$. With $\theta(\hat{\mu}_i), \theta(y_i)$ denoting the canonical parameters of the GLM under investigation and the saturated model, respectively, the *deviance* is given by

$$\begin{aligned} D(\boldsymbol{y}, \hat{\boldsymbol{\mu}}) &= -\phi 2\{l(\boldsymbol{y}; \hat{\boldsymbol{\mu}}, \phi) - l(\boldsymbol{y}; \boldsymbol{y}, \phi)\} \\ &= 2\sum_{i=1}^{n}\{y_i(\theta(y_i) - \theta(\hat{\mu}_i)) - (b(\theta(y_i)) - b(\theta(\hat{\mu}_i)))\}/a_i. \end{aligned}$$

$D(\boldsymbol{y}, \hat{\boldsymbol{\mu}})$ is known as deviance of the model under consideration while $D^+(\boldsymbol{y}, \hat{\boldsymbol{\mu}}) = D(\boldsymbol{y}, \hat{\boldsymbol{\mu}})/\phi$ is the so-called *scaled deviance*. The deviance is linked to the likelihood ratio statistic

TABLE 3.2: Deviances for several distributions.

| Distribution | Deviance |
|---|---|
| Normal | $D(\boldsymbol{y}, \hat{\boldsymbol{\mu}}) = \sum_{i=1}^{n} (y_i - \hat{\mu}_i)^2$ |
| Gamma | $D(\boldsymbol{y}, \hat{\boldsymbol{\mu}}) = 2 \sum_{i=1}^{n} -\log\left(\frac{y_i}{\hat{\mu}_i}\right) + \left[\frac{(y_i - \hat{\mu}_i)}{\hat{\mu}_i}\right]$ |
| Inverse Gaussian | $D(\boldsymbol{y}, \hat{\boldsymbol{\mu}}) = \sum_{i=1}^{n} (y_i - \hat{\mu}_i)^2 / (\hat{\mu}_i^2 y_i)$ |
| Bernoulli | $D(\boldsymbol{y}, \hat{\boldsymbol{\mu}}) = 2 \sum_{i=1}^{n} y_i \log\left(\frac{y_i}{\hat{\mu}_i}\right) + (1 - y_i) \log\left(\frac{1 - y_i}{1 - \hat{\mu}_i}\right)$ |
| Poisson | $D(\boldsymbol{y}, \hat{\boldsymbol{\mu}}) = 2 \sum_{i=i}^{n} y_i \log\left(\frac{y_i}{\hat{\mu}_i}\right) - [(y_i - \hat{\mu}_i)]$ |

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

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 $\mathrm{SSE}/\sigma^2$. For the Bernoulli distribution, one has $\theta(\mu_i) = \log(\mu_i/(1 - \mu_i))$ and one obtains $D(\boldsymbol{y}, \hat{\boldsymbol{\mu}}) = 2 \sum_i d(y_i, \hat{\pi}_i)$, where $d(y_i, \hat{\pi}_i) = -\log(1 - |y_i - \hat{\pi}_i|)$ (for more details see Section 4.2). In the cases of the Poisson and the Gamma deviances, the last term given in brackets $[\ldots]$ can be 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 difference of deviances 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 $D(\boldsymbol{y}, \hat{\boldsymbol{\mu}}) = (n - p)\hat{\sigma}^2 \sim \sigma^2 \chi^2(n - p)$. 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.

This is different if one considers the deviance of the binomial distribution or the Poisson distribution. The binomial distribution may be seen as a replication version of the Bernoulli distribution. Thus, when the number of replications increases, $n_i \to \infty$, the proportion $y_i$ is asymptotically normally distributed. For the Poisson distribution, asymptotic normality of the observations follows if $\mu_i \to \infty$ for each observation. In these cases the $\chi^2$-distribution may be used as an approximation (see Section 3.8).

### 3.7.2 Analysis of Deviance and the Testing of Hypotheses

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

#### Analysis of Deviance

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

$$D(M) = -\phi 2\{l(\boldsymbol{y}, \hat{\boldsymbol{\mu}}; \phi) - l(\boldsymbol{y}, \boldsymbol{y}; \phi)\},$$
$$D(\tilde{M}) = -\phi 2\{l(\boldsymbol{y}, \tilde{\boldsymbol{\mu}}; \phi) - l(\boldsymbol{y}, \boldsymbol{y}; \phi)\}.$$

The difference of deviances

$$D(\tilde{M}|M) = D(\tilde{M}) - D(M) = -2\phi\{l(\boldsymbol{y}, \tilde{\boldsymbol{\mu}}; \phi) - l(\boldsymbol{y}, \hat{\boldsymbol{\mu}}; \phi)\} \tag{3.11}$$

compares the fits of models $\tilde{M}$ and $M$. The difference of scaled deviances $D(\tilde{M}|M)/\phi$ is eqiuivalent 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 $\tilde{M}$ into

$$D(\tilde{M}) = D(\tilde{M}|M) + D(M).$$

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

$$\text{SSE}(\tilde{M}) = \text{SSE}(\tilde{M}|M) + \text{SSE}(M)$$

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

$$\frac{\{\text{SSE}(\tilde{M}) - \text{SSE}(M)\}/s}{\hat{\sigma}^2} \sim \text{F}(s, n-p),$$

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

$$\frac{D(\tilde{M}) - D(M)}{\phi} = \frac{D(\tilde{M}|M)}{\phi},$$

which under mild restrictions is asymptotically $\chi^2(s)$-distributed. This means that the difference

$$D(\tilde{M}) - D(M) = D(\tilde{M}|M)$$

has an asymptotically a $\phi\chi^2(s)$-distribution.

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

| | df | cond. deviance | df |
|---|---|---|---|
| $D(\tilde{M})$ | $n-p+s$ | | |
| $D(M)$ | $n-p$ | $D(\tilde{M}\|M)$ | $s$ |

It should be noted that only the difference of deviances $D(\tilde{M}|M)$ has asymptotically a $\phi\chi^2(s)$-distribution. The degrees of freedom of $D(\tilde{M})$ have the basic structure "number of observations minus number of fitted parameters." In $\tilde{M}$, 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 $D(\tilde{M})$ and $D(M)$ themselves are asymptotically distributed with $D(\tilde{M}) \sim \chi^2(N-p+s)$, and $\chi^2(N-p)$, where $N$ denotes the number of grouped observations (see Section 3.8). While $D(\tilde{M})$ and $D(M)$ are different for grouped and ungrouped data, the difference $D(\tilde{M}) - D(M)$ 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

$$\begin{array}{ccccc} \mathrm{SSE}(\tilde{M}) & = & \mathrm{SSE}(\tilde{M}|M) & + & \mathrm{SSE}(M) \\ \sigma^2\chi^2(n-p+s) & & \sigma^2\chi^2(s) & & \sigma^2\chi^2(n-p) \\ \text{if } \tilde{M} \text{ holds} & & \text{if } \tilde{M} \text{ holds} & & \text{if } M \text{ holds} \end{array}$$

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

$$\begin{array}{ccccc} D(\tilde{M}) & = & D(\tilde{M}|M) & + & D(M) \\ \phi\chi^2(N-p+s) & & \phi\chi^2(s) & & \phi\chi^2(N-p) \\ \text{if } \tilde{M} \text{ holds} & & \text{if } \tilde{M} \text{ holds} & & \text{if } M \text{ holds} \\ \text{grouped data} & & & & \text{grouped data} \end{array}$$

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 $(D(M_i) - D(M_{i+1}))/\phi$. The deviance of the most restrictive model is given as sum of these differences:

$$\begin{aligned} D(M_1) = & \quad (D(M_1) - D(M_2)) + (D(M_2) - D(M_3)) \\ & + \ldots + (D(M_{m-1}) - D(M_m)) + D(M_m) \\ = & \quad D(M_1|M_2) + \ldots + D(M_{m-1}|M_m) + D(M_m). \end{aligned}$$

Thus the discrepancy of the model $M_1$ is the sum of the "conditional" deviances $D(M_i|M_{i+1}) = D(M_i) - D(M_{i+1})$ and the discrepancy between the most general model $M_m$ and the saturated model. However, when one starts from a model $M_m$ and considers sequences of simpler models, one should be aware that different sequences of submodels are possible (see Section 4.4.2).

### 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

$$H_0 : \boldsymbol{C\beta} = \boldsymbol{\xi} \quad \text{against} \quad H_1 : \boldsymbol{C\beta} \neq \boldsymbol{\xi},$$

where $\text{rank}(\boldsymbol{C}) = s$. Alternative test statistics that can be used are the Wald test and the score statistic.

#### Wald Test

The *Wald statistic* has the form

$$w = \big(\boldsymbol{C}\hat{\boldsymbol{\beta}} - \boldsymbol{\xi}\big)^T \left[\boldsymbol{C}\, \boldsymbol{F}^{-1}(\hat{\boldsymbol{\beta}})\, \boldsymbol{C}^T\right]^{-1} \big(\boldsymbol{C}\hat{\boldsymbol{\beta}} - \boldsymbol{\xi}\big).$$

It uses the weighted distance between the unrestricted estimate $\boldsymbol{C}\hat{\boldsymbol{\beta}}$ of $\boldsymbol{C\beta}$ and its hypothetical value $\boldsymbol{\xi}$ under $H_0$. The weight is derived from the distribution of the difference $(\boldsymbol{C}\hat{\boldsymbol{\beta}} - \boldsymbol{\xi})$, for which one obtains asymptotically $\text{cov}(\boldsymbol{C}\hat{\boldsymbol{\beta}} - \boldsymbol{\xi}) = \boldsymbol{C}\, \boldsymbol{F}^{-1}(\hat{\boldsymbol{\beta}})\, \boldsymbol{C}^T$. Therefore, $w$ is the squared length of the standardized estimate $(\boldsymbol{C}\, \boldsymbol{F}^{-1}(\hat{\boldsymbol{\beta}})\, \boldsymbol{C}^T)^{-1/2}(\boldsymbol{C}\hat{\boldsymbol{\beta}} - \boldsymbol{\xi})$, 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 program 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 $\boldsymbol{C}$ is $\boldsymbol{C} = (0, 0, \ldots, 1, \ldots, 0)$. Then the Wald statistic has the simple form

$$w = \frac{\beta_j^2}{\hat{a}_{jj}},$$

where $\hat{a}_{jj}$ is the $j$th diagonal element of the estimated inverse Fisher matrix $\boldsymbol{F}^{-1}$. Since $w$ is asymptotically $\chi^2(1)$-distributed, one may also consider the square root,

$$z = \sqrt{w} = \frac{\beta_j}{\sqrt{\hat{a}_{jj}}},$$

which follows asymptotically a standard normal distribution. Thus, for single parameters, program packages usually give the standard error $\sqrt{\hat{a}_{jj}}$ and the $p$-value based on $z$.

#### Score Statistic

The score statistic is based on the following consideration: The score function $\boldsymbol{s}(\boldsymbol{\beta})$ for the unrestricted model is the zero vector if it is evaluated at the unrestricted ML estimate $\hat{\boldsymbol{\beta}}$. If, however, $\hat{\boldsymbol{\beta}}$ is replaced by the MLE $\tilde{\boldsymbol{\beta}}$ under $H_0$, $\boldsymbol{s}(\tilde{\boldsymbol{\beta}})$ 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*,

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

which is the squared weighted score function evaluated at $\tilde{\boldsymbol{\beta}}$.

An advantage of the Wald and score statistics is that they are properly defined for models with overdispersion since only the first and second moments are involved. All test statistics have the same asymptotic distribution. If they are differing strongly, that may be seen as a hint that the conditions for asymptotic results may not hold. A survey on asymptotics for test statistics was given by Fahrmeir (1987).

**Test Statistics for Linear Hypotheses**

$$H_0 : \boldsymbol{C\beta} = \boldsymbol{\xi} \qquad H_1 : \boldsymbol{C\beta} \neq \boldsymbol{\xi}$$
$$\text{with} \quad \operatorname{rank}(\boldsymbol{C}) = s$$

**Likelihood Ratio Statistic**

$$\begin{aligned}\lambda &= -2\{l(\boldsymbol{y}; \tilde{\boldsymbol{\mu}}, \hat{\phi}) - l(\boldsymbol{y}, \hat{\boldsymbol{\mu}}, \hat{\phi})\} \\ &= (D(\boldsymbol{y}; \tilde{\boldsymbol{\mu}}, \hat{\phi}) - D(\boldsymbol{y}, \hat{\boldsymbol{\mu}}, \hat{\phi}))/\hat{\phi}\end{aligned}$$

**Wald Statistic**

$$w = (\boldsymbol{C}\hat{\boldsymbol{\beta}} - \boldsymbol{\xi})^T (\boldsymbol{C}\boldsymbol{F}^{-1}(\hat{\boldsymbol{\beta}})\boldsymbol{C}^T)^{-1}(\boldsymbol{C}\hat{\boldsymbol{\beta}} - \boldsymbol{\xi})$$

**Score Statistic**

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

**Approximation**

$$\lambda, w, u \sim \chi^2(s)$$

## 3.8 Goodness-of-Fit for Grouped Observations

It has already been mentioned that for grouped observations the deviance has an asymptotic $\chi^2$-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 $\tilde{M}$ may be chosen instead of the more general model $M$, where $\tilde{M} \subset M$. The test statistics may be seen as tools to investigate the fit of model $\tilde{M}$ *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 $M$ 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 $\boldsymbol{x}_i$, independent replications $y_{i1}, \ldots, y_{in_i}$ are observed, the mean across replications $\bar{y}_i = \sum_j y_{ij}/n_i$ again represents a GLM and the deviance for the means $\bar{y}_1, \ldots, \bar{y}_N$ may be used. For grouped data with response $\bar{y}_i$, $i = 1, \ldots, N$, the essential difference is that the scale parameter is given as $\phi_i = \phi/n_i$, where $\phi$ is the dispersion for the single observations. Since

grouped observations $y_i, \ldots y_{in_i}$ share the same predictor value $\boldsymbol{x}_i$, the log-likelihood is given by

$$l = \sum_{i=1}^{N}\sum_{j=1}^{n_i}(y_{ij}\theta_i - b(\theta_i))/\phi + c(y_{ij},\phi) = \sum_{i=1}^{N}\frac{\bar{y}_i\theta_i - b(\theta_i)}{\phi/n_i} + \sum_{i=1}^{N}\sum_{j=1}^{n_i}c(y_{ij},\phi).$$

Thus maximization with respect to $\boldsymbol{\beta}$ yields the same results if $l$ is maximized in the ungrouped or grouped form. The deviance, however, changes for grouped data because the saturated model for data $(\bar{y}_i, x_i), i = 1, \ldots, N$, means that only the $N$ observations $\bar{y}, \ldots, \bar{y}_N$ have to be fitted perfectly.

With $\bar{\boldsymbol{y}}^T = (\bar{y}_1, \ldots, \bar{y}_N)$, $\hat{\boldsymbol{\mu}}^T = (\hat{\boldsymbol{\mu}}_1, \ldots, \hat{\boldsymbol{\mu}}_N)$, and $\hat{\boldsymbol{\mu}}_i = h(\boldsymbol{x}_i^T\hat{\boldsymbol{\beta}})$, where $\boldsymbol{\mu}_i = \mathrm{E}(\bar{y}_i) = \mathrm{E}(y_{ij})$, $j = 1, \ldots, n_i$, the deviance for grouped observations has the form

$$\begin{aligned} D(\bar{\boldsymbol{y}}, \hat{\boldsymbol{\mu}}) &= -\,\phi\,2\,\{l(\bar{\boldsymbol{y}}; \hat{\boldsymbol{\mu}}, \phi) - l(\bar{\boldsymbol{y}}; \bar{\boldsymbol{y}}, \phi)\} \\ &= 2\sum_{i=1}^{N}\{\bar{y}_i(\theta(\bar{y}_i) - \theta(\hat{\boldsymbol{\mu}}_i)) - (b(\theta(\bar{y}_i)) - b(\theta(\hat{\boldsymbol{\mu}}_i)))\}n_i. \end{aligned}$$

The deviances for various distributions are given in Table 3.4. For ungrouped data with $N = n, n_i = 1$, one obtains the deviances as given in Table 3.2. The grouped deviance for Bernoulli variables is equivalent to the deviance of the binomial distribution. The reason is obvious because the binomial distribution implicitly assumes replications.

TABLE 3.4: Deviances for grouped observations.

| Distribution | Deviance for grouped observations |
|---|---|
| Normal | $D(\bar{\boldsymbol{y}}, \hat{\mu}) = \sum_{i=1}^{N} n_i(\bar{y}_i - \hat{\mu}_i)^2$ |
| Gamma | $D(\bar{\boldsymbol{y}}, \hat{\mu}) = 2\sum_{i=1}^{N} n_i\left(-\log\left(\frac{\bar{y}_i}{\hat{\mu}_i}\right) + \frac{\bar{y}_i - \hat{\mu}_i}{\hat{\mu}_i}\right)$ |
| Inverse Gaussian | $D(\bar{\boldsymbol{y}}, \hat{\mu}) = \sum_{i=1}^{N} n_i(\bar{y}_i - \hat{\mu}_i)^2/(\hat{\mu}_i^2\bar{y}_i)$ |
| Bernoulli | $D(\bar{\boldsymbol{y}}, \hat{\mu}) = 2\sum_{i=1}^{N}\left(n_i\bar{y}_i\log\left(\frac{\bar{y}_i}{\hat{\mu}_i}\right) + n_i(1-\bar{y}_i)\log\left(\frac{1-\bar{y}_i}{1-\hat{\mu}_i}\right)\right)$ |
| Poisson | $D(\bar{\boldsymbol{y}}, \hat{\mu}) = 2\sum_{i=1}^{N} n_i\left(\bar{y}_i\log\left(\frac{\bar{y}_i}{\hat{\mu}_i}\right) - (\bar{y}_i - \hat{\mu}_i)\right)$ |

The advantage of replications or grouped data is that for this kind of data the scaled deviance or likelihood ratio statistic $D(\bar{\boldsymbol{y}}, \hat{\boldsymbol{\mu}})/\phi$ provides a goodness-of-fit statistic that may be used to test if the model is appropriate for the data. The maximal likelihood $l(\bar{\boldsymbol{y}}; \bar{\boldsymbol{y}}, \phi)$ is the likelihood of a model with $N$ parameters, one per covariate value $\boldsymbol{x}_i$, where only the assumption of distribution with independent, identically distributed responses is made. Thus $D(\bar{\boldsymbol{y}}, \hat{\boldsymbol{\mu}})/\phi$ may be used to test the current model, implying the form of the linear predictors and a specific link function, against the distribution model. Under fixed cells asymptotics ($N$ fixed, $n_i \to \infty, n_i/n \to c_i, c_i > 0$) and regularity conditions one obtains for $D(\bar{\boldsymbol{y}}, \hat{\boldsymbol{\mu}})/\phi$ a limiting $\chi^2$-distribution with $N - p$ degrees of freedom, where $p$ denotes the dimension of the predictor $\boldsymbol{x}_i$. If $D(\bar{y}, \hat{\boldsymbol{\mu}})/\phi$ is larger than the $1 - \alpha$ quantile of $\chi^2(N - p)$, the model is questionable as a tool for investigating the connection between covariates and responses.

It should be noted that the difference between likelihoods for different models and therefore the analysis of variance yields the same results for ungrouped and grouped modeling. For two

models $\tilde{M} \subset M$ and corresponding fits $\hat{\mu}_i, \tilde{\mu}_i$ , the difference $D(\bar{\boldsymbol{y}}, \tilde{\boldsymbol{\mu}}) - D(\bar{\boldsymbol{y}}, \hat{\boldsymbol{\mu}})$ for grouped data are the same as the difference $D(\boldsymbol{y}, \tilde{\boldsymbol{\mu}}) - D(\boldsymbol{y}, \hat{\boldsymbol{\mu}})$ for ungrouped observations. Therefore $(D(\bar{\boldsymbol{y}}, \tilde{\boldsymbol{\mu}}) - D(\bar{\boldsymbol{y}}, \hat{\boldsymbol{\mu}}))/\phi$ is asymptotically $\chi^2(s)$-distributed, where $s$ is the difference between the number of parameters in $M$ and $\tilde{M}$.

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

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

which has the form

$$D(\tilde{M}) = D(\tilde{M}|M) + D(M),$$

where $\tilde{M}$ stands for the linear model and $M$ for a model where only $y_{ij} = \mu_i + \varepsilon_{ij}$ with $\varepsilon_{ij} \sim \mathrm{N}(0, \sigma^2)$ is assumed. Since in computing $D(M)$ no assumption on linearity is assumed, it is also called the *pure error sum of squares* and $D(\tilde{M}|M)$ the *lack of fit sum of squares*. By use of the mean squares one obtains for $\boldsymbol{H}_0 : \boldsymbol{\mu} = \boldsymbol{X}\boldsymbol{\beta}$ against $H_1 : \boldsymbol{\mu} \neq \boldsymbol{X}\boldsymbol{\beta}$ the $F$-statistic

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

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 $D(\tilde{M}|M)$ 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:

$$\chi^2_P = \sum_{i=1}^{N} \frac{(\bar{y}_i - \hat{\mu}_i)^2}{v(\hat{\mu}_i)/n_i}, \tag{3.12}$$

where $\bar{y}_i$ is the mean for grouped observations, $\hat{\mu}_i$ is the estimated mean, and $v(\hat{\mu}_i)$ is the corresponding variance function that is linked to the variance by $\mathrm{var}(y_i) = v(\mu_i)\phi/n_i$. If fixed cells asymptotics applies ($N$ fixed, $n_i \to \infty$, $n_i/n \to c_i, c_i > 0$), $\chi^2_P$ is asymptotically $\chi^2$-distributed, that is, $\chi^2_P$ has an approximately $\phi\chi^2(N-p)$-distribution. 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 $\chi^2_p/\hat{\phi}_N = N - p$.

Goodness-of-Fit for Grouped Observations

Deviance

$$D = -\phi\, 2 \sum_{i=1}^{N} \{l(\bar{y}_i; \hat{\mu}_i, \phi) - l(\bar{y}_i; \bar{y}_i, \phi)\}$$

Pearsons $\chi^2$

$$\chi^2 = \sum_{i=1}^{N} n_i \frac{(\bar{y}_i - \hat{\mu}_i)^2}{v(\hat{\mu}_i)}$$

Fixed cells asymptotic ($N$ fixed, $n_i \to \infty$, $n_i/n \to c_i, c_i > 0$)

$$D,\ \chi^2 \text{ approximatively } \phi\, \chi^2(N-p)$$

## 3.9 Computation of Maximum Likelihood Estimates

Maximum likelihood estimates are obtained by solving the equation $\boldsymbol{s}(\hat{\boldsymbol{\beta}}) = \mathbf{0}$. 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

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

where in $\boldsymbol{D}$, $\boldsymbol{\Sigma}$, and $\boldsymbol{W}$ the dependence on $\boldsymbol{\beta}$ is suppressed (see Section 3.6).

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

$$s(\boldsymbol{\beta}) \approx s_{\text{lin}}(\boldsymbol{\beta}) = s(\hat{\boldsymbol{\beta}}^{(k)}) + \frac{\partial s(\hat{\boldsymbol{\beta}}^{(k)})}{\partial \boldsymbol{\beta}} (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}^{(k)}).$$

Instead of solving $s(\hat{\boldsymbol{\beta}}) = 0$ one solves $s_{\text{lin}}(\hat{\boldsymbol{\beta}}) = 0$, yielding

$$\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}^{(k)} - (\frac{\partial s(\hat{\boldsymbol{\beta}}^{(k)})}{\partial \boldsymbol{\beta}})^{-1} s(\hat{\boldsymbol{\beta}}^{(k)}).$$

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

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

or, by using the observed information matrix $\boldsymbol{F}_{\text{obs}}(\boldsymbol{\beta}) = -H(\boldsymbol{\beta})$,

$$\hat{\boldsymbol{\beta}}^{(k+1)} = \hat{\boldsymbol{\beta}}^{(k)} + \boldsymbol{F}_{\text{obs}}(\hat{\boldsymbol{\beta}}^{(k)})^{-1} s(\hat{\boldsymbol{\beta}}^{(k)}).$$

Iterations are carried out until the changes between successive steps are smaller than a specified threshold $\varepsilon$. Iteration is stopped if

$$\| \hat{\boldsymbol{\beta}}^{(k+1)} - \hat{\boldsymbol{\beta}}^{(k)} \| \ / \ \| \hat{\boldsymbol{\beta}}^{(k)} \| < \varepsilon.$$

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 $\boldsymbol{F}_{\text{obs}}$ is replaced by the expected information $\boldsymbol{F}(\boldsymbol{\beta}) = \text{E}(\boldsymbol{F}_{\text{obs}}(\boldsymbol{\beta}))$ (or $\boldsymbol{H}(\beta)$ by $-\boldsymbol{F}(\boldsymbol{\beta})$), yielding

$$\hat{\boldsymbol{\beta}}^{(k+1)} = \hat{\boldsymbol{\beta}}^{(k)} + \boldsymbol{F}(\hat{\boldsymbol{\beta}}^{(k)})^{-1} s(\hat{\boldsymbol{\beta}}^{(k)}). \tag{3.13}$$

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

$$\tilde{\eta_i}(\hat{\boldsymbol{\beta}}) = \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}} + \left( \frac{\partial h(\eta_i)}{\partial \eta} \right)^{-1} (y_i - \mu_i(\hat{\boldsymbol{\beta}}))$$

and $\tilde{\boldsymbol{\eta}}(\hat{\boldsymbol{\beta}})^T = (\tilde{\eta_1}(\hat{\boldsymbol{\beta}}), \ldots, \tilde{\eta_n}(\hat{\boldsymbol{\beta}}))$ denote the vector of pseudo-observations given by $\tilde{\boldsymbol{\eta}}(\hat{\boldsymbol{\beta}}) = \boldsymbol{X}\hat{\boldsymbol{\beta}} + \boldsymbol{D}(\hat{\boldsymbol{\beta}})^{-1}(\boldsymbol{y} - \hat{\boldsymbol{\mu}})$. One obtains by simple substitution

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

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

For a canonical link one obtains $\boldsymbol{F}_n(\boldsymbol{\beta}) = \boldsymbol{X}^T \boldsymbol{W} \boldsymbol{X}$ with $\boldsymbol{W} = \boldsymbol{\phi}^{-1} \boldsymbol{\Sigma} \boldsymbol{\phi}^{-1}$, $\boldsymbol{\phi} = \text{diag}(\phi_1, \ldots, \phi_n)$ and score function $s(\boldsymbol{\beta}) = \boldsymbol{X}^T \boldsymbol{\phi}^{-1}(\boldsymbol{y} - \boldsymbol{\mu})$ and therefore

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

which corresponds to least-squares fitting

$$\hat{\boldsymbol{\beta}}^{(k+1)} = (\boldsymbol{X}^T \boldsymbol{W} \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{W} \boldsymbol{\eta}(\hat{\boldsymbol{\beta}}^{(k)})$$

with $\tilde{\boldsymbol{\eta}}(\boldsymbol{\beta}) = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{W}^{-1} \boldsymbol{\phi}^{-1}(\boldsymbol{y} - \boldsymbol{\mu})$. If $\boldsymbol{\phi} = \phi \boldsymbol{I}$, one obtains

$$\hat{\boldsymbol{\beta}}^{(k+1)} = \hat{\boldsymbol{\beta}}^{(k)} + \phi(\boldsymbol{X}^T \boldsymbol{\Sigma} \boldsymbol{X})^{-1} \boldsymbol{X}^T (\boldsymbol{y} - \boldsymbol{\mu}).$$

## 3.10 Hat Matrix for Generalized Linear Models

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

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

At convergence one obtains

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T \boldsymbol{W}(\hat{\boldsymbol{\beta}}) \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{W}(\hat{\boldsymbol{\beta}}) \tilde{\boldsymbol{\eta}}(\hat{\boldsymbol{\beta}}).$$

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

$$\boldsymbol{W}^{T/2} \tilde{\boldsymbol{\eta}}(\hat{\boldsymbol{\beta}}) = \boldsymbol{W}^{T/2} \boldsymbol{X} \boldsymbol{\beta} + \tilde{\boldsymbol{\varepsilon}},$$

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

$$\boldsymbol{H} = \boldsymbol{W}^{T/2} \boldsymbol{X} (\boldsymbol{X}^T \boldsymbol{W} \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{W}^{1/2}.$$

Since the matrix $\boldsymbol{H}$ is idempotent and symmetric, it may be seen as a projection matrix for which $\text{tr}(\boldsymbol{H}) = \text{rank}(\boldsymbol{H})$ holds. Moreover, one obtains for the diagonal elements of $\boldsymbol{H} = (h_{ij})$ $0 \leq h_{ii} \leq 1$ and $\text{tr}(\boldsymbol{H}) = p$ (if $X$ has full rank).

It should be noted that, in contrast to the normal regression model, the hat matrix depends on $\hat{\boldsymbol{\beta}}$ because $\boldsymbol{W} = \boldsymbol{W}(\hat{\boldsymbol{\beta}})$. The equation $\boldsymbol{W}^{T/2}\hat{\boldsymbol{\eta}} = \boldsymbol{H}\boldsymbol{W}^{T/2}\tilde{\boldsymbol{\eta}}(\boldsymbol{\beta})$ shows how the hat matrix maps the adjusted variable $\tilde{\boldsymbol{\eta}}(\boldsymbol{\beta})$ into the fitted values $\hat{\boldsymbol{\eta}}$. Thus $\boldsymbol{H}$ may be seen as the matrix that maps the adjusted observation vector $\boldsymbol{W}^{T/2}\tilde{\boldsymbol{\eta}}$ into the vector of "fitted" values $\boldsymbol{W}^{T/2}\hat{\boldsymbol{\eta}}$, which is a mapping on the transformed predictor space.

For the linear model the hat matrix represents a simple projection having the form $\hat{\boldsymbol{\mu}} = \boldsymbol{H}\boldsymbol{y}$. In the case of generalized linear models, it may be shown that approximatively

$$\boldsymbol{\Sigma}^{-1/2}(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}) \simeq \boldsymbol{H}\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{y} - \boldsymbol{\mu}) \tag{3.14}$$

holds, where $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}})$. Thus $\boldsymbol{H}$ may be seen as measure of the influence of $\boldsymbol{y}$ on $\hat{\boldsymbol{\mu}}$ in standardized units of changes. From (4.6) follows

$$\hat{\boldsymbol{\mu}} - \boldsymbol{\mu} \simeq \boldsymbol{\Sigma}^{1/2}\boldsymbol{H}\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{y} - \boldsymbol{\mu}), \tag{3.15}$$

such that the influence in unstandardized units is given by the projection matrix $\boldsymbol{\Sigma}^{1/2}\boldsymbol{H}\boldsymbol{\Sigma}^{-1/2}$, which is idempotent but not symmetric. Note that for the normal regression model with an identity link one has $\boldsymbol{W} = \boldsymbol{I}/\sigma^2$, $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$, and (4.6) and (3.15) hold exactly. From (3.15) one derives the approximation

$$\text{cov}(\boldsymbol{y} - \hat{\boldsymbol{\mu}}) \simeq \boldsymbol{\Sigma}^{1/2}(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{\Sigma}^{T/2}.$$

An alternative property is obtained from considering the estimating equation $\boldsymbol{X}^T\hat{\boldsymbol{W}}\hat{\boldsymbol{D}}^{-T}(\boldsymbol{y} - \hat{\boldsymbol{\mu}}) = \boldsymbol{0}$, which yields directly

$$\boldsymbol{P}_{\boldsymbol{X},\boldsymbol{W}}\hat{\boldsymbol{D}}^{-T}\hat{\boldsymbol{\mu}} = \boldsymbol{P}_{\boldsymbol{X},\boldsymbol{W}}\hat{\boldsymbol{D}}^{-T}\boldsymbol{y}, \tag{3.16}$$

with $\boldsymbol{P}_{\boldsymbol{X},\boldsymbol{W}} = \boldsymbol{X}(\boldsymbol{X}^T\hat{\boldsymbol{W}}(\hat{\boldsymbol{\beta}})\boldsymbol{X})^{-1}\boldsymbol{X}^T\hat{\boldsymbol{W}}$ being the projection on the subspace span $(\boldsymbol{Z})$ with respect to the matrix $\boldsymbol{W}$, which means an orthogonal projection with respect to the product $\boldsymbol{x}_1^T\boldsymbol{W}\boldsymbol{x}_2$, $\boldsymbol{x}_1, \boldsymbol{x}_2 \in I\!R^n$. $\boldsymbol{P}_{\boldsymbol{Z},\boldsymbol{W}}$ is idempotent but not symmetric. Thus the projections of the transformed values $\hat{\boldsymbol{D}}^{-T}\boldsymbol{y}$, $\hat{\boldsymbol{D}}^{-T}\hat{\boldsymbol{\mu}}$ on the $\eta$-space are identical. From (3.16) one obtains easily

$$\boldsymbol{H}\boldsymbol{W}^{T/2}\boldsymbol{D}^{-T}\hat{\boldsymbol{\mu}} = \boldsymbol{H}\boldsymbol{W}^{T/2}\boldsymbol{D}^{-T}\boldsymbol{y},$$

which yields

$$\boldsymbol{H}\boldsymbol{\Sigma}^{-1/2}\hat{\boldsymbol{\mu}} = \boldsymbol{H}\boldsymbol{\Sigma}^{-1/2}\boldsymbol{y},$$

meaning that the orthogonal projections (based on $\boldsymbol{H}$) of standardized values $\boldsymbol{\Sigma}^{-1/2}\hat{\boldsymbol{\mu}}$, $\boldsymbol{\Sigma}^{-1/2}\boldsymbol{y}$ are identical. With $\boldsymbol{\chi} = \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{y} - \hat{\boldsymbol{\mu}})$ denoting the standardized residual, one has

$$\boldsymbol{H}\boldsymbol{\chi} = 0 \text{ and } (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{\chi} = \boldsymbol{\chi}.$$

There is a strong connection to the $\chi^2$-statistic since $\chi^2 = \boldsymbol{\chi}^T\boldsymbol{\chi}$. The matrix $\boldsymbol{H}$ has the form $\boldsymbol{H} = (\boldsymbol{W}^{T/2}\boldsymbol{X})(\boldsymbol{X}^T\boldsymbol{W}^{1/2}\boldsymbol{W}^{T/2}\boldsymbol{X})^{-1}(\boldsymbol{X}^T\boldsymbol{W}^{1/2})$, which shows that the projection is into the subspace that is spanned by the columns of $\boldsymbol{X}^T\boldsymbol{W}^{1/2}$. The essential difference from ordinary linear regression is that the hat matrix does depend not only on the design but also on the fit.

## 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

$$\mathrm{E}(y_i|\boldsymbol{x}_i) \;=\; \mu_i \;=\; h(\boldsymbol{x}^T\boldsymbol{\beta}), \qquad \mathrm{var}(y_i|\boldsymbol{x}_i) \;=\; \sigma_i^2(\mu_i) \;=\; \phi v(\mu_i), \tag{3.17}$$

where $v(\mu)$ 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:

$$\boldsymbol{s}(\hat{\boldsymbol{\beta}}) = \boldsymbol{X}^T\boldsymbol{D}(\hat{\boldsymbol{\beta}})\boldsymbol{\Sigma}^{-1}(\hat{\boldsymbol{\beta}})(\boldsymbol{y}-\boldsymbol{\mu}) = \boldsymbol{0}. \tag{3.18}$$

The ML estimation equation uses the link (in $\boldsymbol{D}(\hat{\boldsymbol{\beta}})$) and the variance function (in $\boldsymbol{\Sigma}^{-1}(\hat{\boldsymbol{\beta}})$), 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

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

with the specifications (3.17) as a quasi-score function with the corresponding estimation equation $\boldsymbol{s}_Q(\hat{\boldsymbol{\beta}}) = \boldsymbol{0}$. It is also possible to construct a quasi-likelihood function $Q(\boldsymbol{\beta},\phi)$ that has the derivative $\boldsymbol{s}_Q(\boldsymbol{\beta}) = \partial Q/\partial\boldsymbol{\beta}$ (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:

$$\boldsymbol{F}_Q(\boldsymbol{\beta}) = \sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i^T\left(\frac{\partial h(\eta_i)}{\partial\eta}\right)^2/\sigma_i^2 = \boldsymbol{X}^T\boldsymbol{D}\boldsymbol{\Sigma}^{-1}\boldsymbol{D}\boldsymbol{X} = \boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X}$$

(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 $v(\mu_i)$ 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 $\mathrm{var}(y_i) = \mu_i = \exp(\boldsymbol{x}_i^T\boldsymbol{\beta})$. In quasi-likelihood approaches one might assume that the variance is $\mathrm{var}(y_i) = \phi\mu_i = \phi\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})$, 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).

More flexible models allow the variance function to depend on additional parameters (e.g., Nelder and Pregibon, 1987). Then the variance function has the form $\phi v(\mu;\alpha)$, where $\alpha$ is an additional unknown parameter. An example is $v(\mu;\alpha) = \mu^\alpha$. For fixed $\alpha$, the quasi-likelihood estimate $\hat{\boldsymbol{\beta}}$ (and an estimate $\hat{\phi}$) is obtained by solving the corresponding estimation equation. Estimation of $\alpha$, given $\hat{\boldsymbol{\beta}}$ and $\hat{\phi}$, can be carried out by the method of moments. Cycling between the two steps until convergence gives a joint estimation procedure. Asymptotic results for $\hat{\boldsymbol{\beta}}$ remain valid if $\alpha$ is replaced by a consistent estimate $\hat{\alpha}$.

Within quasi-likelihood approaches one can also model that the dispersion parameter depends on covariates. Then one assumes $\mu = h(\boldsymbol{x}^T\boldsymbol{\beta})$, $\phi = \phi(\boldsymbol{z}^T\boldsymbol{\gamma})$, $\text{var}(y) = \phi v(\mu)$, where $\boldsymbol{z}$ is a vector of covariates affecting the dispersion parameter. Cycling between the estimating equation for $\boldsymbol{\beta}$ and an estimation equation for $\boldsymbol{\gamma}$ yields estimates for both parameters. Alternatively, a joint estimation of parameters can be obtained by the general techniques for fitting likelihood and quasi-likelihood models described by Gay and Welsch, 1988. Consistent estimation of $\beta$ and $\alpha$ requires that not only the mean but also the dispersion parameter be correctly specified (for details see Pregibon, 1984; Nelder and Pregibon, 1987; Efron, 1986; McCullagh and Nelder, 1989; Nelder, 1992).

## 3.12 Further Reading

*Surveys and Books*. A source book on generalized linear models is McCullagh and Nelder (1989). Multivariate extensions are covered in Fahrmeir and Tutz (2001). Shorter introductions were given by Dobson (1989) and Firth (1991). A Bayesian perspective on generalized linear models is outlined in Dey et al. (2000). More recently, a general treatment of regression, including GLMs was given by Fahrmeir et al. (2011).

*Quasi-likelihood*. Quasi-likelihood estimates were considered by Wedderburn (1974), McCullagh (1983), and McCullagh and Nelder (1989). Efficiency of quasi-likelihood estimates was investigated by Firth (1987). A rigorous mathematical treatment was given by Heyde (1997). Asymptotic properties were also studied by Xia et al. (2008).

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

## 3.13 Exercises

3.1 Let independent observations $y_{i1},\ldots,y_{in_i}$ have a fixed covariate vector $\boldsymbol{x}_i$ with $n_i$ denoting the sample size at covariate value $\boldsymbol{x}_i$. The model to be examined has the form $\mu_i = \mu_{ij} = \text{E}(y_{ij}) = h(\boldsymbol{x}_i^T\boldsymbol{\beta})$.

(a) Let observations be binary with $y_{ij} \in \{0,1\}$. Show that the distribution of the mean $\bar{y}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} y_{ij}$ also has the form of a simple exponential family. Compare the canonical parameter and the dispersion parameter with the values of the exponential family of the original binary response.

(b) Show that the mean $\bar{y}_i$ always has the form of a simple exponential family if the mean of $y_{ij}$ has the form of a simple exponential family.

3.2 Let observations $(y_i,\boldsymbol{x}_i)$, $i = 1,\ldots,n$, with binary response $y_i \in \{0,1\}$ be given. The used models are the logit model $P(y_i = 1|\boldsymbol{x}_i) = \Phi(\boldsymbol{x}_i^T\boldsymbol{\beta})$, with $F$ denoting the logistic distribution function, and the probit model $P(y_i = 1|\boldsymbol{x}_i) = \Phi(\boldsymbol{x}_i^T\boldsymbol{\beta})$, with $\Phi$ denoting the standard normal distribution function. Give the log-likelihood function and derive the score function $\boldsymbol{s}(\boldsymbol{\beta})$, the matrix of derivatives $\boldsymbol{D}(\boldsymbol{\beta})$, and the variance matrix $\boldsymbol{\Sigma}(\boldsymbol{\beta})$ 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 $\text{cov}(\hat{\boldsymbol{\beta}}) \approx (\boldsymbol{X}^T \text{diag}(\hat{\boldsymbol{\mu}})\boldsymbol{X})^{-1}$, where $\log(\hat{\boldsymbol{\mu}}) = \boldsymbol{X}\boldsymbol{\beta}$.

3.4 Let the independent observations $y_{i1}, \ldots, y_{in_i}$, observed at predictor value $\boldsymbol{x}_i$, follow a Poisson distribution, $y_{ij} \sim \text{P}(\lambda_{ij})$. Then one obtains for the sum $\tilde{y}_i = \sum_{j=1}^{n_i} y_{ij} \sim \text{P}(\tilde{\lambda}_i)$, where $\tilde{\lambda}_i = \sum_j \lambda_{ij}$. Discuss modeling strategies for the observations. Consider in particular models for single variables $y_{ij}$, for accumulated counts $\tilde{y}_i$, and for average counts $\tilde{y}_i/n_i$.

3.5 Let $(y_i, \boldsymbol{x}_i)$ denote independent observations. A linear model with log-transformed responses is given by $\log(y_i) = \boldsymbol{x}_i^T\boldsymbol{\beta} + \epsilon_i$, $\epsilon_i \sim \text{N}(0, \sigma^2)$. Compare the model to the GLM

$$y_i|\boldsymbol{x}_i \sim \text{N}(\mu_i, \sigma^2) \qquad \text{and} \qquad \mu_i = \exp(\boldsymbol{x}_i^T\boldsymbol{\beta}),$$

and explain the difference between the two models.

3.6 The R Package *catdat* provides the data set *rent*.

(a) Find a GLM with response *rent* (net rent in Euro) and the explanatory variables *size* (size in square meters) and *rooms* (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 binary regression model. In this section we will discuss modeling issues for the more general binary regression model

$$P(y_i = 1|\boldsymbol{x}_i) = \pi(\boldsymbol{x}_i) = h(\boldsymbol{x}_i^T\boldsymbol{\beta}),$$

where the response function $h$ is a fully specified function, which in the case of the logit model is the logistic distribution function $h(\eta) = \exp(\eta)/(1 + \exp(\eta))$. 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 regression 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

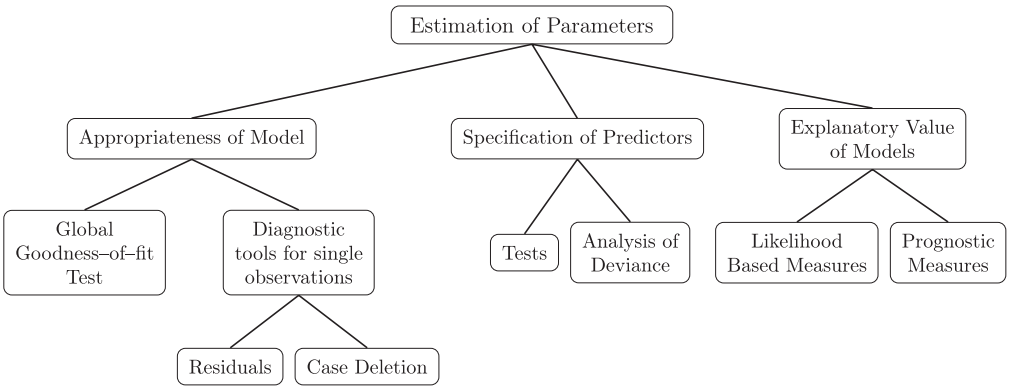


FIGURE 4.1: Aspects of regression modeling.

simplified. On the other hand, the specification determines the goodness-of-fit of the model. There is also a strong link between the specification of the linear predictor and the explanatory value of the covariates. Only the focus is different. While the specification of the linear predictor aims at finding an adequate form of the covariates and at reducing the variables to a set of variables that is really needed, the explanatory value of a model aims at quantifying the effect of the covariates within the model.

Since the estimation of parameters plays such a prominent role in modeling, the first section of this chapter is devoted to estimation, in particular maximum likelihood (ML) estimation. Since most of the tools considered in this chapter are likelihood-based, we will defer alternative estimation concepts to later sections. In this chapter the focus is on tools; therefore, only part of the model structures will be discussed. For example, the specification of the link function and weaker distributional assumptions will be considered later, in Chapter 5.

We will consider here in particular tools for the following aspects:

- Discrepancy between data and fit: global goodness-of-fit of a model (Section 4.2)
- Diagnostic tools for single observations (Section 4.3)
- Specification of the linear predictor (Section 4.4)
- Explanatory value of covariates (Section 4.6)

## 4.1 Maximum Likelihood Estimation

When a model is assumed to represent a useful relationship between an observed response variable and several explanatory variables, the first step in inference is the estimation of the unknown parameters. In the linear logit model

$$P(y_i = 1|\boldsymbol{x}_i) = \frac{\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})}{1 + \exp(\boldsymbol{x}_i^T\boldsymbol{\beta})},$$

as well as in the more general model $\pi(\boldsymbol{x}_i) = h(\boldsymbol{x}_i^T\boldsymbol{\beta})$, the parameters to be estimated are the regression parameters $\boldsymbol{\beta}$. In the following we consider the more general model, where $h$ is a fully specified function. The logit model uses $h(\eta) = \exp(\eta)/(1 + \exp(\eta))$. Alternative response functions like the normal distribution function, which yields the probit model, will be considered in Chapter 5.

The most widely used general method of estimation in binary regression models is maximum likelihood. The basic principle is to construct the likelihood of the unknown parameters for the sample data. The likelihood represents the joint probability or probability density of the observed data, considered as a function of the unknown parameters.

In the following we will distinguish between the case of single binary responses and the more general case of scaled binomials (or proportions) $\bar{y}_i$. A proportion $\bar{y}_i$ is computed from $n_i$ independent binary observations observed at the same measurement point $\boldsymbol{x}_i$.

## Single Binary Responses

For independent observations $(y_i, \boldsymbol{x}_i), i = 1, \ldots, n$, with $y_i \in \{0, 1\}$, the *likelihood* for the conditional responses $y_i|\boldsymbol{x}_i$ is given by

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \pi(\boldsymbol{x}_i)^{y_i}(1 - \pi(\boldsymbol{x}_i))^{1-y_i}.$$

Each term in the product represents the probability that $y_i$ is observed since $\pi(\boldsymbol{x}_i)^{y_i}(1 - \pi(\boldsymbol{x}_i))^{1-y_i}$ simplifies to $\pi(\boldsymbol{x}_i)$ if $y_i = 1$ and to $1 - \pi(\boldsymbol{x}_i)$ if $y_i = 0$. The product is built since the observations $y_1, \ldots, y_n$, given $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$, are considered as independent. The maximum likelihood estimates of $\boldsymbol{\beta}$ are those values $\hat{\boldsymbol{\beta}}$ that maximize the likelihood. It is usually more convenient to maximize the log-likelihood rather than the likelihood itself. Since the logarithm is a strictly monotone transformation, the obtained values $\hat{\boldsymbol{\beta}}$ will be the same. Therefore, the *log-likelihood*

$$l(\boldsymbol{\beta}) = \log(L(\boldsymbol{\beta})) = \sum_{i=1}^{n} l_i(\boldsymbol{\beta}) = \sum_{i=1}^{n} y_i \log(\pi(\boldsymbol{x}_i)) + (1 - y_i) \log(1 - \pi(\boldsymbol{x}_i))$$

is used. The value $\hat{\boldsymbol{\beta}}$ that maximizes $l(\boldsymbol{\beta})$ can be obtained by solving the system of equations $\partial l(\boldsymbol{\beta})/\partial \boldsymbol{\beta} = \mathbf{0}$. It is common to consider the derivatives that yield the equations as functions of $\boldsymbol{\beta}$. One considers the so-called *score function*:

$$\boldsymbol{s}(\boldsymbol{\beta}) = \partial l(\boldsymbol{\beta})/\partial \boldsymbol{\beta} = (\partial l(\boldsymbol{\beta})/\partial \beta_1, \ldots, \partial l(\boldsymbol{\beta})/\partial \beta_p)^T,$$

which has the form

$$\boldsymbol{s}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \boldsymbol{s}_i(\boldsymbol{\beta}) = \sum_{i=1}^{n} \boldsymbol{x}_i \frac{\partial h(\eta_i)}{\partial \eta} \frac{(y_i - \pi(\boldsymbol{x}_i))}{\sigma_i^2},$$

where $\eta_i = \boldsymbol{x}_i^T \boldsymbol{\beta}$ and $\sigma_i^2 = var(y_i) = \pi(\boldsymbol{x}_i)(1 - \pi(\boldsymbol{x}_i))$. The maximum likelihood (ML) estimate is then found by solving $\boldsymbol{s}(\hat{\boldsymbol{\beta}}) = \mathbf{0}$. The system of equations has to be solved iteratively (see Section 3.9).

For the logit model one obtains by simple calculation

$$\frac{\partial l(\boldsymbol{\beta})}{\partial \beta_j} = \sum_{i=1}^{n} x_{ij}(y_i - \pi(\boldsymbol{x}_i)),$$

yielding the score function $\boldsymbol{s}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \boldsymbol{x}_i(y_i - \pi(\boldsymbol{x}_i))$. Therefore, the likelihood equations to be solved are

$$\sum_{i=1}^{n} x_{ij} y_i = \sum_{i=1}^{n} x_{ij} \pi(\boldsymbol{x}_i)),$$

which equate the sufficient statistics $\sum_{i=1}^{n} x_{ij}y_i$, $j = 1\ldots, p$ for $\boldsymbol{\beta}$ to their expected values (Exercise 4.3).

In maximum likelihood theory, the asymptotic variance of the estimator is determined by the *information* or *Fisher matrix* $\boldsymbol{F}(\boldsymbol{\beta}) = E(-\frac{\partial^2 l(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T})$. As derived in Chapter 3, $\boldsymbol{F}(\boldsymbol{\beta})$ is given by $\boldsymbol{F}(\boldsymbol{\beta}) = \boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X}$, where $\boldsymbol{X}$ with $\boldsymbol{X}^T = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ is the design matrix and $\boldsymbol{W} = \text{Diag}\,((\frac{\partial h(\eta_1)}{\partial \eta})^2/\sigma_1^2, \ldots, (\frac{\partial h(\eta_n)}{\partial \eta})^2/\sigma_n^2)$ is a diagonal weight matrix.

For the logit model one obtains the simpler form $\boldsymbol{F}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \boldsymbol{x}_i\boldsymbol{x}_i^T\sigma_i^2 = \boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X}$ with weight matrix $\boldsymbol{W} = (\sigma_1^2, \ldots, \sigma_n^2)$. An approximation to the covariance of $\hat{\boldsymbol{\beta}}$ is given by

$$\hat{cov}(\hat{\boldsymbol{\beta}}) \approx \boldsymbol{F}(\boldsymbol{\beta})^{-1} = (\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X})^{-1},$$

where in practice $\boldsymbol{W}$ is replaced by $\hat{\boldsymbol{W}}$, which denotes the evaluation of $\boldsymbol{W}$ at $\hat{\boldsymbol{\beta}}$.

## Grouped Data: Estimation for Binomially Distributed Responses

In many applications, for given values of the covariates, several independent binary responses are observed. Since $P(y = 1|\,\boldsymbol{x}) = \pi(\boldsymbol{x})$ is assumed to depend on $\boldsymbol{x}$ only, the mean is assumed to be the same for all the binary observations collected at this value. More formally, let $y_{i1}, \ldots, y_{in_i}$ denote the independent dichotomous responses collected at value $\boldsymbol{x}_i$ and let $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$ denote the distinct values of covariates or measurement points where responses are observed. The model has the form

$$P(y_{ij} = 1|\boldsymbol{x}_i) = h(\boldsymbol{x}_i^T\boldsymbol{\beta}), j = 1, \ldots, n_i,$$

or simpler $\pi(\boldsymbol{x}_i) = h(\boldsymbol{x}_i^T\boldsymbol{\beta})$, where $\pi(\boldsymbol{x}_i) = P(y_{ij} = 1|\boldsymbol{x}_i), j = 1, \ldots, n_i$.

For the purpose of estimation one may use the original binary variables $y_{i1}, \ldots, y_{in_i}$, $i = 1, \ldots, N$, $(y_{ij} \sim B(1, \pi(\boldsymbol{x}_i))$ or, equivalently, the binomial distribution of $y_{i1} + \cdots + y_{in_i} \sim B(n_i, \pi(\boldsymbol{x}_i))$. For the collection of binary variables the likelihood has the form

$$\begin{aligned} L(\boldsymbol{\beta}) &= \prod_{i=1}^{N}\prod_{j=1}^{n_i} \pi(\boldsymbol{x}_i)^{y_{ij}}(1 - \pi(\boldsymbol{x}_i))^{1-y_{ij}} \\ &= \prod_{i=1}^{N} \pi(\boldsymbol{x}_i)^{y_i}(1 - \pi(\boldsymbol{x}_i))^{n_i-y_i}, \end{aligned} \tag{4.1}$$

where $y_i = y_{i1} + \cdots + y_{in_i}$ denotes the number of successes. The likelihood for the number of successes $y_i \sim B(n_i, \pi(\boldsymbol{x}_i))$ has the form

$$L_{bin}(\boldsymbol{\beta}) = \prod_{i=1}^{N} \left( \begin{array}{c} n_i \\ y_i \end{array} \right) \pi(\boldsymbol{x}_i)^{y_i}(1 - \pi(\boldsymbol{x}_i))^{n_i-y_i}.$$

The binary observations likelihood $L(\boldsymbol{\beta})$ and the binomial likelihood $L_{bin}$ differ in the binomial factor, which is irrelevant in maximization because it does not depend on $\boldsymbol{\beta}$. Therefore, the relevant part of the log-likelihood (omitting the constant) is

$$\begin{aligned} l(\boldsymbol{\beta}) = \log(L(\boldsymbol{\beta})) &= \sum_{i=1}^{N} y_i \log(\pi(\boldsymbol{x}_i)) + (n_i - y_i)\log(1 - \pi(\boldsymbol{x}_i)) \\ &= \sum_{i=1}^{N} n_i\{\bar{y}_i \log\left(\frac{\pi(\boldsymbol{x}_i)}{1 - \pi(\boldsymbol{x}_i)}\right) + \log(1 - \pi(\boldsymbol{x}_i))\}, \end{aligned} \tag{4.2}$$

where $\bar{y}_i = y_i/n_i$ denotes the relative frequency. The score function is given by $\boldsymbol{s}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \boldsymbol{x}_i(\partial h(\eta_i)/\partial\eta)(\bar{y}_i - \pi(\boldsymbol{x}_i))/\mathrm{var}(\bar{y}_i)$, and the Fisher matrix, which is needed for the standard errors, is

$$\boldsymbol{F}(\boldsymbol{\beta}) = E(-\partial l^2(\boldsymbol{\beta})/\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T) = \sum_{i=1}^{N} n_i \frac{(\partial h(\boldsymbol{x}_i^T\boldsymbol{\beta})/\partial\eta)^2}{h(\boldsymbol{x}_i^T\boldsymbol{\beta})(1 - h(\boldsymbol{x}_i^T\boldsymbol{\beta}))}\boldsymbol{x}_i\boldsymbol{x}_i^T.$$

For the logit model one obtains the score function $\boldsymbol{s}(\boldsymbol{\beta}) = \sum_{i=1}^{n} n_i\boldsymbol{x}_i(\bar{y}_i - \pi(\boldsymbol{x}_i))$ and the Fisher matrix

$$\boldsymbol{F}(\boldsymbol{\beta}) = \sum_{i=1}^{N} n_i h(\boldsymbol{x}_i^T\boldsymbol{\beta})(1 - h(\boldsymbol{x}_i^T\boldsymbol{\beta}))\boldsymbol{x}_i\boldsymbol{x}_i^T.$$

An approximation to the covariance of $\hat{\boldsymbol{\beta}}$ is again given by the inverse information matrix.

### Asymptotic Properties

Under regularity assumptions, ML estimators for binary regression models have several favorable asymptotic properties ($n \to \infty$, where $n = n_1 + \cdots + n_N$ in the binomial case). The ML estimator exists and is unique asymptotically; it is consistent and asymptotically normally distributed with $\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \boldsymbol{F}(\hat{\boldsymbol{\beta}})^{-1})$. Moreover, it is asymptotically efficient compared to a wide class of other estimates. General results specifying the necessary regularity conditions were given by Haberman (1977) and Fahrmeir and Kaufmann (1985).

### Existence of Maximum Likelihood Estimates

For a finite sample size it may happen that ML estimates do not exist. It is easy to construct data structures for which the estimates tend to infinity. For a univariate predictor, let the data be separated such that all data with a predictor below a fixed value $c$ have response 0 and above that value response 1. Then the best fit is obtained when the parameter for the predictor becomes infinitely large. The general case was investigated by Albert and Anderson (1984) and Santner and Duffy (1986). They call a dataset *completely separated* if there exists a vector $\boldsymbol{\theta}$ such that

$$\boldsymbol{x}_i^T\boldsymbol{\theta} > 0 \text{ if } y_i = 1 \text{ and } \boldsymbol{x}_i^T\boldsymbol{\theta} < 0 \text{ if } y_i = 0$$

hold for $i = 1, \ldots, n$, where $\boldsymbol{x}_i$ contains an intercept. It is called *quasi-completely separated* if there exists a vector $\boldsymbol{\theta}$ such that

$$\boldsymbol{x}_i^T\boldsymbol{\theta} \geq 0 \text{ if } y_i = 1 \text{ and } \boldsymbol{x}_i^T\boldsymbol{\theta} \leq 0 \text{ if } y_i = 0$$

holds for $i = 1, \ldots, n$. A dataset is said to have *overlap* if there is no complete separation and no quasi-complete separation. They show that the ML estimate exists if and only if the dataset has overlap. Thus, from a geometric point of view, ML estimates exist if there is no hyperplane that separates the 0 and 1 responses. Christmann and Rousseeuw (2001) showed how to compute the smallest number of observations that need to be removed to make the ML estimate non-existent. This number of observations is called the regression depth and measures the amount of separation between 0 and 1 responses. For literature on alternative estimators, in particular robust procedures, see Section 4.7.

## Estimation Conditioned on Predictor Values

ML estimation as considered in the previous section is conditional on the $\boldsymbol{x}$-values. Although both variables $y$ and $\boldsymbol{x}$ can be random, estimating the parameters of a model for $P(y = 1|\boldsymbol{x})$

does not depend on the marginal distribution of $\boldsymbol{x}$. Therefore, the ML estimates have the same form in cases where one has a total sample of $iid$ observations $(y_i, \boldsymbol{x}_i)$ or a sample of responses conditional on $\boldsymbol{x}$-values.

In applications one also finds samples conditional on the response. In such a stratified sample one observes $\boldsymbol{x}$-values given $y = 1$ and $\boldsymbol{x}$-values given $y = 0$. A common case is case-control studies in biomedicine, where $y = 1$ refers to cases and $y = 0$ are controls. In both populations the potential risk factors $\boldsymbol{x}$ are observed given the population. In econometrics, this type of sampling is often called *choice-based sampling*, referring to the sampling of characteristics of a choice maker given his or her choice was made.

The association between response $y$ and predictor $\boldsymbol{x}$ as captured in the logit model also can be inferred from samples that are conditional on responses. Let us consider the most simple case of one binary predictor. Therefore, one has $y \in \{0, 1\}$ and $x \in \{0, 1\}$. The coefficient $\beta$ in the logit model $\text{logit}(P(y = 1|x)) = \beta_0 + x\beta$ is given by $\beta = \log(\gamma)$, where $\gamma$ is the odds ratio, which contains the association between $y$ and $x$. However, the odds ratio $\gamma$ can be given in two forms:

$$\gamma = \frac{P(y = 1|x = 1)/P(y = 0|x = 1)}{P(y = 1|x = 0)/P(y = 0|x = 0)} = \frac{P(x = 1|y = 1)/P(x = 0|y = 1)}{P(x = 1|y = 0)/P(x = 0|y = 0)}.$$

The first form corresponds to the logit model $\text{logit}(P(y = 1|x)) = \beta_0 + x\beta$, the second form to the logit model $\text{logit}(P(x = 1|y)) = \tilde{\beta}_0 + y\tilde{\beta}$. For the latter model, which models response $x$ given $y$, the parameter that contains the association between these variables is the same, $\tilde{\beta} = \log(\gamma) = \beta$. Therefore, ML estimation of the latter model, based on a sample given $y$, yields an estimate of the coefficient $\beta$ of the original logit model $\text{logit}(P(y = 1|x)) = \beta_0 + x\beta$. Asymptotic properties hold for the transformed model $\text{logit}(P(x = 1|y)) = \tilde{\beta}_0 + y\tilde{\beta}$.

In general, the use of estimators for samples that are conditional on $y$ may be motivated by the specific structure of the logit model. Therefore, we go back to the derivation of the binary logit model to assume that predictors are normally distributed given $y = r$ (Section 2.2.2). With $f(\boldsymbol{x}|r)$ denoting the density given $y = r$ and $p(r) = P(y = r)$ denoting the marginal probability, it follows from Bayes' theorem that

$$P(y = 1|\boldsymbol{x}) = \frac{\exp\{\log([p(1)f(\boldsymbol{x}|1)]/[p(0)f(\boldsymbol{x}|0)])\}}{1 + \exp\{\log([p(1)f(\boldsymbol{x}|1)]/[p(0)f(\boldsymbol{x}|0)])\}}.$$

Therefore, the linear logit model holds if

$$\log \frac{p(1)f(\boldsymbol{x}|1)}{p(0)f(\boldsymbol{x}|0)} = \beta_0 + \boldsymbol{x}^T\boldsymbol{\beta},$$

holds. The equivalent form,

$$\log \frac{f(\boldsymbol{x}|1)}{f(\boldsymbol{x}|0)} = \beta_0 - \log(p(1)/p(0)) + \boldsymbol{x}^T\boldsymbol{\beta},$$

shows that a logit model holds if $\log(f(\boldsymbol{x}|1)/f(\boldsymbol{x}|0))$ has a linear form that contains the term $\boldsymbol{x}^T\boldsymbol{\beta}$ and only the intercept depends on the marginal probabilities. The essential point is that the marginals determine only the intercept. Thus, for identical densities but different marginal probabilities, the coefficient $\boldsymbol{\beta}$, which measures the association between $y$ and $\boldsymbol{x}$, is unchanged. The argument holds more generally when the linear term $\boldsymbol{x}^T\boldsymbol{\beta}$ is replaced by a function $\eta(\boldsymbol{x}^T, \boldsymbol{\beta})$; linearity in $\boldsymbol{x}$ is just the most prominent case. It is one of the strengths of the logit model that the parameter $\boldsymbol{\beta}$ does not depend on the marginal probabilities.

Nevertheless, the likelihood for a given $y$ differs from the likelihood given predictors. By using $f(\boldsymbol{x}_i|y_i) = P(y_i|\boldsymbol{x}_i)f(\boldsymbol{x}_i)/p(y_i)$, one obtains for the log-likelihood conditional on $y$

$$l_{cond} = \sum_{i=1}^{n} \log(P(y_i|\boldsymbol{x}_i)) + \log(f(\boldsymbol{x}_i)) - \log(p(y_i)).$$

The first term on the right-hand side is equivalent to the conditional log-likelihood given $\boldsymbol{x}$-values. The second term corresponds to the marginal distribution of $\boldsymbol{x}$, which can be maximized by the empirical distribution. The third term refers to the marginal distribution of $y$, which is fixed by the sampling and changes the intercept. Of course it has to be shown that maximization of $l_{cond}$, which includes non-parametric maximization of the marginal distribution of $\boldsymbol{x}$, yields coefficient estimates $\boldsymbol{\beta}$ that have the properties of ML estimates. For more details of choice-based sampling see Prentice and Pyke (1979), Carroll et al. (1995), and Scott and Wild (1986).

## 4.2 Discrepancy between Data and Fit

### 4.2.1 The Deviance

Before drawing inferences from a fitted model it is advisable to critically assess the fit. Thus, when fitting a regression model one usually wants some measure for the discrepancy between the fitted model and the observations. It should be obvious that the sum of the squared residuals $\sum_i (y_i - \hat{\pi}_i)^2$ that is used in normal regression models is inappropriate because it is designed for a symmetric distribution like the normal and in addition assumes homogeneous variances. It does not account for the specific type of distribution (and noise) when responses are binary. If the unknown parameters are estimated by maximum likelihood, a common measure for the discrepancy between the data and the fitted model that is specific for the underlying distribution of responses is the deviance. The deviance may be seen as a comparison between the data fit and the perfect fit.

The deviance is strongly related to basic concepts of statistics. A basic test statistic that is used to evaluate nested models is the *likelihood ratio statistic*:

$$\lambda = -2 \log \frac{L(\text{submodel})}{L(\text{model})},$$

where $L$(model) represents the maximal likelihood when a model is fit and $L$(submodel) represents the maximal likelihood if a more restrictive model, a so-called submodel, is fit. By considering the submodel as the binary regression model and the model as the most general possible model (with perfect fit), one considers

$$\begin{aligned} \lambda &= -2\{\log\, L(\text{fitted submodel}) - \log\, L(\text{fitted model})\} \\ &= -2\{l(\text{fitted submodel}) - l(\text{fitted model})\}. \end{aligned}$$

The most general model that produces a perfect fit is often called the *saturated model*. It is assumed to have as many parameters as observations: therefore it is the perfect fit.

#### Deviance for Binary Response

Suppose that the data are given by $(y_i, \boldsymbol{x}_i), i = 1, \ldots, n$, where $y_i \in \{0, 1\}$. Let $l(\boldsymbol{y}; \hat{\boldsymbol{\pi}})$ denote the log-likelihood for the fitted model with $\boldsymbol{y}^T = (y_1, \ldots, y_n)$, $\hat{\boldsymbol{\pi}}^T = (\hat{\pi}_1, \ldots, \hat{\pi}_n)$, $\hat{\pi}_i = \hat{\pi}(\boldsymbol{x}_i) = h(\boldsymbol{x}_i^T \hat{\boldsymbol{\beta}})$. The perfectly fitting model is represented by the likelihood $l(\boldsymbol{y}; \boldsymbol{y})$, where

the fitted values are the observations themselves. Then the *deviance* for the binary responses is given by the difference of $l(\boldsymbol{y}, \boldsymbol{y})$ and $l(\boldsymbol{y}, \hat{\boldsymbol{\pi}})$:

$$\begin{aligned} D(\boldsymbol{y}, \hat{\boldsymbol{\pi}}) &= 2\{l(\boldsymbol{y}, \boldsymbol{y}) - l(\boldsymbol{y}, \hat{\boldsymbol{\pi}})\} \\ &= 2\left\{\sum_{i=1}^{n} y_i \log\left(\frac{y_i}{\hat{\pi}_i}\right) + (1-y_i)\log(\frac{1-y_i}{1-\hat{\pi}_i})\right\} \\ &= -2\sum_{i=1}^{n}\{y_i \log(\hat{\pi}_i) + (1-y_i)\log(1-\hat{\pi}_i)\}, \end{aligned} \tag{4.3}$$

where the convention $0 \cdot \infty = 0$, is used. Since $l(\boldsymbol{y}, \boldsymbol{y}) = 0$, the deviance reduces to $D(\boldsymbol{y}, \hat{\boldsymbol{\pi}}) = -2l(\boldsymbol{y}, \hat{\boldsymbol{\pi}})$. An alternative form of the deviance is

$$D(\boldsymbol{y}, \hat{\boldsymbol{\pi}}) = 2\sum_{i=1}^{n} d(y_i, \hat{\pi}_i), \tag{4.4}$$

where

$$d(y_i, \hat{\pi}_i) = \begin{cases} -\log(\hat{\pi}_i) & y_i = 1 \\ -\log(1-\hat{\pi}_i) & y_i = 0. \end{cases}$$

The simpler form,

$$d(y_i, \hat{\pi}_i) = -\log(1 - |y_i - \hat{\pi}_i|),$$

shows that for binary data the deviance implicitly uses the difference between observations and fitted values. From the latter form it is also seen that $D(\boldsymbol{y}, \hat{\boldsymbol{\pi}}) \geq 0$ and $D(\boldsymbol{y}, \hat{\boldsymbol{\pi}}) = 0$ only if $\hat{\boldsymbol{\pi}} = \boldsymbol{y}$.

In extreme cases the deviance degenerates as a measure of goodness-of-fit. Consider the simple case of a single sample and $n$ independent Bernoulli variables. Let $n_1$ denote the number of "hits" $(y_i = 1)$ and $n_2 = n - n_1$ the number of observations $y_i = 0$. Since $\hat{\pi}_i$ is the same for all observations, one obtains with $\hat{\pi} = n_1/n$

$$\begin{aligned} D(\boldsymbol{y}, \hat{\boldsymbol{\pi}}) &= -2(n_1 \log(\hat{\pi}) \ + n_2 \log(1-\hat{\pi})) \\ &= -2n\{\hat{\pi}\log(\hat{\pi}) + (1-\hat{\pi})\log(1-\hat{\pi})\}, \end{aligned}$$

which is completely determined by the relative frequency $\hat{\pi}$ and therefore does not reflect the goodness-of-fit. The asymptotic distribution depends heavily on $\pi$ with a degenerate limit as $n \to \infty$.

Another concept that is related to the deviance is the Kullback-Leibler distance, which is a general directed measure for the distance between two distributions (Appendix D). For discrete distributions with support $\{z_1, \ldots, z_m\}$ and the vectors of probabilities for the two probability functions given by $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_m)^T$ and $\boldsymbol{\pi}^* = (\pi_1^*, \ldots, \pi_m^*)^T$, it has the form

$$KL(\boldsymbol{\pi}, \boldsymbol{\pi}^*) = \sum_{i=1}^{m} \pi_i \log(\frac{\pi_i}{\pi_i^*}).$$

Considering the data as a degenerate mass function $(1-y_i, y_i)$ with support $\{0, 1\}$, one obtains

$$D(\boldsymbol{y}, \hat{\boldsymbol{\pi}}) = 2\sum_{i=1}^{n} KL((1-y_i, y_i), (1-\hat{\pi}_i, \hat{\pi}_i)).$$

The maximum likelihood estimator may be motivated as a minimum distance estimator that minimizes the sums of the Kullback-Leibler distances considered in the last equation.

## Deviance for Proportions

Let the data be given in the form $(y_i, \boldsymbol{x}_i), i = 1, \ldots, N$, where the $y_i$'s are binomially distributed. Each response variable $y_i \sim B(n_i, \pi(\boldsymbol{x}_i))$ can be thought of as being composed from $n_i$ binary variables. Therefore one should distinguish between the total number of (binary) observations $n = n_1 + \cdots + n_N$ and the number of observed binomials. The model-specific probabilities $\pi_i = \pi(\boldsymbol{x}_i)$ depend only on the measurement points $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$; thus only $N$ potentially differing probabilities are specified.

Let $\bar{\boldsymbol{y}}^T = (\bar{y}_1, \ldots, \bar{y}_N)$ denote the observations where $\bar{y}_i = y_i/n_i$ represents the relative frequencies that are linked to $y_i = y_{i1} + \cdots + y_{in_i} \sim B(n_i, \pi_i)$, the number of successes for the repeated trials at measurement point $\boldsymbol{x}_i$. Then the corresponding difference of log-likelihoods based on (4.2) is given by

$$D(\bar{\boldsymbol{y}}, \hat{\boldsymbol{\pi}}) = 2(l(\bar{\boldsymbol{y}}, \bar{\boldsymbol{y}}) - l(\bar{\boldsymbol{y}}, \hat{\boldsymbol{\pi}})) = 2 \sum_{i=1}^{N} n_i \{\bar{y}_i \log(\frac{\bar{y}_i}{\hat{\pi}_i}) + (1 - \bar{y}_i) \log(\frac{1 - \bar{y}_i}{1 - \hat{\pi}_i})\}.$$

The essential difference between the deviance for single binary observations and the deviance for binomial distributions is in the definition of the saturated model. While in the first case observations $y_1, \ldots, y_n$ are fitted perfectly by the saturated model, in the latter case the means $\bar{y}_1, \ldots, \bar{y}_N$ are fitted perfectly. This has severe consequences when trying to use the deviance as a measure of the goodness-of-fit of the model.

## Deviance as Goodness-of-Fit Statistic

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, this does not hold for the binary variables deviance $D(\boldsymbol{y}, \hat{\boldsymbol{\pi}})$ if $n \to \infty$. The reason is that the degrees of freedom are not fixed; they increase with the sample size. Thus there is no benchmark (in the form of an approximate distribution) to which the absolute value of $D(\boldsymbol{y}, \hat{\boldsymbol{\mu}})$ may be compared. The case is different for binomial distributions. If $N$, the number of measurement points, is fixed, and $n_i \to \infty$ for $i = 1, \ldots, N$, then the degrees of freedom are fixed, and, if the model holds, $D(\bar{\boldsymbol{y}}, \hat{\boldsymbol{\pi}})$ is under weak conditions asymptotically $\chi^2$-distributed with $D(\bar{\boldsymbol{y}}, \hat{\boldsymbol{\pi}}) \sim^{(a)} \chi^2(N - dim(\boldsymbol{x}))$, where $dim(\boldsymbol{x})$ denotes the length of the vector $\boldsymbol{x}$ that is equivalent to the number of estimated parameters. Thus, for $n_i$ sufficiently large, the $\chi^2$-distribution provides a benchmark to which the value of $D(\bar{\boldsymbol{y}}, \hat{\boldsymbol{\pi}})$ may be compared. Usually the model is considered to show an unsatisfactory fit, if $D(\bar{\boldsymbol{y}}, \hat{\boldsymbol{\pi}})$ is larger than the $1 - \alpha$ quantile of the $\chi^2$-distribution for user-specified significance value $\alpha$.

Nevertheless, since program packages automatically give the deviance, it is tempting to compare the value of the deviance to the (often absurdly high) degrees of freedom that are found for ungrouped binary data (see Example 4.1). Although the deviance does not provide a goodness-of-fit statistic for simple binary observations, it is useful in residual analysis and informal comparisons of link functions (see Sections 4.3 and 5.1).

TABLE 4.1: Main effects model for unemployment data.

| | Estimate | Std. Error | $z$-value | Pr(>\|$z$\|) |
|---|---|---|---|---|
| Intercept | 1.2140 | 0.2032 | 5.97 | 0.0000 |
| Age | −0.0312 | 0.0060 | −5.18 | 0.0000 |
| Gender | 0.6710 | 0.1393 | 4.82 | 0.0000 |

Example 4.1: Unemployment

In a study on the duration of unemployment with sample size $n = 982$ we distinguish between short-term unemployment ($\leq 6$ months) and long-term unemployment ($> 6$ months). Short-term unemployment is considered as success ($y = 1$). The response and the covariates are gender (1: male; 0: female) and age ranging from 16 to 61 years of age. From the estimates of coefficients in Table 4.1 it is seen that the probability of short-term unemployment is larger for males than for females and decreases with age. The deviance for ungrouped data, where the response is a 0-1 variable, is 1224.1 on 979 $df$. That deviance cannot be considered as a goodness-of-fit statistic. However, if data are grouped with the response represented by the binomial distribution given a fixed combination of gender and age, the deviance is 87.16 on 88 $df$. For the grouped case, 92 combinations of gender and age effects are possible; $N = 91$ of these combinations are found in the data; therefore, the $df$ are $91 - 3$ since three parameters have been fitted. In the grouped case the deviance has an asymptotic distribution and one finds that the main effect model is acceptable. □

Example 4.2: Commodities in Household

In Table 4.2 the fit of two linear logit models is shown. In the first model the response is "car in household," and in the second model it is "personal computer (pc) in household." For both models the only covariate is net income and only one-person households are considered. Since the responses are strictly binary, the deviance cannot be considered as a goodness-of-fit statistic that is asymptotically $\chi^2$-distributed. Thus the values of the deviance cannot be compared to the degrees of freedom. But since the number of observations and the linear predictor are the same for both models, one might compare the deviance of the two models in an informal way. The linear logit model seems to fit the data better when the response "pc in household" is considered rather than "car in household." However, it is not evaluated whether the fit is significantly better. □

TABLE 4.2: Effects of net income on alternative responses for linear logit model.

| | $\hat{\beta}_0$ | $\hat{\beta}$ | Deviance | $df$ |
|---|---|---|---|---|
| Car | $-2.42$ | 0.0019 | 1497.7 | 1294 |
| PC | $-3.88$ | 0.0011 | 614.8 | 1294 |

## 4.2.2 Pearson Statistic

When proportions are considered and the $n_i$ is large, the deviance yields a goodness-of-fit statistic. An alternative statistic in this setting is the *Pearson statistic*:

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

where $\hat{\pi}_i = h(\boldsymbol{x}_i^T \hat{\boldsymbol{\beta}})$. If $N$ is fixed and $n_i \to \infty$ for $i = 1, \ldots, N$, then $\chi^2_P$ is also asymptotically $\chi^2$-distributed with $N - dim(\boldsymbol{x})$ degrees of freedom given the model holds.

The Pearson statistic is familiar from introductory texts on statistics. In the analysis of contingency tables it usually takes the form

$$\chi^2 = \sum_{\text{cells}} \frac{(\text{observed cell counts} - \text{expected cell counts})^2}{\text{expected cell counts}}, \tag{4.5}$$

where the expected cell counts are computed under the assumption that the model under investigation holds. Thus one considers the discrepancy between the actual counts and what is to be expected if the model holds. The denominator is a weight on the squared differences that takes differences less serious if the expected cell counts are large. In fact, this weighting scheme is a standardization to obtain the asymptotic $\chi^2$-distribution. This is seen from showing that $\chi^2_P$ has the form (4.5). The fitting of the binary regression model $\pi_i = h(\boldsymbol{x}_i^T\boldsymbol{\beta})$ may be seen within the framework of contingency analysis with $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$ refering to the rows and $y = 0, y = 1$ refering to the columns. Then the probabilities within one row (corresponding to $\boldsymbol{x}_i$) are given by $\pi_i, 1 - \pi_i$; the cell counts are given by $n_i\bar{y}_i, n_i - n_i\bar{y}_i$; and the (estimated) expected cell counts are given by $n_i\hat{\pi}_i, n_i - n_i\hat{\pi}_i$. From (4.5) one obtains

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

which, after simple derivation, turns out to be equivalent to $\chi^2_P$. The essential difference between the representation (4.5) and $\chi^2_P$ is that in the form (4.5) the sum is across all cells of the corresponding contingency table, while in $\chi^2_P$ the sum is only across the different values of $\boldsymbol{x}_i$, which correspond to rows.

The deviance and Pearson's $\chi^2$ have the same asymptotic distribution. More concisely, one postulates for fixed $N$ that $n = n_1 + \cdots + n_N \to \infty$ and $n_i/N \to \lambda_i$, where $\lambda_i \in (0, 1)$. To distinguish these assumptions from the common asymptotical conditions where only $n \to \infty$ is assumed, one uses the term *fixed cells asymptotics*, referring to the assumption that the number of rows $N$ is fixed. If the values of the deviance $D$ and $\chi^2_P$ differ strongly, this may be taken as a hint that the requirement of fixed cells asymptotics does not hold and both test statistics are not reliable (see also Section 8.6.2, where alternative asymptotic concepts that hold for the more general power-divergence family of goodness-of-fit statistics are briefly discussed).

**Goodness-of-Fit Statistics for Proportions**

Pearson Statistic

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

Deviance

$$D = 2\sum_{i=1}^{N} n_i \left\{ \bar{y}_i \log\left(\frac{\bar{y}_i}{\hat{\pi}_i}\right) + (1 - \bar{y}_i) \log\left(\frac{1 - \bar{y}_i}{1 - \hat{\pi}_i}\right) \right\}$$

Approximation $(n_i/N \to \lambda_i)$

$$\chi^2_P, D \overset{(a)}{\sim} \chi^2(N - p)$$

A general problem with global or omnibus tests to assess the fit of a parametric model is that large values of the test statistic indicate lack-of-fit but do not show the particular reason for the lack-of-fit. It is only by comparing fits that one may get some insight about its nature. In Section 4.4 nested models are compared that differ in the predictor. In that case the difference of deviances has a fixed asymptotic distribution. In the case of different link functions, considered in Section 5.1, the models are not nested, but a comparison of the goodness-of-fit measures may show which link function has the best fit.

### 4.2.3 Goodness-of-Fit for Continuous Predictors

The deviance and Pearson's $\chi^2_P$ may be used to assess the adequacy of the model when the response is binomial with not too small cell counts. They cannot be used when predictors are continuous since then cell counts would always be one and there would be no benchmark in the form of a $\chi^2$-statistic available (see also Example 4.2). An easy way out seems to be to categorize all of the continuous variables. However, then one looses power and it works only in the case of one or two continuous variables. When more continuous variables are in the model it is not clear how to find categories that contain enough observations. Several methods have been proposed to assess the fit of a model when continuous variables are present. We briefly consider some of them in the following.

#### Hosmer-Lemeshow Test

Hosmer and Lemeshow (1980) proposed a Pearson-type test statistic where the categorizing is based on the fitted values. One orders the responses according to the fitted probabilities and then forms $N$ equally sized groups. Thus the quantiles of the response values determine the groups. More precisely, the $n/N$ observations with the smallest fitted probabilities form the first group, the next $n/N$ observations form the second group, and so on. Hosmer and Lemeshow (1980) proposed to use $N = 10$ groups that are called "deciles of risk."

Let $y_{ij}$ denote the $j$th observation in the $i$th group, where $i = 1, \ldots, N, j = 1, \ldots, n_i$. Then the average of the observations of the $i$th group, $\bar{y}_i = \sum_{j=1}^{n_i} y_{ij}/n_i$, is compared to the average of the fitted probabilities, $\hat{\pi}_i = \sum_{j=1}^{n_i} \hat{\pi}_{ij}/n_i$, where $\hat{\pi}_{ij}$ denotes the fitted value for observation $y_{ij}$ in a Pearson-type statistic:

$$\chi^2_{HL} = \sum_{i=1}^{N} n_i \frac{(\bar{y}_i - \hat{\pi}_i)^2}{\hat{\pi}_i(1 - \hat{\pi}_i)}.$$

The test statistic has a rather complicated distribution, but Hosmer and Lemeshow (1980) showed by simulations that the asymptotic distribution can be approximated by a $\chi^2$-distribution with $df = N - 2$.

Since the grouping is based on the model that is assumed to hold, one cannot expect good power for the Hosmer-Lemeshow statistic, and indeed the test statistic has moderate power. More importantly, as for all global tests, a large value of the test statistic indicates lack-of-fit but does not show if, for example, the linear term has been misspecified.

A methodology that is similar to the Hosmer-Lemeshow approach but makes use of categorical variables in the model has been proposed by Pulkstenis and Robinson (2002). An adjusted Hosmer-Lemeshow test that uses an alternative standardization was proposed by Pigeon and Heyse (1999). Kuss (2002) showed that Hosmer-Lemeshow-type tests are numerically unstable.

#### Alternative Tests

Alternative test strategies may be based on the score test. Brown (1982) embedded the logit model into a family of models and proposed a test for the specific parameters that select the logit model. Tsiatis (1980) divided the space of the covariate into distinct regions and tested if a term that is constant in these regions may be omitted from the linear predictor.

An alternative approach for goodness-of-fit tests uses non-parametric regression techniques. Azzalini et al. (1989) proposed a pseudo-likelihood ratio test by using a kernel-smoothed estimate of the response probabilities. For binomial data, $y_i \sim B(n_i, \pi(x_i))$, a smooth

nonparametric estimate of the response probability (for a unidimensional predictor) is

$$\hat{\pi}(x) = \sum_{i=1}^{N} y_i K(\frac{x - x_i}{h}) / \sum_{i=1}^{N} n_i K(\frac{x - x_i}{h}).$$

The hypotheses to be investigated are

$$\begin{aligned} H_0: &\quad \pi(x) = \pi(x, \boldsymbol{\beta}) \text{ specified by a parametric model} \\ H_1: &\quad \pi(x) \text{ is a smooth function.} \end{aligned}$$

The corresponding pseudo-likelihood ratio statistic is

$$\sum_{i=1}^{N} y_i \log(\frac{\hat{\pi}(x_i)}{\pi(x, \hat{\boldsymbol{\beta}})}) + (n_i - y_i) \log(\frac{1 - \hat{\pi}(x_i)}{1 - \pi(x, \hat{\boldsymbol{\beta}})}),$$

where $\hat{\pi}(x)$ is the smoothed estimate based on a selected smoothing parameter $h$. Since the test statistic is not asymptotically $\chi^2$, the null hypothesis behavior of the test statistic is examined by simulating data from the fitted parametric model.

Rather than estimating the probability by smoothing techniques, one may also use smooth estimates of the standardized residuals. LeCessie and van Houwelingen (1991) smooth the (ungrouped) Pearson residuals $r(x_i) = r_P(y_i, \hat{\pi}_i)$ (see next section) to obtain $\tilde{r}(x) = \sum_i r(x_i) K(\frac{x-x_i}{h}) / \sum_i K(\frac{x-x_i}{h})$. The test statistic has the form $T = \sum_i \tilde{r}(x_i) v(x_i)$, where $v(x_i) = \sum_i \{K(x - x_i)/h)\}^2 / \sum_i K((x - x_i)/h)^2$ is the inverse of the variance of the smoothed residuals. LeCessie and van Houwelingen (1991) derive the mean and the variance and demonstrate that the test statistic may be approximated by a normal distribution. The equivalence to a score test in a random effects model was shown by LeCessie and van Houwelingen (1995). The approaches may be extended to multivariate predictors by using multivariate kernels but are restricted to few dimensions since kernel-based estimates for higher dimensions suffer from the curse of dimensionality (see Section 10.1.4).

## 4.3 Diagnostic Checks

Goodness-of-fit tests provide only global measures of the fit of a model. They tell nothing about the reason for a bad fit. Regression diagnostics aims at identifying reasons for it. Diagnostic measures should in particular identify observations that are not well explained by the model as well as those that are influential for some aspect of the fit.

### 4.3.1 Residuals

The goodness-of-fit statistics from the preceding section provide global measures for the discrepancy between the data and the fit. In particular, if the fit is bad, one wants to know if all the observations contribute to the lack-of-fit or if the effect is due to just some observations. Residuals measure the agreement between single observations and their fitted values and help to identify poorly fitting observations that may have a strong impact on the overall fit of the model. In the following we consider responses from a binomial distribution.

For scaled binomial data the *Pearson residual* has the form

$$r_P(\bar{y}_i, \hat{\pi}_i) = \frac{\bar{y}_i - \hat{\pi}_i}{\sqrt{\hat{\pi}_i(1 - \hat{\pi}_i)/n_i}}.$$

It is just the raw residual scaled by the estimated standard deviation of $\bar{y}_i$. It is also the signed square root of the corresponding component of the Pearson statistics that has the form $\chi^2_P = \sum_i r_P(\bar{y}_i, \hat{\pi}_i)^2$. For small $n_i$ the distribution of $r_P(\bar{y}_i, \hat{\pi}_i)$ is rather skewed, an effect that is ameliorated by using the transformation to *Anscombe* residuals:

$$r_A(\bar{y}_i, \hat{\pi}_i) = \sqrt{n_i}\frac{t(\bar{y}_i) - [t(\hat{\pi}_i) + (\hat{\pi}_i(1-\hat{\pi}_i))^{-1/3}(2\hat{\pi}_i - 1)/6n_i]}{(\hat{\pi}_i(1-\hat{\pi}_i))^{1/6}},$$

where $t(u) = \int_0^u s^{-1/3}(1-s)^{-1/3}ds$ (see Pierce and Schafer, 1986). Ansombe residuals consider an approximation to

$$\frac{t(\bar{y}_i) - E(t(\bar{y}_i))}{\sqrt{var(t(\bar{y}_i))}}$$

by use of the delta method, which yields

$$\frac{t(\bar{y}_i) - t(E(t(\bar{y}_i)))}{t'(E(\bar{y}_i))\sqrt{var(\bar{y}_i)}}.$$

The Pearson residual cannot be expected to have unit variance because the variance of the residual has not been taken into account. The standardization in $r_P(\bar{y}_i; \hat{\pi}_i)$ just uses the estimated standard deviation of $\bar{y}_i$. As shown in Section 3.10, the variance of the residual vector may be approximated by $\boldsymbol{\Sigma}^{1/2}(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{\Sigma}^{T/2}$, where the hat matrix is given by $\boldsymbol{H} = \boldsymbol{W}^{T/2}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{W}^{1/2}$. Therefore, when looking for ill-fitting observations, one prefers the *standardized Pearson residuals*:

$$r_{P,s}(\bar{y}_i, \hat{\pi}_i) = \frac{\bar{y}_i - \hat{\pi}_i}{\sqrt{(1-h_{ii})\hat{\pi}_i(1-\hat{\pi}_i)/n_i}},$$

where $h_{ii}$ denotes the $i$th diagonal element of $\boldsymbol{H}$. The standardized Pearson residuals are simply the Pearson residuals divided by $\sqrt{1-h_{ii}}$.

Alternative residuals derive from the deviance. From the basic form of the deviance $D = \sum_i r_D(y_i, \hat{\pi}_i)^2$ one obtains

$$r_D(\bar{y}_i, \hat{\pi}_i) = sign(\bar{y}_i - \hat{\pi}_i)\sqrt{n_i\{\bar{y}_i \log\left(\frac{\bar{y}_i}{\hat{\pi}_i}\right) + (1-\bar{y}_i)\log\left(\frac{1-\bar{y}_i}{1-\hat{\pi}_i}\right)\}},$$

where $sign(\bar{y}_i - \hat{\pi}_i)$ is 1 when $\bar{y}_i \geq \hat{\pi}_i$ and is $-1$ when $\bar{y}_i < \hat{\pi}_i$. For the special case $n_i = 1$ it simplifies to

$$r_D(y_i, \hat{\pi}_i) = sign(y_i - \hat{\pi}_i)\sqrt{-\log(1 - |y_i - \hat{\pi}_i|)}.$$

A transformation that yields a better approximation to the normal distribution is the *adjusted residuals*:

$$r_D\hat{a}(\bar{y}_i, \hat{\pi}_i) = r_D(\bar{y}_i, \hat{\pi}_i) + (1 - 2\hat{\pi}_i)/\sqrt{n_i\hat{\pi}_i(1-\hat{\pi}_i) * 36}.$$

*Standardized deviance residuals* are obtained by dividing by $\sqrt{1-h_{ii}}$ :

$$r_{D,s}(\bar{y}_i, \hat{\pi}_i) = \frac{r_D(\bar{y}_i, \hat{\pi}_i)}{\sqrt{1-h_{ii}}}.$$

Typically residuals are visualized in a graph. In an index plot the residuals are plotted against the observation number, or index. It shows which observations have large values and may be considered outliers. For finding systematic deviations from the model it is often more informative to plot the residuals against the fitted linear predictor. If one suspects that particular

variables should be transformed before being included in the linear predictor, one may also plot residuals against the ordered fitted values of that explanatory variable.

An alternative graph compares the standardized residuals to the order statistic of an $N(0, 1)$-sample. In this plot the residuals are ordered and plotted against the corresponding quantiles of a normal distribution. If the model is correct and residuals can be expected to be approximately normally distributed (depending on local sample size), the plot should show approximately a straight line as long as outliers are absent.

### Example 4.3: Unemployment

In a study on the duration of unemployment with sample size $n = 982$ we distinguish between short-term unemployment ($\leq 6$ months) and long-term unemployment ($> 6$ months). For illustration, a linear logit model is fitted with the covariate age, ranging from 16 to 61 years of age. Figure 4.2 shows the fitted response function. It is seen that in particular for older unemployed persons, the fitted values tend to be larger than the observed proportions. The effect is also seen in the plot of residuals against fitted values in Figure 4.3. The quantile plot in Figure 4.3 shows a rather straight line but the slope is rather steep, indicating that the variability of deviances differs from that of a standard normal distribution. A less restrictive nonparametric fit of the data is considered in Example 10.1. □

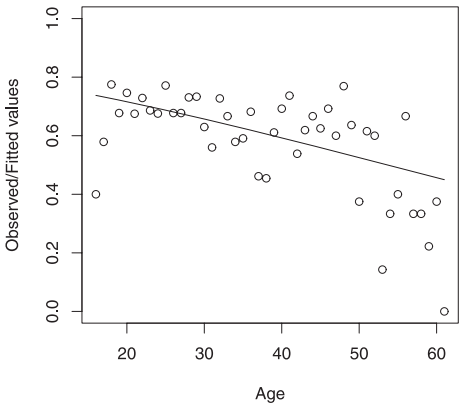


FIGURE 4.2: Observations and fitted probabilities for unemployment data plotted against age.

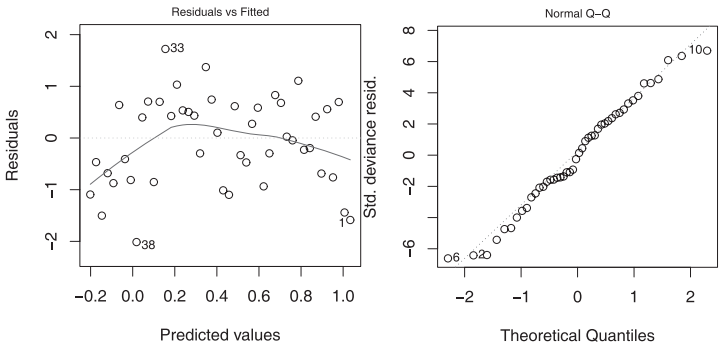


FIGURE 4.3: Deviance residuals for unemployment data plotted against fitted values (left) and quantile plot (right).

TABLE 4.3: Short- and long-term unemployment depending on age

| Observ. | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Age | | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 25 | 26 | 27 | 28 | 29 | 30 | 31 |
| | 1 | 2 | 11 | 31 | 42 | 50 | 54 | 43 | 35 | 25 | 27 | 21 | 21 | 19 | 22 | 17 | 14 |
| | 0 | 3 | 8 | 9 | 20 | 17 | 26 | 16 | 16 | 12 | 8 | 10 | 10 | 7 | 8 | 10 | 11 |
| Observ. | | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 25 | 26 | 27 | 28 | 29 | 30 | 31 | 32 |
| Age | | 32 | 33 | 34 | 35 | 36 | 37 | 38 | 39 | 40 | 41 | 42 | 43 | 44 | 45 | 46 | 47 |
| | 1 | 8 | 14 | 11 | 13 | 15 | 6 | 5 | 11 | 9 | 14 | 7 | 13 | 4 | 10 | 9 | 9 |
| | 0 | 3 | 7 | 8 | 9 | 7 | 7 | 6 | 7 | 4 | 5 | 6 | 8 | 2 | 6 | 4 | 6 |
| Observ. | | 33 | 34 | 35 | 36 | 37 | 38 | 39 | 40 | 41 | 42 | 43 | 44 | 45 | 46 | | |
| Age | | 48 | 49 | 50 | 51 | 52 | 53 | 54 | 55 | 56 | 57 | 58 | 59 | 60 | 61 | | |
| | 1 | 10 | 7 | 3 | 8 | 3 | 1 | 2 | 2 | 2 | 3 | 2 | 2 | 3 | 0 | | |
| | 0 | 3 | 4 | 5 | 5 | 2 | 6 | 4 | 3 | 1 | 6 | 4 | 7 | 5 | 1 | | |

### Example 4.4: Food-Stamp Data

Data that have often been used in diagnostics of binary data are the food-stamp data from Künsch et al. (1989), which consist of $n = 150$ persons, 24 of whom participated in the federal food-stamp program. The response indicates participation, and the predictor variables represent the dichotomous variables tenancy (TEN), supplemental income (SUP), as well as the log-transformation of the monthly income, $\log(\text{monthly income} + 1)$ (LMI). Künsch et al. (1989) show that two values are poorly accounted for by the logistic model and are most influential for the ML fit. If these two observations and four other observations are left out, the data have no overlap and the ML estimate does not exist. Figure 4.4 shows quantile plots for Pearson and Anscombe residuals. It is seen that Anscombe residuals are much closer to the normal distribution than Pearson residuals. □

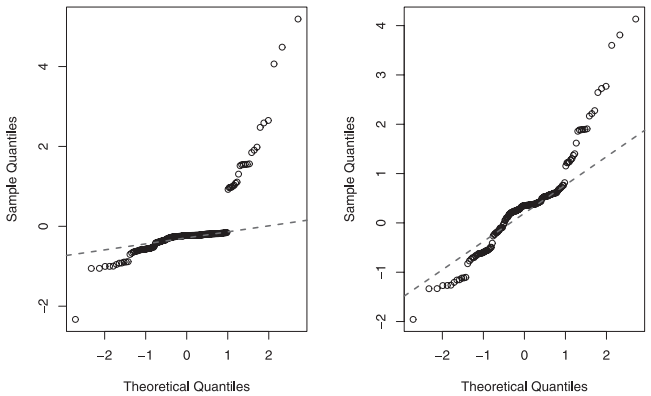


FIGURE 4.4: Pearson (left) and Anscombe (right) residuals for food-stamp data.

## 4.3.2 Hat Matrix and Influential Observations

The iteratively reweighted least-squares fitting (see Section 3.9) that can be used to compute the ML estimate has the form

$$\hat{\boldsymbol{\beta}}^{(k+1)} = (\boldsymbol{X}^T \boldsymbol{W}(\hat{\boldsymbol{\beta}}^{(k)})\boldsymbol{X})^{-1}\boldsymbol{X}^T \boldsymbol{W}(\hat{\boldsymbol{\beta}}^{(k)})\tilde{\boldsymbol{\eta}}(\hat{\boldsymbol{\beta}}^{(k)})$$

with the adjusted variable $\tilde{\boldsymbol{\eta}}(\hat{\boldsymbol{\beta}}) = \hat{\boldsymbol{\eta}} + \boldsymbol{D}(\hat{\boldsymbol{\beta}})^{-1}(\boldsymbol{y} - \boldsymbol{\mu}(\hat{\boldsymbol{\beta}}))$, where $\hat{\boldsymbol{\eta}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$ and the diagonal matrices $\boldsymbol{W} = \text{Diag}\,((\partial h(\hat{\eta}_1)/\eta)^2/\sigma_1^2, \ldots, (\partial h(\hat{\eta}_n)/\partial\eta)^2/\sigma_n^2)$ and $\boldsymbol{D}(\hat{\boldsymbol{\beta}}) = (\partial h(\hat{\eta}_1)/\partial\eta, \ldots, \partial h(\hat{\eta}_n)/\partial\eta)$. At convergence one obtains

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{W}(\hat{\boldsymbol{\beta}})\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{W}(\hat{\boldsymbol{\beta}})\tilde{\boldsymbol{\eta}}(\hat{\boldsymbol{\beta}}).$$

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

$$\boldsymbol{W}^{1/2}\tilde{\boldsymbol{\eta}}(\hat{\boldsymbol{\beta}}) = \boldsymbol{W}^{1/2}\boldsymbol{X}\boldsymbol{\beta} + \tilde{\boldsymbol{\varepsilon}},$$

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

$$\boldsymbol{H} = \boldsymbol{W}^{1/2}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{W}^{1/2}.$$

Since the matrix $\boldsymbol{H}$ is idempotent and symmetric, it may be seen as a projection matrix for which $\text{tr}(\boldsymbol{H}) = \text{rank}(\boldsymbol{H})$ holds. Moreover, one obtains for the diagonal elements of $\boldsymbol{H} = (h_{ij})$ $0 \leq h_{ij} \leq 1$ and $\text{tr}(\boldsymbol{H}) = p$ (if $\boldsymbol{X}$ has full rank).

The equation $\boldsymbol{W}^{1/2}\hat{\boldsymbol{\eta}} = \boldsymbol{H}\boldsymbol{W}^{1/2}\tilde{\boldsymbol{\eta}}(\boldsymbol{\beta})$ shows how the hat matrix maps the adjusted variable $\tilde{\boldsymbol{\eta}}(\boldsymbol{\beta})$ into the fitted values $\hat{\boldsymbol{\eta}}$. Thus $\boldsymbol{H}$ may be seen as the matrix that maps the adjusted observation vector $\boldsymbol{W}^{1/2}\tilde{\boldsymbol{\eta}}$ into the vector of "fitted" values $\boldsymbol{W}^{1/2}\hat{\boldsymbol{\eta}}$, which is a mapping on the transformed predictor space. Moreover, it may be shown that approximately

$$\boldsymbol{\Sigma}^{-1/2}(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}) \simeq \boldsymbol{H}\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{y} - \boldsymbol{\mu}) \tag{4.6}$$

holds, where $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}})$. Thus $\boldsymbol{H} = (h_{ij})$ may be seen as a measure of the influence of $\boldsymbol{y}$ on $\hat{\boldsymbol{\mu}}$ in standardized units of changes.

In summary, large values of $h_{ii}$ should be useful in detecting influential observations. But it should be noted that, in contrast to the normal regression model, the hat matrix depends on $\hat{\boldsymbol{\beta}}$ because $\boldsymbol{W} = \boldsymbol{W}(\hat{\boldsymbol{\beta}})$. The essential difference from an ordinary linear regression is that the hat matrix does depend not only on the design but also on the fit.

A further property that is not too hard to derive is

$$\boldsymbol{H}\boldsymbol{\Sigma}^{-1/2}\hat{\boldsymbol{\mu}} = \boldsymbol{H}\boldsymbol{\Sigma}^{-1/2}\boldsymbol{y},$$

meaning that the orthogonal projection (based on $\boldsymbol{H}$) of standardized values $\boldsymbol{\Sigma}^{-1/2}\hat{\boldsymbol{\mu}}$, $\boldsymbol{\Sigma}^{-1/2}\boldsymbol{y}$ are identical. With $\boldsymbol{\chi} = \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{y} - \hat{\boldsymbol{\mu}})$ denoting the standardized residual, one has

$$\boldsymbol{H}\boldsymbol{\chi} = \boldsymbol{0} \text{ and } (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{\chi} = \boldsymbol{\chi}.$$

There is a strong connection to the Pearson $\chi_P^2$ statistic since $\chi_P^2 = \boldsymbol{\chi}^T\boldsymbol{\chi}$. The matrix $\boldsymbol{H}$ has the form $\boldsymbol{H} = (\boldsymbol{W}^{T/2}\boldsymbol{X})(\boldsymbol{X}^T\boldsymbol{W}^{1/2}\boldsymbol{W}^{T/2}\boldsymbol{X})^{-1}(\boldsymbol{X}^T\boldsymbol{W}^{1/2})$, which shows that the projection is into the subspace that is spanned by the columns of $\boldsymbol{X}^T\boldsymbol{W}^{1/2}$.

### 4.3.3 Case Deletion

A strategy to investigate the effect of single observations on the parameter estimates is to compare the estimate $\hat{\beta}$ with the estimate $\hat{\boldsymbol{\beta}}_{(i)}$, obtained from fitting the model to the data without the $i$th observation. An overall measure that includes all the components of the vector of coefficients is due to Cook (1977). *Cook's distance* for observation $i$ has the form

$$c_i = (\hat{\boldsymbol{\beta}}_{(i)} - \hat{\boldsymbol{\beta}})^T \,\text{cov}(\hat{\boldsymbol{\beta}})^{-1}(\hat{\boldsymbol{\beta}}_{(i)} - \hat{\boldsymbol{\beta}}) = (\hat{\boldsymbol{\beta}}_{(i)} - \hat{\boldsymbol{\beta}})^T\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X}(\hat{\boldsymbol{\beta}}_{(i)} - \hat{\boldsymbol{\beta}}).$$

It may be seen as a confidence interval displacement diagnostic resulting from the exclusion of the $i$th observation. It is derived from the asymptotic confidence region for $\hat{\boldsymbol{\beta}}$ given by the likelihood distance $-2\{l(\boldsymbol{\beta}) - l(\hat{\boldsymbol{\beta}})\} = c$. By approximation of $l(\boldsymbol{\beta})$ by a second-order Taylor approximation at $\hat{\boldsymbol{\beta}}$ one obtains

$$c_i = (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^T \operatorname{cov}(\hat{\boldsymbol{\beta}})^{-1}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}).$$

Cook's distance is obtained by replacing $\boldsymbol{\beta}$ by $\hat{\boldsymbol{\beta}}_{(i)}$ and using an estimate of $\operatorname{cov}(\hat{\boldsymbol{\beta}})^{-1}$ that is composed from previously defined elements.

Computation of $\hat{\boldsymbol{\beta}}_{(i)}$ requires an iterative procedure for each observation. This may be avoided by using approximations. One may approximate $\hat{\boldsymbol{\beta}}_{(i)}$ by a one-step estimate, which is obtained by performing one Fisher scoring step starting from $\hat{\boldsymbol{\beta}}$. The corresponding approximation has the form

$$c_{i,1} = h_{ii} r_{P,i}^2/(1 - h_{ii})^2,$$

where $r_{P,i} = r_P(\bar{y}_i, \hat{\pi}_i)$ is the Pearson residual for the $i$th observation and $h_{ii}$ is the $i$th diagonal element of the hat matrix $\boldsymbol{H}$. The approximation is based on the one-step approximation of $\hat{\boldsymbol{\beta}}_{(i)}$:

$$\hat{\boldsymbol{\beta}}_{(i)} \approx \hat{\boldsymbol{\beta}} - w_{ii}^{1/2} r_{P,i}(\boldsymbol{X}^T \boldsymbol{W} \boldsymbol{X})^{-1} \boldsymbol{x}_i/(1 - h_{ii}),$$

where $\boldsymbol{x}_i$ is the covariate vector of observation $i$ and $w_{ii}$ is the $i$th diagonal element of $\boldsymbol{W}$.

Large values of Cook's measure $c_i$ or its approximation indicate that the $i$th observation is influential. Its presence determines the value of the parameter vector. A useful way of presenting this measure of influence is in an index plot.

### Example 4.5: Unemployment

Cook's distances for unemployment data (Figure 4.5) show that observations $33, 38, 44$, which correspond to ages $48, 53, 59$, are influential. All three observations are rather far from the fit. □

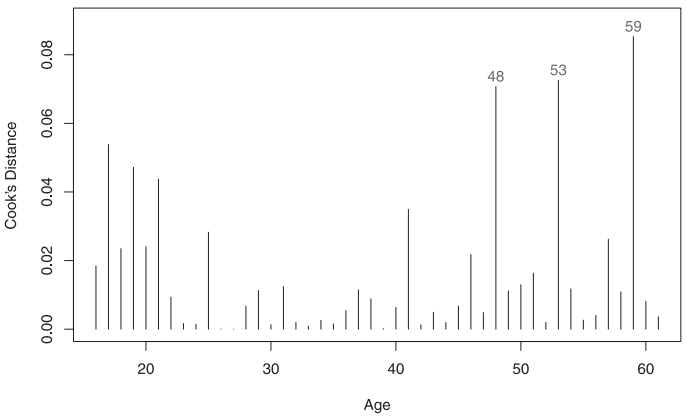


FIGURE 4.5: Cook distance for unemployment data.

### Example 4.6: Exposure to Dust (Non-Smokers)

In a study on the effects of dust on bronchitis conducted in a German plant, the observed covariates were mean dust concentration at working place in $mg/m^3$ (dust), duration of exposure in years (years), and

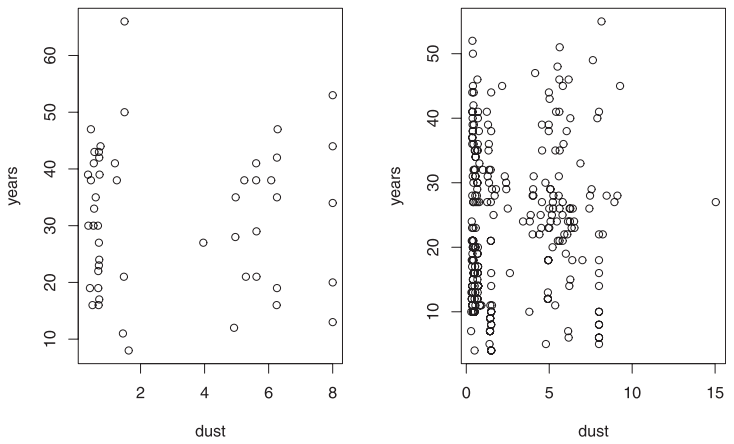


FIGURE 4.6: Dust exposure data, non-smokers with bronchitis (left panel), non-smokers without bronchitis (right panel).

smoking (1: yes; 0: no). Bronchitis is considered the binary response (1: present; 0: not present). The total sample was $n = 1246$. In previous analysis the focus has been on the estimation of threshold limiting values (Ulm, 1991; Küchenhoff and Ulm, 1997).

In the following we will first consider the subsample of non-smokers. Figure 4.6 shows the observations for workers with bronchitis and without bronchitis. It is seen that there is one person with an extreme value in the observation space. Since interaction terms and quadratic terms turn out not to contribute significantly to the model fit, in the following the main effect logit model is used. Table 4.4 shows the estimated coefficient for the main effects model. While years of exposure is highly significant, dust concentration is not significant in the subsample of non-smokers. Figure 4.7 shows Cook's distances for the subsample. Observations that show large values of Cook's distance are observations $730, 1175, 1210$, which have values $(1.63, 8), (8, 32), (8, 13)$ for (dust, years); all three observations correspond to persons with bronchitis. The observations are not extreme in the range of years, which is the influential variable. The one observation $(15.04, 27)$, which is very extreme in the observation space, corresponds to observation $1245$, which is the last in the plot of Cook's distances and has a rather small value of Cook's distance. The fit without that observation, as given in Table 4.5, shows that the coefficient for concentration of dust

TABLE 4.4: Main effects model for dust exposure data (non-smokers).

| | Estimate | Std. Error | $z$-Value | Pr($>\|z\|$) |
|---|---|---|---|---|
| Intercept | $-3.1570$ | 0.4415 | $-7.15$ | 0.0000 |
| Dust | 0.0053 | 0.0564 | 0.09 | 0.9248 |
| Years | 0.0532 | 0.0132 | 4.04 | 0.0001 |

TABLE 4.5: Main effects model for dust exposure data without one observation (non-smokers).

| | Estimate | Std. Error | $z$-Value | Pr($>\|z\|$) |
|---|---|---|---|---|
| Intercept | $-3.1658$ | 0.4419 | $-7.16$ | 0.0000 |
| Dust | 0.0120 | 0.0580 | 0.21 | 0.8361 |
| Years | 0.0529 | 0.0131 | 4.03 | 0.0001 |

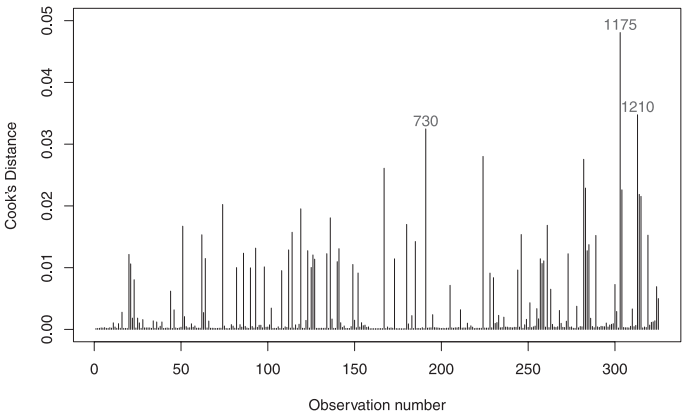


FIGURE 4.7: Cook distances for dust data (non-smokers).

has distinctly changed. However, the variable dust shows no significant effect, and therefore it is only a consequence that the Cook's distance is small. □

### Example 4.7: Exposure to Dust

In the following the exposure data, including non-smokers, is used. It turns out that for the full dataset concentration of dust, years of exposure, and smoking are significantly influential (see Table 4.6). Again interaction effects between the covariates can be omitted. Figure 4.8 shows the observation space for all

TABLE 4.6: Main effects model for dust exposure data.

| | Estimate | Std. Error | $z$-Value | Pr($>\|z\|$) |
|---|---|---|---|---|
| (Intercept) | $-3.0479$ | 0.2486 | $-12.26$ | 0.0000 |
| Dust | 0.0919 | 0.0232 | 3.95 | 0.0001 |
| Years | 0.0402 | 0.0062 | 6.47 | 0.0000 |
| smoking | 0.6768 | 0.1744 | 3.88 | 0.0001 |

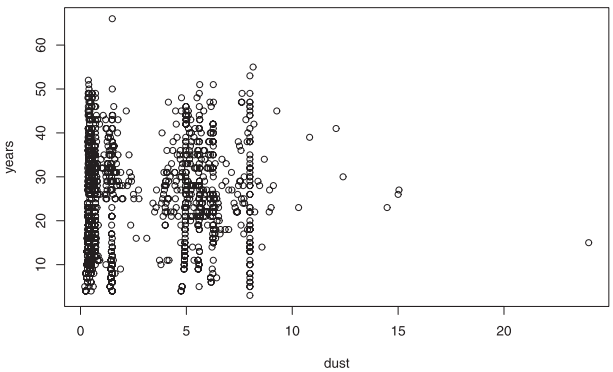


FIGURE 4.8: Dust exposure data.

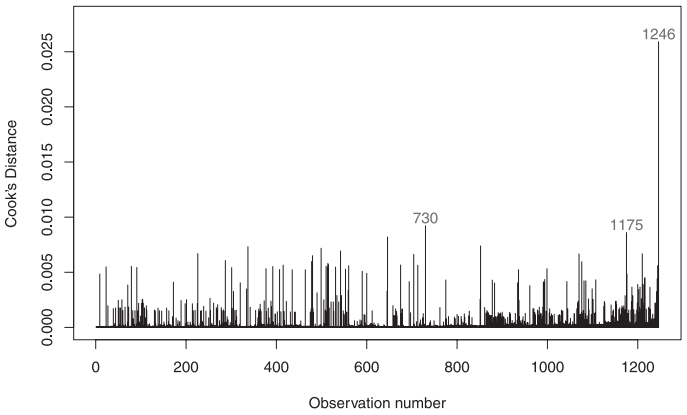


FIGURE 4.9: Cook distances for dust exposure data.

of the observations. It is seen that also in the full dataset one observation is positioned very extreme in the observation space. That observation (1246) is seen to have an extreme value of Cook's distance (Figure 4.9). When that extreme value is excluded, coefficient estimates for the variables years and smoking are similar to the estimates for the full dataset. However, coefficients differ for the variable concentration of dust by about 8% (see Table 4.7). Since observation 1246 is very far away from the data and the mean exposure is a variable that is not easy to measure exactly, one might suspect that the value of that variable is not trustworthy. One might consider the observation an outlier that should be omitted since it yields biased estimates. □

TABLE 4.7: Main effects model for dust exposure data without one observation.

| | Estimate | Std. Error | $z$-Value | Pr($>\|z\|$) |
|---|---|---|---|---|
| (Intercept) | −3.0620 | 0.2491 | −12.29 | 0.0000 |
| Dust | 0.0992 | 0.0239 | 4.15 | 0.0000 |
| Years | 0.0398 | 0.0062 | 6.40 | 0.0000 |
| Smoking | 0.6816 | 0.1745 | 3.91 | 0.0001 |

## 4.4 Structuring the Linear Predictor

The structuring of the linear predictor, and in particular the coding of categorical predictors, have already been considered briefly in Section 1.4. In the following we again have a look at the linear predictor and introduce a notation scheme due to Wilkinson and Rogers (1973).

### 4.4.1 The Linear Predictor: Continuous Predictors, Factors, and Interactions

The parametric binary regression model

$$\pi(\boldsymbol{x}) = P(y = 1|\boldsymbol{x}) = h(\boldsymbol{x}^T\boldsymbol{\beta})$$

that is considered here contains the linear predictor $\eta = \boldsymbol{x}^T\boldsymbol{\beta}$. In the simplest case the linear predictor contains the $p$ variables $x_1, \ldots, x_p$ in a main effect model of the form

$$\eta = \beta_0 + x_1\beta_1 + \cdots + x_p\beta_p.$$

However, frequently one finds that some interaction terms are necessary in the predictor, yielding a linear predictor of the form

$$\eta = \beta_0 + x_1\beta_1 + \cdots + x_p\beta_p + x_1x_2\beta_{12} + x_1x_3\beta_{13} + \cdots .$$

When some of the covariates are continuous, polynomial terms also may be included in the linear predictor, which means that the predictor is still linear in the parameters but not linear in the variables. If, for example, $x_1$ is continuous, a more flexible predictor containing non-linear effects of $x_1$ is

$$\eta = \beta_0 + x_1\beta_1 + \cdots + x_p\beta_p + x_1^2\beta_1^{(2)} + x_1^3\beta_1^{(3)} \cdots .$$

In many applications the influential variables are factors that may take a finite number of values. As in classical regression, these categorical covariates are included in the predictor in the form of dummy variables. When a categorical variable or factor $A$ takes values in $\{1, \ldots, k\}$ the linear predictor has the form

$$\eta = \beta_0 + x_{A(1)}\beta_{A(1)} + \ldots + x_{A(k-1)}\beta_{A(k-1)},$$

where $x_{A(i)}$ are dummy variables. In (0-1)-coding one has $x_{A(i)} = 1$ if $A = i$ and $x_{A(i)} = 0$ otherwise, which implies that $A = k$ is used as a reference category. When an additional continuous covariate $x$ is available, the predictor of the main effects model becomes

$$\eta = \beta_0 + x\beta_x + x_{A(1)}\beta_{A(1)} + \cdots + x_{A(k-1)}\beta_{A(k-1)}. \tag{4.7}$$

The inclusion of an interaction between the continuous predictor and the categorical variables means that all the products $xx_{A(i)}$ are included, yielding the more complicated predictor

$$\begin{aligned}\eta = \beta_0 + x\beta_x + x_{A(1)}\beta_{A(1)} + \cdots + x_{A(k-1)}\beta_{A(k-1)} + xx_{A(1)}\beta_{x,A(1)} + \cdots \\ + xx_{A(k-1)}\beta_{x,A(k-1)}.\end{aligned} \tag{4.8}$$

Interaction of this form means that the effect of one covariate is modified by the other. From the restructuring

$$\eta = \beta_0 + \cdots + x(\beta_x + x_{A(1)}\beta_{x,A(1)} + \cdots) + \cdots$$

one sees that the slope of $x$ is modified by the categorical variable, yielding different slopes for different categories of $A$. The restructuring

$$\eta = \beta_0 + \cdots + x_{A(i)}(\beta_{A(i)} + x\beta_{x,A(i)}) + \cdots$$

shows how the effects of variable $A$ are modified by the value of $x$. Interactions between two factors $A \in \{1, \ldots, k_A\}$ and $B \in \{1, \ldots, k_B\}$ may be modeled by including all products of dummy variables $x_{A(i)}x_{B(j)}$. For the interpretation of interactions that is familiar from linear modeling, one should have in mind that the the effects do not refer directly to the mean $\mathrm{E}(y|\boldsymbol{x})$ but to the transformed mean $g(\mathrm{E}(y|\boldsymbol{x}))$, which in the case of the logit model is the logits $\log(\pi(\boldsymbol{x}))/(1 - \pi(\boldsymbol{x}))$. In particular, for categorical predictors, interaction parameters have simple interpretations in terms of odds ratios (Exercise 4.4).

Since the form of the linear predictor is rather tedious to write down, it is helpful to use a notation that is due to Wilkinson and Rogers (1973). The notation uses operators to combine predictors in model formula terms. Let $x, z$ denote continuous, metrically scaled variables and $A, B, C$ denote factors.

### Basic Model Term

The model terms $X$ and $A$ stand for themselves. However, the linear predictor that is built depends on the type of variable. For a continuous variable $x$, the model term $X$ means that $x\beta_x$ is included. For the factor $A$, the term to be included is a set of dummy variables with corresponding weights $x_{A(i)}\beta_{A(i)}$.

### The + Operator

The + operator in a model formula means that the corresponding terms are added in the algebraic expression of the predictor. Therefore $X + A$ means that the predictor has the form (4.7).

### The Dot Operator

The dot operator is used to form products of the constituent terms. For example, the model term $X.Z$ refers to the algebraic expression $xz\beta_{x,z}$, whereas $X.A$ refers to the algebraic expression containing all products $xx_{A(i)}$, which is the interaction term in (4.8). Of course $X.Z$ is equivalent to $Z.X$. Polynomial terms of continuous variables my be built by $X.X$, referring to $x^2\beta$. The dot operator dominates the + operator, so that $A.B + X$ is equivalent to $(A.B) + X$. The notation of higher interaction terms is straightforward. $A.B.C$ denotes inclusion of the set of products $x_{A(i)}x_{B(j)}x_{C(k)}$. Of course it helps that the dot operator is commutative, that is, $(A.B).C$ is equivalent to $A.(B.C)$. It should be noted that sometimes different notations are used for the dot operator. For example, the statistical software R uses the notation $A : B$ for $A.B$.

### The Crossing Operator

The crossing operator is helpful for including marginal effects. The model term $A * B$ is an abbreviation for $A+B+A.B$. Similarly, the term $A*B*C$ means that the main effects and all two factor interactions are included. It is abbreviated as $A+B+C+A.B+A.C+B.C+A.B.C$. Crossing is distributive, meaning that $A * (B + C)$ is equivalent to $A * B + A * C$ (and to $A + B + C + A.B + A.C$).

TABLE 4.8: Wilkinson-Rogers notation for linear predictions with metric covariates $X, Z$ and factors $A, B$.

| | Model Term | Linear Predictor | Linear Predictor |
|---|---|---|---|
| | | | $A = i, B = j$ |
| Single | $X$ | $x\beta_x$ | $x\beta_x$ |
| | $A$ | $\sum_s x_{A(s)}\beta_{A(s)}$ | $\beta_{A(i)}$ |
| Addition | $X + Z$ | $x\beta_x + z\beta_z$ | $x\beta_x + z\beta_z$ |
| | $A + B$ | $\sum_s x_{A(s)}\beta_{A(s)} + \sum_s x_{B(s)}\beta_{B(s)}$ | $\beta_{A(i)} + \beta_{B(j)}$ |
| | $X + A$ | $x\beta_x + \sum_s x_{A(s)}\beta_{A(s)}$ | $x\beta_x + \beta_{A(i)}$ |
| Interaction | $X.Z$ | $xz\beta_{x,z}$ | $xz\beta_{xz}$ |
| | $A.B$ | $\sum_{s,r} x_{A(s)}x_{B(r)}\beta_{AB(s,r)}$ | $\beta_{AB(ij)}$ |
| | $X.A$ | $\sum_s xx_{A(s)}\beta_{xA(s)}$ | $x\beta_{xA(i)}$ |
| Hierarchical interaction | $X * Z$ | $x\beta_x + z\beta_z + xz\beta_{xz}$ | $x\beta_x + z\beta_z + xz\beta_{xz}$ |
| | $A * B$ | $\sum_s x_{A(s)}\beta_{A(s)} + \sum_r x_{B(r)}\beta_{B(r)} + \sum_{s,r} x_{A(s)}x_{B(r)}\beta_{AB(s,r)}$ | $\beta_{A(i)} + \beta_{B(j)} + \beta_{AB(ij)}$ |
| | $X * A$ | $x\beta_x + \sum_s x_{A(s)}\beta_{A(s)} + \sum_s xx_{A(s)}\beta_{xA(s)}$ | $x\beta_x + \beta_{A(i)} + x\beta_{xA(i)}$ |

Table 4.8 shows model terms in the notation of Wilkinson-Rogers together with the corresponding linear predictors for continuous covariates $X, Z$ and factors $A, B$. It should be noted that the interactions built by these model terms are a specific form of parametric interaction. Alternative concepts of interactions may be derived (see also Section 10.3.3 for smooth interaction terms).

### 4.4.2 Testing Components of the Linear Predictor

Most interesting testing problems concerning the linear predictor are linear hypotheses of the form

$$H_0 : \boldsymbol{C}\boldsymbol{\beta} = \boldsymbol{\xi} \text{ against } H_1 : \boldsymbol{C}\boldsymbol{\beta} \neq \boldsymbol{\xi},$$

where $\boldsymbol{C}$ is a fixed matrix of full rank $s \leq p$ and $\boldsymbol{\xi}$ is a fixed vector. In the simplest case one tests if one parameter can be omitted by considering

$$H_0 : \beta_j = 0 \text{ against } H_1 : \beta_j \neq 0,$$

which has the form of a linear hypothesis with $\boldsymbol{C} = (0, \ldots, 1, \ldots, 0)$ and $\boldsymbol{\xi} = 0$. The general no-effects hypothesis

$$H_0 : \boldsymbol{\beta} = \mathbf{0} \text{ against } H_1 : \boldsymbol{\beta} \neq \mathbf{0}$$

corresponds to a linear hypothesis with $\boldsymbol{C}$ denoting the unit matrix and $\boldsymbol{\xi} = \mathbf{0}$. If one wants to test if a factor $A$ has no effect, one has to test simultaneously if all the corresponding parameters are zero:

$$H_0 : \beta_{A(1)} = \cdots = \beta_{A(k-1)} = 0 \text{ against } H_1 : \beta_{A(j)} \neq 0 \text{ for one } j.$$

It is easily seen that in this case one also tests linear hypotheses. General test statistics for linear hypotheses have already been given in Section 3.7.2. Therefore, in the following the tests are considered only briefly.

#### Likelihood Ratio Statistic and the Analysis of Deviance

Let $M$ denote the model with linear predictor $\eta = \boldsymbol{x}^T\boldsymbol{\beta}$ and $\tilde{M}$ denote the submodel that is constrained by $\boldsymbol{C}\boldsymbol{\beta} = \boldsymbol{\xi}$. By using the notation of Section 4.2, the likelihood ratio statistic that compares models $\tilde{M}$ and $M$ has the form

$$\lambda = -2\{l(\boldsymbol{y}, \tilde{\boldsymbol{\pi}}) - l(\boldsymbol{y}, \hat{\boldsymbol{\pi}})\},$$

where $\tilde{\boldsymbol{\pi}}^T = (\tilde{\pi}_1, \ldots, \tilde{\pi}_n)$, $\tilde{\pi}_i = h(\boldsymbol{x}_i^T\tilde{\boldsymbol{\beta}})$, denotes the fit of submodel $\tilde{M}$ and $\hat{\boldsymbol{\pi}}^T = (\hat{\pi}_1, \ldots, \hat{\pi}_n)$, $\hat{\pi}_i = h(\boldsymbol{x}_i^T\hat{\boldsymbol{\beta}})$, denotes the fit of model $M$. Since for a binary distribution the deviances of models $\tilde{M}$ and $M$ are given by $-2l(\boldsymbol{y}, \tilde{\boldsymbol{\pi}})$ and $-2l(\boldsymbol{y}, \hat{\boldsymbol{\pi}})$, respectively, $\lambda$ is equivalent to the difference between deviances:

$$\lambda = D(\boldsymbol{y}, \tilde{\boldsymbol{\pi}}) - D(\boldsymbol{y}, \hat{\boldsymbol{\pi}}) = D(\tilde{M}) - D(M)$$

and has under mild conditions an asymptotic $\chi^2$-distribution with $s = rg(\boldsymbol{C})$ degrees of freedom. It is noteworthy that the difference of deviances yields the same value if computed from binomial responses or from the single binary variables that build the binomial response.

As shown in Section 3.7.2, a sequence of nested models $M_1 \subset M_2 \subset \cdots \subset M_m$ can be tested by considering the difference between successive models $D(M_i|M_{i+1}) = D(M_i) - D(M_{i+1})$ within the decomposition

$$D(M_1) = D(M_1|M_2) + \cdots + D(M_{m-1}|M_m) + D(M_m).$$

TABLE 4.9: Empirical odds and log-odds for duration of unemployment.

| Gender | Education Level | Short Term | Long Term | Odds | Log-Odds |
|---|---|---|---|---|---|
| m | 1 | 97 | 45 | 2.155 | 0.768 |
| | 2 | 216 | 81 | 2.667 | 0.989 |
| | 3 | 56 | 32 | 1.750 | 0.560 |
| | 4 | 34 | 9 | 3.778 | 1.330 |
| w | 1 | 105 | 51 | 2.059 | 0.722 |
| | 2 | 91 | 81 | 1.123 | 0.116 |
| | 3 | 31 | 34 | 0.912 | -0.092 |
| | 4 | 11 | 9 | 1.222 | 0.201 |

TABLE 4.10: Hierarchies for level (L) and gender (G).

| Model | Deviance | df ($p$-value) | Cond. Deviance | df ($p$-value) | Tested Effect |
|---|---|---|---|---|---|
| 1 | 32.886 | 7 (0.000) | | | |
| | | | 6.557 | 3 (0.087) | L(G ignored, L.G ignored) |
| 1+L | 26.329 | 4 (0.000) | | | |
| | | | 18.808 | 1 (0.000) | G(L.G ignored, L taken into account) |
| 1+L+G | 7.521 | 3 (0.057) | | | |
| | | | 7.521 | 3 (0.057) | L.G |
| L∗G | 0 | 0 | | | |
| 1 | 32.886 | 7 (0.000) | | | |
| | | | 17.959 | 1 (0.000) | G(L ignored, L.G ignored) |
| 1+G | 14.928 | 6 (0.021) | | | |
| | | | 7.406 | 3 (0.060) | L(L.G ignored, G taken into account) |
| 1+L+G | 7.521 | 3 (0.057) | | | |
| | | | 7.521 | 3 (0.057) | L.G |
| L∗G | 0 | 0 | | | |

The result is usually presented in an analysis of deviance table that gives the deviances of models, their differences, and the corresponding degrees of freedom (see Example 4.8).

Example 4.8: Duration of Unemployment

With the response duration of unemployment (1: short-term unemployment, less than 6 months; 0: long-term unemployment) and the covariates gender (1: male; 0: female) and level of education (1: lowest, up to 4: highest, university degree), one obtains the data given in Table 4.9. Analysis is based on the grouped data structure given in this table. The saturated logit model is given by

$$\text{logit}(\pi(G, L)) = \beta_0 + x_G\beta_G + x_L\beta_L + x_G x_L\beta_{GL},$$

which can be abbreviated by $G * L$. The testing of $\beta_{GL}, \beta_G, \beta_L$ yields the analysis of deviance table given in Table 4.10. It contains the deviances and the differences of the two sequences of nested models $1 \subset 1 + L \subset 1 + L + G \subset G * L$ and $1 \subset 1 + G \subset 1 + L + G \subset G * L$. In both sequences the intercept model is the strongest and the saturated model is the weakest. Thus different paths between these models can be tested. Starting from the saturated model, the first transition seems possible, since $D = 7.52$ on 3 df, which corresponds to a $p$-value of 0.057. However, in the first sequence the next transition to model $1 + L$ yields the difference of deviances $18.81$ on 1 df; therefore gender cannot be omitted. In the other sequence one considers the transition to model $1 + G$, obtaining the difference of deviances 7.41 on 3 df, which corresponds to a a $p$-value of 0.06. Although simplification seems possible, the model $1 + G$ does

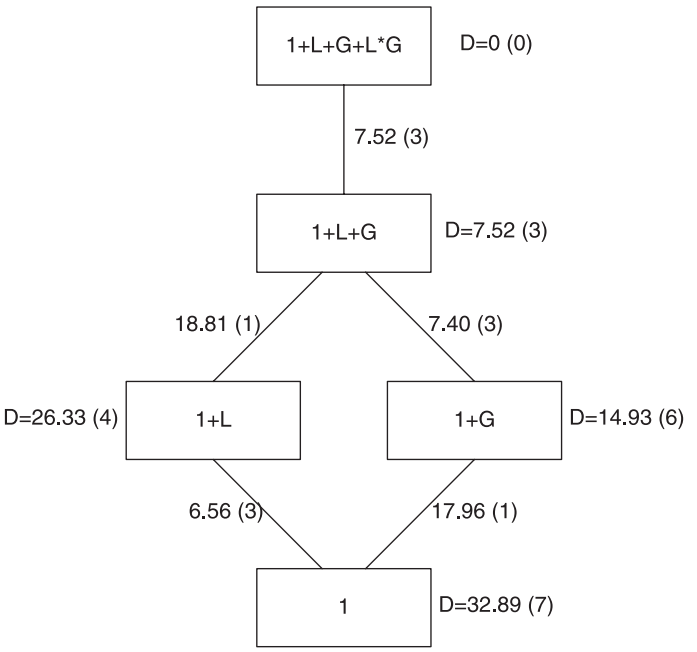


FIGURE 4.10: Analysis of deviance for unemployment data with factors level (L) and gender (G).

not fit will since the deviance 14.93 on 3 df is rather large, corresponding to a $p$-value of 0.02. Therefore, simplification beyond the main effect model does not seem to be warranted. □

### Alternative Test Statistics

Alternatives to the likelihood ratio statistic are the Wald test and the score test for linear hypotheses given by

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

and

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

Asymptotically, all three test statistics, the likelihood ratio statistic, the Wald test, and the score test, have the same distribution $\lambda, w, u \overset{(a)}{\sim} \chi^2(\text{rank}\,\boldsymbol{C})$.

### Contrasts and Quasi-Variances

Let us again consider the influence of a factor $A$ that takes values in $\{1, \ldots, k\}$. The corresponding linear predictor without constraints has the form

$$\eta = \beta_0 + x_{A(1)}\beta_{A(1)} + \ldots + x_{A(k)}\beta_{A(k)}.$$

Since the parameters $\beta_{A(1)}, \ldots, \beta_{A(k)}$ are not identifiable, side constraints have to be used. For example, by choosing the reference category $k$ one sets $\beta_{A(k)} = 0$. Among the full set of parameters $\beta_{A(1)}, \ldots, \beta_{A(k)}$ only contrasts, that is, linear combinations $\boldsymbol{c}^T\boldsymbol{\beta}$ with $\boldsymbol{c}^T = (c_1, \ldots, c_k)$,

TABLE 4.11: Estimates, standard errors, quasi-standard-errors and quasi-variances for duration of unemployment data.

| Level | Estimate $\hat{\beta}_{A(i)}$ | Standard Error | Quasi-Standard Error | Quasi-Variance |
|---|---|---|---|---|
| 1 | −0.170 | 0.305 | 0.124 | 0.015 |
| 2 | −0.273 | 0.295 | 0.097 | 0.009 |
| 3 | −0.637 | 0.323 | 0.163 | 0.026 |
| 4 | 0 | 0 | 0.278 | 0.077 |

$\sum_i c_i = 0$, $\boldsymbol{\beta}^T = (\beta_{A(1)}, \ldots, \beta_{A(k)})$, are identified. A simple contrast is $\beta_{A(i)} - \beta_{A(j)}$, which compares levels $i$ and $j$. For illustration we consider again the effect of the education level on the binary response duration of unemployment (Example 2.6). Table 4.11 shows the estimates together with standard errors for reference category 4. The given standard errors refer to the contrasts $\hat{\beta}_{A(i)} - \hat{\beta}_{A(4)}$ because $\hat{\beta}_{A(4)} = 0$. The disadvantage of the presentation of standard errors for fixed reference category is that all other contrasts are not seen. An alternative presentation of standard errors uses so-called *quasi-variances*, which have been proposed by Firth and De Menezes (2004). Quasi-variances $q_1, \ldots, q_k$ are constructed such that the variance of a contrast, $\text{var}(\boldsymbol{c}^T\hat{\boldsymbol{\beta}})$, is approximately $\sum_{i=1}^k c_i^2 q_i$. The corresponding "quasi-standard-errors" are given by $q_1^{1/2}, \ldots, q_k^{1/2}$. Quasi-variances (or quasi-standard-errors) can be used to determine the variance of contrasts. Consider the quasi-variances given in Table 4.11. For example, the approximate standard error for $\hat{\beta}_{A(1)} - \hat{\beta}_{A(4)}$ can be computed from quasi-variances by $(0.015 + 0.077)^{1/2}$, or from quasi-standard-errors, based on Pythagorean calculation, by $(0.124^2 + 0.278^2)^{1/2}$, yielding 0.303, which is a rather good approximation of the standard error of $\hat{\beta}_{A(1)}$ given in Table 4.11. But with the quasi-variances given in Table 4.11 also standard errors for $\hat{\beta}_{A(1)} - \hat{\beta}_{A(2)}$ can be computed yielding $(0.015 + 0.009)^{1/2} = 0.155$, which is rather large when compared to $\hat{\beta}_{A(1)} - \hat{\beta}_{A(2)} = 0.103$. Quasi-variances are a helpful tool in reporting

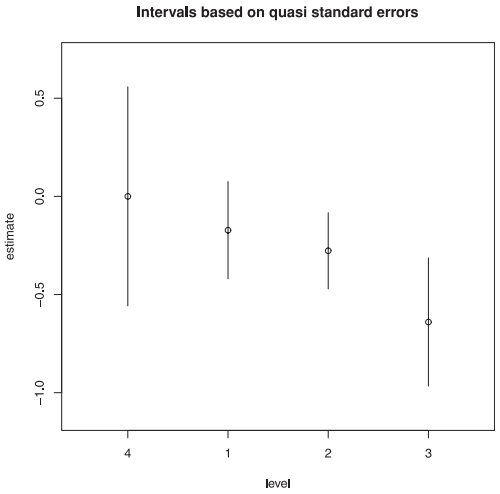


FIGURE 4.11: Quasi-standard-errors for duration of unemployment data.

standard errors for the level of factors. They are useful for investigating contrasts but should not be misinterpreted as standard errors for the parameters themselves. Firth and De Menezes (2004) give conditions under which the approximation is exact and investigate the accuracy in a variety of settings.

### 4.4.3 Ordered Categorical Predictors

The inclusion of categorical covariates in the linear predictor may strongly increase the dimension of the predictor space. Since a factor $A$ that takes values in $\{1, \ldots, k\}$ means that $k-1$ dummy variables have to be included, the number of observations within one factor level may become very small. In particular, if one wants to model interactions of factors, empty cells (for combinations of factors) will occur if the number of levels is large. The consequence frequently is that estimates of effects do not exist. In that case practioners typically join some factor categories to reduce the dimension of the predictor space, but at the cost of loosing information.

The coding in dummy variables for factors is based on the nominal scale of the factor. The categories $\{1, \ldots, k\}$ represent simple labels for levels of the factor; the ordering of the categories is arbitrary and should not be used. In practice, however, the categories of a factor are often ordered, and the variable that constitutes the predictor is an ordered categorical variable. Then the ordering should be used to obtain a more sparse representation and therefore more stable estimates. One way to use the ordering is to define assigned scores to the categories of the factor. Although that "solution" to the dimensional problem is widespread, it is not satisfactory. By assigning scores and using a linear or quadratic model for the scores one definitely assumes a higher scale level for the factor. However, linear or quadratic terms are appropriate only when the influential variable is metrically scaled, albeit discrete. In particular, differences of values of the variable have to be meaningful. For ordered factors this is usually not the case. If the factor levels represent subjective judgements like "strong agreement," "slight agreement," and "strong disagreement," the assigned scores are typically useless. Interpretation depends on the assigned scores, which are to a certain extent arbitrary, and different sets of scores yield different effect strengths.

It is slightly different when the ordinal scale of the factor is due to an underlying continuous variable. For example, often variables like income brackets and categorized age are available and it is known how the categories have been built. For a categorized covariate age, available in categories $[20, 29), [30, 39), \ldots, [60, 80)$, one may build mid-range scores and use them as a substitute for the true underlying age. Then one approximates the unobserved covariate age in the predictor. What works for covariate age may work less well when the predictor is categorized income. Since only categories are available, it is hard to know what values hide in the highest interval, because it has been learned that incomes can be extremely high. Then the mid-range score of the last category is a mere guess. In addition, the score of the highest category is at the boundary of the predictor space and therefore tends to be influential. Thus one has an influential observation but has to choose a score.

An alternative to assigning scores that takes the ordering seriously but does not assume knowledge of the true score is to use penalized estimates where the penalty explicitly uses the ordering of categories. Let the predictor of a single factor $A$ with ordered categories $1, \ldots, k$ be given in the form

$$\eta = \beta_0 + x_{A(2)}\beta_2 + \ldots + x_{A(k)}\beta_k,$$

where $x_{A(2)}, \ldots, x_{A(k)}$ denotes $k-1$ dummy variables in (0–1)-coding, therefore implicitly using category 1 as the reference category. However, instead of maximizing the usual log-likelihood $l(\boldsymbol{\beta})$, estimates of the parameter vector $\boldsymbol{\beta}^T = (\beta_0, \beta_2, \ldots, \beta_k)$ are obtained by

maximizing the *penalized log-likelihood*:

$$l_p(\boldsymbol{\beta}) = l(\boldsymbol{\beta}) - \frac{\lambda}{2} P(\boldsymbol{\beta}),$$

where the penalty term is given by

$$P(\boldsymbol{\beta}) = \sum_{j=2}^{k} (\beta_j - \beta_{j-1})^2$$

with $\beta_1 = 0$. Therefore, the differences of adjacent parameters are penalized with the strength of the penalty determined by $\lambda$. For $\lambda = 0$, maximization of $l_p(\boldsymbol{\beta})$ yields the usual maximum likelihood estimate whereas $\lambda \to \infty$ yields $\hat{\beta}_j = 0, j = 1, \ldots, k$. For an appropriately chosen $\lambda$, the penalty restricts the variability of the parameters across the response categories, thereby assuming a kind of "smooth effect" of the ordinal categories on the dependent variable. The method is strongly related to the penalization techniques considered in Chapter 6.

In matrix form the penalty is given by $P(\boldsymbol{\beta}) = \boldsymbol{\beta}^T \boldsymbol{D}^T \boldsymbol{D} \boldsymbol{\beta} = \boldsymbol{\beta}^T \boldsymbol{K} \boldsymbol{\beta}$, where $\boldsymbol{K} = \boldsymbol{D}^T \boldsymbol{D}$ and $\boldsymbol{D}$ is the $((k-1) \times k)$-matrix:

$$\boldsymbol{D} = \begin{pmatrix} 0 & 1 & 0 & \ldots & & & 0 \\ 0 & -1 & 1 & & & & \vdots \\ 0 & & -1 & 1 & & & \\ \vdots & & \ddots & & \ddots & & \\ 0 & 0 & \ldots & 0 & & -1 & 1 \end{pmatrix}. \tag{4.9}$$

The corresponding penalized score function is

$$\boldsymbol{s}_p(\boldsymbol{\beta}) = \sum_{i=1}^{n} \boldsymbol{x}_i \frac{\partial h(\eta_i)}{\partial \eta} (y_i - \mu_i)/\sigma_i^2 - \lambda \boldsymbol{K} \boldsymbol{\beta}.$$

The estimation equation $\boldsymbol{s}_p(\hat{\boldsymbol{\beta}}_p) = \mathbf{0}$ may be solved by iterative pseudo-Fisher scoring:

$$\hat{\boldsymbol{\beta}}_p^{(k+1)} = \hat{\boldsymbol{\beta}}_p^{(k)} + \boldsymbol{F}_p(\hat{\boldsymbol{\beta}}_p^{(k)})^{-1} \boldsymbol{s}_p(\hat{\boldsymbol{\beta}}_p^{(k)})^{-1},$$

where $\boldsymbol{F}_p(\boldsymbol{\beta}) = \boldsymbol{F}(\boldsymbol{\beta}) - \lambda \boldsymbol{K}$ and $\boldsymbol{F}(\boldsymbol{\beta}) = E(-\partial^2 l / \partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T)$. Approximate covariances are obtained by the sandwich matrix

$$\text{cov}(\hat{\boldsymbol{\beta}}_p) \approx (\boldsymbol{F}(\hat{\boldsymbol{\beta}}_p) + \lambda \boldsymbol{K})^{-1} \boldsymbol{F}(\hat{\boldsymbol{\beta}}_p) (\boldsymbol{F}(\hat{\boldsymbol{\beta}}_p) + \lambda \boldsymbol{K})^{-1}.$$

An advantage of the penalized estimate $\hat{\boldsymbol{\beta}}_p$ is that it will exist even in cases where the unpenalized ML estimate $\hat{\boldsymbol{\beta}}$ does not exist. Unconstrained estimates may not exist even in the simple case of a single factor. For the logit model, unconstrained estimates are given by $\hat{\beta}_0 = \log(p_1/(1-p_1)), \hat{\beta}_j = \log(p_j/(1-p_j)) - \hat{\beta}_0, j = 2, \ldots, p$, with $p_j$ denoting the observed relative frequencies in category $j$. Therefore, estimates do not exist if one of the relative frequencies is zero or one. In the more general case, where the intercept term is replaced by a linear term $\boldsymbol{x}^T \boldsymbol{\beta}_x$, which contains an additional vector of predictors, non-existence will occur more frequently.

### Alternative Representation

Before demonstrating the advantages of penalized estimates, an alternative representation of the approach by use of split-coding is given. Let us separate the weights on predictors from the intercept by considering the decomposition $\boldsymbol{\beta}^T = (\beta_0, \boldsymbol{\beta}_c^T)$. In analogy the matrix $\boldsymbol{D}$ is decomposed into $\boldsymbol{D} = [\boldsymbol{0}|\boldsymbol{D}_c]$. Then the penalty may be given as

$$J(\boldsymbol{\beta}) = \boldsymbol{\beta}_c^T \boldsymbol{K}_c \boldsymbol{\beta}_c, \tag{4.10}$$

where $\boldsymbol{K}_c = \boldsymbol{D}_c^T \boldsymbol{D}_c$ and $\boldsymbol{\beta}_c^T = (\beta_2, \ldots, \beta_k)$. The penalized maximization problem can be transformed into a problem with a simpler penalty by use of the inverse $\boldsymbol{D}_c^{-1}$, which is a lower triangular matrix with 1 on and below the diagonal. By using the parametrization $\tilde{\boldsymbol{\beta}} = \boldsymbol{D}_c \boldsymbol{\beta}_c$ with $\tilde{\boldsymbol{\beta}}^T = (\tilde{\beta}_1, \ldots, \tilde{\beta}_{k-1})$ the penalty $\boldsymbol{\beta}_c^T \boldsymbol{K}_c \boldsymbol{\beta}_c$ turns into the simpler form

$$J(\boldsymbol{\beta}) = \tilde{\boldsymbol{\beta}}^T \tilde{\boldsymbol{\beta}} = \sum_{i=1}^{k-1} \tilde{\beta}_i^2,$$

which is a ridge penalty on the transformed parameters. The corresponding transformation of the design matrix is based on the decomposition $\boldsymbol{X} = [\boldsymbol{1}|\boldsymbol{X}_c]$. One obtains $\boldsymbol{X}\boldsymbol{\beta} = [\boldsymbol{1}\beta_0|\boldsymbol{X}_c \boldsymbol{D}_c^{-1} \tilde{\boldsymbol{\beta}}]$. Then one row of the design matrix $\boldsymbol{X}_c \boldsymbol{D}_c^{-1}$ is given by the vector $(\tilde{x}_{A(1)}, \ldots, \tilde{x}_{A(k-1)}) = (x_{A(2)}, \ldots, x_{A(k)}) \boldsymbol{D}_c^{-1}$, where the dummies $\tilde{x}_{A(i)}$ are given in split-coding:

$$\tilde{x}_{A(i)} = \begin{cases} 1 & \text{if } A > i \\ 0 & \text{otherwise,} \end{cases}$$

which is a coding scheme that distinguishes between categories $\{1, \ldots, i\}$ and $\{i+1, \ldots, k\}$ (for split-coding compare Section 1.4.1). The transformation $\tilde{\boldsymbol{\beta}} = \boldsymbol{D}_c \boldsymbol{\beta}_c$ yields the parameters $\tilde{\beta}_1 = \beta_2$, $\tilde{\beta}_2 = \beta_3 - \beta_2$, $\tilde{\beta}_{k-1} = \beta_k - \beta_{k-1}$, which are used in the corresponding predictor $\eta = \beta_0 + \sum_{i=1}^{k-1} \tilde{x}_{A(i)} \tilde{\beta}_i$. The transformation shows that the smoothness penalty $\boldsymbol{\beta}^T \boldsymbol{K} \boldsymbol{\beta}$ can be represented as a ridge penalty for the parameters corresponding to split-coding, and thus smoothness across categories is transformed into penalizing transitions between groups of adjacent categories.

#### Example 4.9: Simulation

In a small simulation study the penalized regression approach is compared to pure dummy coding and a binary regression model with a linear predictor that takes the group labels as (metric) independent variable. The underlying model is the logit model $P(y = 1) = \exp(\boldsymbol{x}^T \boldsymbol{\beta})/(1 + \exp(\boldsymbol{x}^T \boldsymbol{\beta}))$. As true values of $\boldsymbol{\beta}$ we assume an approximately linear structure (Figure 4.12, top left) and an obviously non-linear coefficient vector (Figure 4.12, bottom left). In each of the 100 simulations $n = 330$ values of $\boldsymbol{x}$ were generated. Figure 4.12 shows the mean squared error for the estimation of $\boldsymbol{\beta}$. It is seen that the linear predictor performs well when the underlying structure is approximately linear but fails totally when the structure is non-linear. Simple dummy coding should adapt to both scenarios, but due to estimation uncertainty it performs worse than the linear predictor approach in the first scenario but better in the second scenario. The penalized estimate with smoothed dummies outperforms both approaches distinctly. It adapts well to the approximately linear and to the distinctly non-linear structure. □

So far we assumed a single independent variable, but models with several predictors are an obvious extension. Only the penalty matrix has to be modified. Now one uses a block-diagonal structure with the blocks given by the penalty matrix for a single ordered predictor. In the following example three ordinal predictors are used.

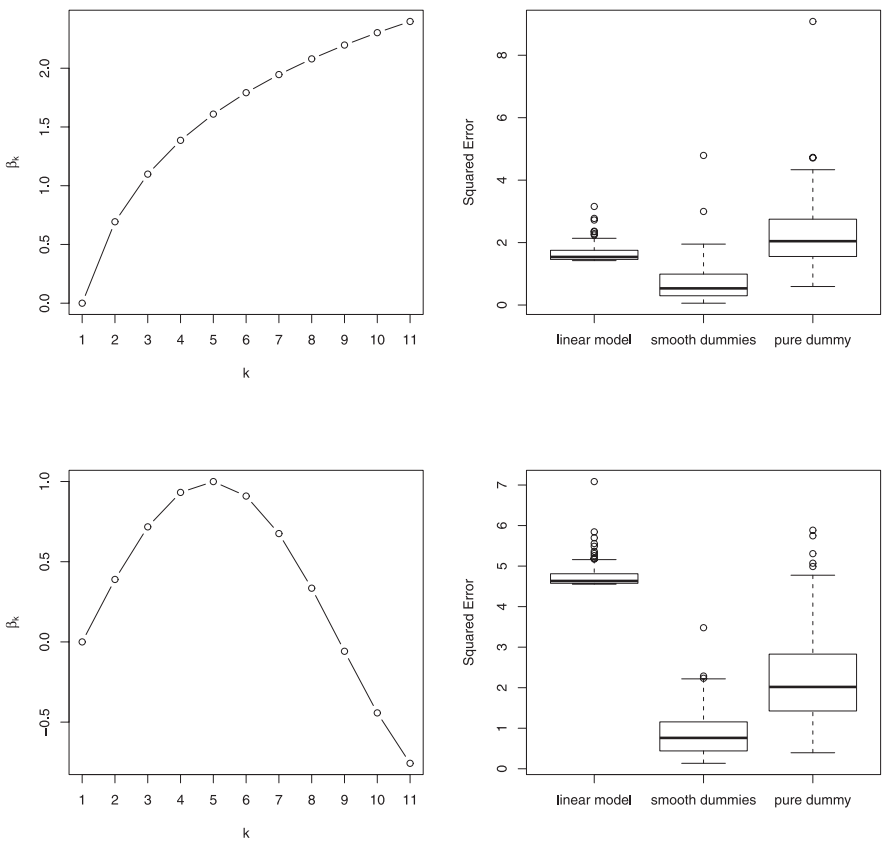


FIGURE 4.12: True coefficient vectors (left) and squared errors (right) for the considered methods after 100 simulation runs with $\sigma^2 = 2$.

### Example 4.10: Choice of Coffee Brand

In this example the binary response is coffee brand, which is only separated into cheap coffee from a German discounter and real branded products. The ordinal predictors are monthly income (in four categories), social class (in five categories), and age group in five categories, below 25, 25–39, 40–49, 50–59, above 59). While one might consider some sort of midpoints for age group and monthly income, it is hardly imaginable for social class. Table 4.12 shows the estimated coefficients of corresponding dummy variables for a logit model with group labels as predictors. The comparison refers to linear modeling, pure dummy and smoothed ordered dummies. It can be seen that penalization yields much less variation in the coefficients for single variables when compared to pure dummy coding.

To investigate the methods' performance in terms of prediction accuracy, the data were split randomly into training ($n = 100$) and test ($m = 100$) data. The training data are used to fit the model, the test set for evaluation only. As a measure of prediction accuracy we take the sum of squared deviance residuals (SSDR) on the test set. It is remarkable that the pure dummy model cannot be fitted due to complete data separation in 68% of the splits. Figure 4.13 summarizes the results in terms of SSDR after 200 random splits. It is seen that smooth dummies are to be preferred, in particular because the fit of pure dummies exists only in 32% of the splits. □

As the previous example shows, ML estimates often do not exist when simple dummy coding is used. Penalized estimates have the advantage that estimates exist under weaker

TABLE 4.12: Coefficients of corresponding dummy variables, estimated by the use of a (generalized) linear model, i.e., logit model, with group labels as predictors, penalized regression types I and II ($\lambda = 10$ in each case), and a logit model based on pure dummy coding.

| | | Linear Model | Smooth Dummies | Pure Dummy |
|---|---|---|---|---|
| Intercept | | −0.36 | −0.81 | −0.38 |
| Income | 2 | 0.02 | −0.05 | −0.13 |
| | 3 | 0.03 | −0.02 | 0.29 |
| | 4 | 0.05 | −0.04 | 0.17 |
| Social Class | 2 | −0.28 | −0.14 | −0.92 |
| | 3 | −0.56 | −0.31 | −1.39 |
| | 4 | −0.84 | −0.39 | −1.28 |
| | 5 | −1.12 | −0.56 | −1.96 |
| Age Group | 2 | −0.10 | 0.06 | 0.79 |
| | 3 | −0.20 | −0.09 | −0.29 |
| | 4 | −0.30 | 0.05 | 0.84 |
| | 5 | −0.40 | −0.18 | −0.04 |

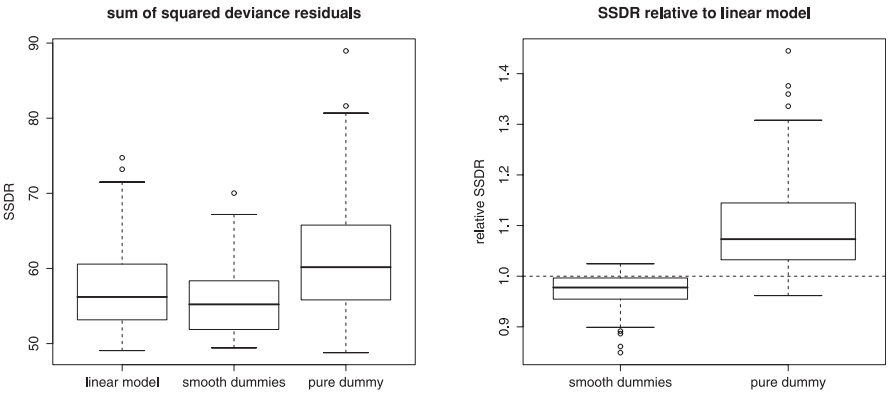


FIGURE 4.13: Performance (in terms of SSDR) of a (generalized) linear regression on the group labels, penalized regression for smooth dummy coefficients, and a pure dummy model (for the latter only the 68% successful estimates have been used) (left). Observed values for all considered methods; (right) SSDR values relative to linear model.

assumptions. It may be seen as an disadvantage that an additional tuning parameter has to be selected. However, this is easily done for example by cross-validation or minimization of information criteria like the $AIC$. For details and more simulations and applications, see Gertheiss and Tutz (2009b).

The ridge-type penalty ensures the existence of estimates and yields a smooth estimation of effects for ordered categorical predictors. An alternative strategy is to search for the categories that actually have different effects, or, in other words, to identify the categories that can be collapsed with respect to the dependent variable. In the case of ordered categorical predictors, collapsing naturally refers to adjacent categories. This selection and collapsing approach for ordered predictors is considered within selection procedures in Section 6.5.

## 4.5 Comparing Non-Nested Models

The analysis of deviance is a useful tool that allows one to distinguish between relevant and irrelevant terms in the linear predictor. The underlying strategy is based on a comparison of the nested models. One typically compares the model that contains the term in question to the model in which the term is omitted.

Analysis of deviance cannot be used if the models to be compared are non-nested, because differences of deviances have no standard distribution. If two models are not nested, one should distinguish between two cases. When the two models have the same number of parameters one can compute goodness-of-fit statistics, if available, and at least find out which model shows a better fit to the data. Although there is no benchmark for the comparison in the form of a standard distribution, it shows how strong the goodness-of-fit varies across models. However, if in addition the models to be compared have different numbers of parameters, goodness-of-fit will tend to favor the model that contains more parameters because the additional flexibility of the model will yield a better fit. Various criteria that take the number of parameters into account have been proposed. A widely used criterion is Akaike's information criterion which he named $AIC$, for "an information criterion" (Akaike, 1973). It is defined by

$$AIC = -2\log(L(\hat{\boldsymbol{\beta}})) + 2 \cdot (\text{number of fitted parameters}),$$

where $\log(L(\hat{\boldsymbol{\beta}}))$ denotes the log-likelihood of the fitted model evaluated at the ML estimate $\hat{\boldsymbol{\beta}}$. The second term may be seen as a penalty accounting for the number of fitted parameters. For binary response models with ungrouped data the deviance is given by $D(\boldsymbol{y}, \hat{\boldsymbol{\pi}}) = -2\log(L(\hat{\boldsymbol{\beta}}))$ and one obtains

$$AIC = D(\boldsymbol{y}, \hat{\boldsymbol{\pi}}) + 2 \cdot (\text{number of fitted parameters}).$$

Therefore, in the $AIC$ criterion the number of fitted parameters is added to the deviance as a measure for the discrepancy between the data and the fit. The correction to the deviance may be derived by asymptotic reasoning. $AIC$ has been shown to be an approximately unbiased estimate of the mean log-density of a new independent data set (see also Appendix D). A careful investigation of $AIC$ and alternative model choice criteria was given by Burnham and Anderson (2002). While $AIC$ is an information-theoretic measure based on Kullback-Leibler distance, the $BIC$ (for Bayesian information criterion) has been derived in a Bayesian context by Schwarz (1978). It has the form

$$BIC = -2\log(L(\hat{\boldsymbol{\beta}})) + \log(n) \cdot (\text{number of fitted parameters}).$$

$AIC$ and $BIC$ are not directly comparable since the underlying targets differ. Based on its derivation as an unbiased estimate of the mean density of a new dataset with equal sample size, $AIC$ is specific for the sample size at hand. In contrast, derivation of $BIC$ assumes a true generating model, independent of sample size, although selection of the true model is obtained only asymptotically. For a comparison of these selection criteria including multimodel inference, see Burnham and Anderson (2004).

**Akaike's Criterion**

$$AIC = -2\log(L) + 2 \cdot (\text{number of fitted parameters})$$

**Bayesian Information Criterion**

$$BIC = -2\log(L) + \log(n) \cdot (\text{number of fitted parameters})$$

Although a comparison of non-nested models by testing is not straightforward, some methods were proposed in the econometric literature; see, for example, Vuong (1989) , who proposed likelihood ratio tests for model selection with tests on non-nested hypotheses. An alternative strategy is to compare the models in terms of prediction accuracy. When the sample is split several times into a learning dataset (for fitting of the model) and a test dataset (for evaluating prediction performance) one chooses the model that has the better performance (see Chapter 15).

## 4.6 Explanatory Value of Covariates

When using a regression model one is usually interested in describing the strength of the relation between the dependent variable and the covariates. In classical linear regression the most widely used measure is the squared multiple correlation coefficient $R^2$, also called the *coefficient of determination*. The strength of $R^2$ is its simple interpretation as the proportion of variation explained by the regression model. As a descriptive measure it may be derived from the partioning of squared residuals:

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{\mu}_i - \bar{y})^2 + \sum_{i=1}^{n}(\hat{\mu}_i - y_i)^2, \tag{4.11}$$

where $\bar{y}$ denotes the mean across observations $y_1, \ldots, y_n$ and $\hat{\mu}_i = \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}$ denotes the fitted values when $\hat{\boldsymbol{\beta}}$ is estimated by least squares and $\boldsymbol{x}_i$ contains an intercept. The term on the left-hand side is the total sum of squares (SST), the first term on the right-hand side corresponds to the sum of squares explained by regression (SSR), and the second term is the error sum of squares (SSE). The empirical *coefficient of determination* $R^2$ is defined by

$$R^2 = \frac{\text{SSR}}{\text{SST}} = \frac{\sum_i(\hat{\mu}_i - \bar{y})^2}{\sum_i(y_i - \bar{y})^2}.$$

With SST being the total variation without covariates, $R^2$ gives the proportion of variation explained by the covariates. If $\hat{\mu}_i$ is obtained by least-squares fitting, $R^2$ is equivalent to the squared correlation between the observations $y_i$ and fitted values $\hat{\mu}_i$ (for more on $R^2$ see also Section 1.4.5).

The partitioning of squared residuals (4.11) may be seen as an empirical version of the decomposition of variance into "between variance" $\text{var}(\text{E}(y|\boldsymbol{x}))$ (explained by regression on $\boldsymbol{x}$) and "within variance" $\text{E}(\text{var}(y|\boldsymbol{x}))$ (or error):

$$\text{var}(y) = \text{var}(\text{E}(y|\boldsymbol{x})) + \text{E}(\text{var}(y|\boldsymbol{x})),$$

which holds without distributional assumptions for random variables $y, \boldsymbol{x}$. Thus $R^2$ is an empirical version of the true or theoretical proportion of explained variance:

$$R_T^2 = \text{var}(\text{E}(y|\boldsymbol{x}))/\,\text{var}(y) = \{\text{var}(y) - \text{E}(\text{var}(y|\boldsymbol{x}))\}/\,\text{var}(y),$$

which represents the population coefficient of determination. For the population measure $R_T^2$ one has a similar property as for the empirical coefficient of determination. It represents the proportion of explained variance but also equals the squared correlation between the random variables $y$ and $E(y|\boldsymbol{x})$:

$$R_T^2 = \text{var}(\text{E}(y|\boldsymbol{x}))/\,\text{var}(y) = \text{cor}(y, \text{E}(y|\boldsymbol{x}))^2. \tag{4.12}$$

Therefore, in the linear model the empirical coefficient of determination $R^2$, obtained from least-squares fitting, as well as the empirical correlation coefficient are estimates of $R^2_T$.

The extension to GLMs has some pitfalls. Although the equivalence (4.12) holds more generally for GLMs, the empirical correlation coefficient between the data and values fitted by ML estimation is not the same as $R^2 = \mathrm{SSR}\,/\,\mathrm{SST}$. Moreover, in GLMs it is more difficult to interpret $R^2_T$ as a proportion of the explained variation. For example, when responses are binary, the mean and variance of the response are strictly linked. Therefore, $\mathrm{var}(y|\boldsymbol{x})$ is not fixed but varies with $\boldsymbol{x}$. This is different from the case of a normally distributed vector $(y, \boldsymbol{x})$ for which $\mathrm{var}(y|\boldsymbol{x})$ is a constant value and therefore the variability of $\mathrm{E}(y|\boldsymbol{x})$ and $\mathrm{var}(y|\boldsymbol{x})$, which determine the decomposition of $\mathrm{var}(y)$, are strictly separated. In addition, the population measure $R^2_T$ can be severely restricted to small values, although a strong relation between the predictor and the response is present. Cox and Wermuth (1992) considered a linear regression model for binary responses with a single explanatory variable, $P(y = 1|x) = \pi + \beta(x - \mu_x)$, where $x$ has mean $\mu_x$ and variance $\sigma^2_x$, and $\pi = P(y = 1)$. The value of $R^2_T = \beta^2\sigma^2_x/(\pi(1-\pi))$ is primarily determined by the variance of $x$, and for a sensible choice of the variance $R^2_T = 0.36$ is the largest value that can be achieved. Therefore, explained variation as a proportion of the explained variance is restricted to rather small values.

The different interpretations of the coefficient of determination in the linear model have led to various measures for non-linear models, most of them more descriptive in nature. Although there is no widely accepted direct analog to $R^2$ from least-squares regression, a number of $R^2$ for binary response models are in common use. We give some of them in the following.

### 4.6.1 Measures of Residual Variation

The coefficient of determination is a member of a more general class of measures that aim at the proportional decrease in variation obtained in going from a simple model to a more complex model that includes explanatory variables. The measures have the general form

$$R(M_0|M) = \frac{D(M_0) - D(M)}{D(M_0)}, \tag{4.13}$$

where $D(M)$ measures the (residual) variation of model $M$ and $D(M_0)$ the variation of the simpler model $M_0$, which typically is the intercept model (compare Efron, 1978). The *proportional reduction in variation measure* (4.13) can be seen as a descriptive statistic or a population measure depending on the measure of variation that is used. For example, the empirical coefficient of determination $R^2$ is obtained by using $D(M_0) = SST$, $D(M) = SSE$; the population value is based on the corresponding distribution models and uses $D(M_0) = var(y)$, $D(M) = E(var(y|\boldsymbol{x}))$. However, in most applications measures of the form (4.13) are used as descriptive tools without reference to an underlying population value. The form (4.13) has more generally been referred to as a measure of proportional reduction in loss (or error) that is used to quantify reliability (Cooil and Rust, 1994).

Least squares as measures of variation are linked to the estimation in linear models with normally distributed responses. For generalized linear models the preferred estimation method is maximum likelihood. The corresponding variation measure is the deviance, which explicitly uses the underlying distribution. Let $l(\hat{\boldsymbol{\beta}})$ denote the maximal log-likelihood of the fitted model, $l(\hat{\beta}_0)$ denote the log-likelihood of the intercept model, and $l(\mathrm{sat})$ be the log-likelihood of the saturated model. With $D(M) = D(\hat{\boldsymbol{\beta}}) = -2(l(\hat{\boldsymbol{\beta}}) - l(\mathrm{sat}))$, $D(M_0) = D(\hat{\beta}_0) = -2(l(\hat{\beta}_0) - l(\mathrm{sat}))$ one obtains the deviance-based measure

$$R^2_{dev} = \frac{D(\hat{\beta}_0) - D(\hat{\boldsymbol{\beta}})}{D(\hat{\beta}_0)} = \frac{l(\hat{\boldsymbol{\beta}}) - l(\hat{\beta}_0)}{l(\mathrm{sat}) - l(\hat{\beta}_0)}.$$

$R^2_{dev}$ compares the reduction of the deviance when the regression model is fitted instead of the simple intercept model to the deviance of the intercept model.

For ungrouped binary observations one has $D(\hat{\boldsymbol{\beta}}) = -2l(\hat{\boldsymbol{\beta}})$ and the coefficient is equivalent to McFadden's (1974) *likelihood ratio index* (also called pseudo-$R^2$):

$$R^2_{dev} = \frac{l(\hat{\beta}_0) - l(\hat{\boldsymbol{\beta}})}{l(\hat{\beta}_0)}.$$

It is seen that $0 \leq R^2_{dev} \leq 1$ with

$$R^2_{dev} = 0 \quad \text{if} \quad l(\hat{\beta}_0) = l(\hat{\boldsymbol{\beta}}), \text{that is, if all other parameters have zero estimates:}$$
$$R^2_{dev} = 1 \quad \text{if} \quad D(\hat{\boldsymbol{\beta}}) = 0, \text{that is, if the model shows perfect fit, } \hat{\pi}_i = y_i.$$

$R^2_{dev}$ is directly linked to the likelihood ratio statistic $\lambda$ by $\lambda = R^2_{MF}(-2l(\hat{\beta}_0))$, which has asymptotic $\chi^2$-distribution.

It should be noted that the deviance-based measure depends on the level of aggregation because the deviance for grouped observations differs from the deviance for individual observations. Deviances based on ungrouped data are to be preferred because if one fits a model that contains many predictors, the fit can be perfect for grouped data yielding $D(\hat{\boldsymbol{\beta}}) = 0$, although individual observations are not well explained. As an explanatory measure for future data, which will will come as individual data it is certainly insufficient. For the use of deviances for grouped observations see also Theil (1970) and Goodman (1971).

Although the deviance has some appeal when considering GLMs, alternative measures of variation can be used. One candidate is squared error with $D(M_0) = SST$, $D(M) = SSR$, yielding $R^2_{SE}$ (e.g., Efron, 1978) . For binary observations, the use of the empirical variance $D(M_0) = n\bar{p}(1-\bar{p})$, where $\bar{p}$ is the proportion of ones in the total sample, and $D(M) = \sum_i \hat{\pi}_i(1-\hat{\pi}_i)$ yields Gini's concentration measure:

$$G = \frac{\sum_i \hat{\pi}_i^2 - n\bar{p}^2}{n\bar{p}(1-\bar{p})},$$

which was also discussed by Haberman (1982) (Exercise 4.7).

A prediction-oriented criterion that has been proposed uses the predictions based on $\bar{p}$ (without covariates) and $\hat{\pi}_i$ (with covariates). The variation measures $D$ are defined by

$$D(M) = \sum_i L(y_i, \hat{\pi}_i), \quad D(M_0) = \sum_i L(y_i, \bar{p}),$$

where $L$ is a modified (0–1) loss function with $L(y, \hat{\pi}) = 1$ if $|y - \hat{\pi}| > 0.5$, and $L(y, \hat{\pi}) = 0.5$ if $|y - \hat{\pi}| = 0.5$, and $L(y, \hat{\pi}) = 0$ if $|y - \hat{\pi}| < 0.5$. The corresponding measure $R_{\text{class}}$ compares the number of misclassified observations ("non-hits") obtained without covariates to the number of misclassified observations obtained by using the model $M$, where the simple classification rule $\hat{y} = 1$ if $\hat{\pi}_i > 0.5$ and $\hat{y} = 0$ if $\hat{\pi}_i < 0.5$, with an adaptation for ties, is used. The measure uses the threshold 0.5, which corresponds to prior probabilities $P(y = 1) = P(y = 0)$ in terms of classification (see Chapter 15). The measure has a straightforward interpretation but depends on the fixed threshold. It is also equivalent to Goodman and Kruskal's $\lambda$ (Goodman and Kruskal, 1954); see also van Houwelingen and Cessie (1990). A problem with measures like the number of misclassified errors is that they are a biased measure of the underlying true error rate in future samples. Since the sample is used to estimate a classification rule and to evaluate its performance, it underestimates the true error. The resulting error is also called a reclassification error. Better measures are based on cross-classification or leaving-one-out versions (for details see Section 15.3).

**Measures for Explanatory Value of Covariates**

Pseudo-$R^2$

$$R^2_{dev} = \frac{l(\hat{\beta}_0) - l(\hat{\boldsymbol{\beta}})}{l(\hat{\beta}_0)}.$$

Squared Error

$$R^2_{SE} = 1 - \frac{\sum_{i=1}^n (y_i - \hat{\pi}_i)^2}{\sum_{i=1}^n (y_i - \hat{y})^2}$$

Gini

$$G = \frac{\sum_i \hat{\pi}_i^2 - n\bar{p}^2}{n\bar{p}(1-\bar{p})}.$$

Reduction in Classification Error

$$R_{\text{class}} = \frac{\text{non-hits}(M_0) - \text{non-hits}(M)}{\text{non-hits}(M_0)},$$

Cox and Snell

$$R^2_{LR} = 1 - \left(\frac{L(\hat{\beta}_0)}{L(\hat{\boldsymbol{\beta}})}\right)^{2/n}$$

### 4.6.2 Alternative Likelihood-Based Measures

Since the coefficient of determination has several interpretations, alternative generalizations are possible. Cox and Snell (1989) considered

$$R^2_{LR} = 1 - \left(\frac{L(\hat{\beta}_0)}{L(\hat{\boldsymbol{\beta}})}\right)^{2/n},$$

where $L(\hat{\beta}_0), L(\hat{\boldsymbol{\beta}})$ denote the likelihoods of the two models. When $L(\hat{\beta}_0), L(\hat{\boldsymbol{\beta}})$ are the likelihoods of a normal response model, $R^2_{LR}$ reduces to the standard $R^2$ of a classical linear regression. With a binary regression model $R^2_{LR}$ cannot obtain a value of one even if the model predicts perfectly. Therefore, a correction has been suggested by Nagelkerke (1991). He proposed using $R^2_{corr} = R^2_{LR}/(1 - L(\hat{\beta}_0))^{2/n}$, which is a simple rescaling by using the maximal value $(1 - L(\hat{\beta}_0))^{2/n}$ that can be obtained by $R^2_{LR}$.

### 4.6.3 Correlation-Based Measures

In most program packages, summary measures that are based on the correlation between observations $y_i$ and fit $\hat{\pi}_i$ are given. One candidate is the (squared) correlation between $y_i$ and $\hat{\pi}_i$. Some motivation for the use of the squared correlation is that in the linear model and least-squares fitting the squared correlation is equivalent to the ratio of the explained variation $SSR/SST$, although the equivalence does not hold for GLMs. Zheng and Agresti (2000) prefer the correlation to its square because it has a familiar interpretation and works on the original scale of the observations.

Widely used association measures in logistic regression are rank-based measures. Concordance measures compare pairs of tupels $(y_i, \hat{\pi}_i)$ built from observation $y_i$ and the corresponding prediction $\hat{\pi}_i$. Let a pair be given by $(y_i, \hat{\pi}_i), (y_j, \hat{\pi}_j)$, where $y_i < y_j$. The pair is called *concordant* if the same ordering holds for the predictions, $\hat{\pi}_i < \hat{\pi}_j$; it is *discordant* if $\hat{\pi}_i > \hat{\pi}_j$ holds. The pair is *tied* if $\hat{\pi}_i = \hat{\pi}_j$. If $y_i = y_j$ holds, the pair is also tied. If a pair is concordant, discordant, or tied, it can be expressed by using the sign function: $\text{sign}(d) = 1$ if $d > 0$, $\text{sign}(d) = 0$ if $d = 0$, and $\text{sign}(d) = -1$ if $d < 0$. Then the product $\text{sign}(y_i - y_j)\,\text{sign}(\hat{\pi}_i - \hat{\pi}_j)$ takes value $1$ for concordant pairs and value $-1$ for discordant pairs. Several measures in common use are given in a separate box. The measures are also given in an alternative form since they can be expressed by using $N_c$ for the number of concordant pairs, $N_d$ for the number of discordant pairs, and $N$ for the number of pairs with different observations $y_i \neq y_j$.

**Rank-Based Measures**

Kendall's $\tau_a$

$$\begin{aligned}\tau_a &= \sum_{i<j} \text{sign}(y_i - y_j)\,\text{sign}(\hat{\pi}_i - \hat{\pi}_j)/(n(n-1)/2) \\ &= (N_c - N_d)/(n(n-1)/2),\end{aligned}$$

Somers' $D$

$$\begin{aligned}D_S &= \sum_{i<j} \text{sign}(y_i - y_j)\,\text{sign}(\hat{\pi}_i - \hat{\pi}_j) / \sum_{i<j} \text{sign}(y_i - y_j)^2 \\ &= (N_c - N_d)/N,\end{aligned}$$

Goodman and Kruskal's $\gamma$

$$\begin{aligned}\gamma &= \sum_{i<j} \text{sign}(y_i - y_j)\,\text{sign}(\hat{\pi}_i - \hat{\pi}_j) / \sum_{i<j} \text{sign}(y_i - y_j)^2\,\text{sign}(\hat{\pi}_i - \hat{\pi}_j)^2 \\ &= (N_c - N_d)/(N_c + N_d),\end{aligned}$$

A disadvantage of rank-based association measures is that in many applications they can not distinguish between different link functions. If the predicted values remain monotonic, link functions that yield quite different fits yield the same association value. The correlation coefficient used by Zheng and Agresti (2000) is able to distinguish between link functions because it uses the exact estimate, although it is more sensitive to outliers.

In general, association measures between observations and fit try to quantify how strong the link between $y_i$ and $\hat{\pi}_i$ is. They may be used as descriptive statistics, but one should be aware of what they are measuring. While the correlation coefficient measures linear dependence, rank-based procedures measure the association between ordered values and therefore monotonicity. They reflect in particular goodness-of fit in the sample. One should be cautious with squared values of association measures, considered, for example, by Mittlböck and Schemper (1996). In particular, for squared rank-based measures it is unclear what they are measuring. Moreover, one should not take them at face value because as descriptive statistics they are random variables depending on the sample, a fact that is often ignored. A more sensible approach first

defines a population measure and then finds ways to estimate it. For most useful measures there is an underlying population measure. For example, the empirical correlation coefficient has the theoretic analog $\mathrm{cor}(y, E(y|\boldsymbol{x}))$. Also, Kendall's $\tau_a$ and Somers' $D$ may be seen as estimates of an underlying measure (nicely described by Newson, 2002 ). The advantage of an underlying population value is that one can study the properties of the empirical measures as an estimator. That is a non-trivial task because one cannot expect $(y_i, \hat{\pi}_i)$ to be iid observations even when $(y_i, \boldsymbol{x}_i)$ are iid observations and the empirical measure will not be the best estimator for the population value. Zheng and Agresti (2000) used the population correlation coefficient $\mathrm{cor}(y, E(y|\boldsymbol{x}))$ and investigate bias and MSE for several estimators and show that in particular the cross-validation estimator has poor performance.

In various studies measures have been compared. For example, Mittlböck and Schemper (1996) investigated the performance of most of the measures considered here. They demonstrated that the squared correlation coefficient, $R^2_{SE}$, and $G$ are very similar over a wide range of values and are numerically consistent with the empirical coefficient of determination when a linear model is an appropriate approximation, which means for small values of $R^2$. Squared $\tau_a$, $\gamma$, and $R^2_{\text{class}}$ were found to yield quite different values. However, it cannot be expected that all the measures will behave in the same way since different population measures are behind. Numerical consistency or inconsistency is interesting, but in cases where the linear model is inappropriate nothing can be inferred on the accuracy of a descriptive statistic if what one is trying to estimate is not well defined.

Of course population values should have an intuitively clear interpretation. The Kullback-Leibler distance measure, which is behind the deviance, might not be very convincing for practioners. Simple measures with clear interpretations are measures that represent performance in classification. Behind the number of hits is the probability of a correct classification. Therefore, population measure and estimate refer to interesting quantities. But the number of hits refers to hits in the learning sample and therefore can be an overoptimistic estimate. To infer on the performance in future samples and find appropriate estimates, one should not rely on the empirical analog but find alternative estimators. Prediction-based measures and estimators are considered in more detail in Chapter 15.

General definitions of population measures that quantify the strength of the relation between a response variable and a vector of covariates have been given by Joe (1989), Osius (2004), Soofi et al. (2000). Van der Linde and Tutz (2008) considered $R^2$ measures derived from symmetric Kullback-Leibler discrepancies.

## 4.7 Further Reading

*Robust Estimators*. With the development of diagnostic tools for binary regression models (e.g., Pregibon, 1981; Landwehr et al., 1984; Fowlkes, 1987), estimates that are robust against outliers have been suggested. The resistant fitting procedure proposed by Pregibon (1982) is based on the downgrading of the influence of observations with high residuals. Copas (1988) considered the substantial bias of resistant fitting, which yields numerically larger coefficients, yielding a more extreme fit, closer to 0 or 1. He considered a bias-corrected version and proposed a misclassification model where transpositions between the possible outcomes 0 and 1 happen with a small probability. Carroll and Pederson (1993) studied an estimate that is closely related to Copas' misclassification estimate but which is consistent for the logistic model. Rousseeuw and Christmann (2003) considered estimates that are robust against separation and connected to the approach used by Tutz and Leitenstorfer (2006). An interesting approach to robust fitting by a forward search through the data was proposed by Atkinson and Riani (2000). An alternative form of robustification is the use of shrinkage estimators as considered in Chapter 6.

*Weighted least-squares Estimator*. Grizzle, Starmer, and Koch (1969) proposed an least-squares estimator for categorical responses. The so-called Grizzle-Starmer-Koch approach was very influential in the modeling of categorical response data. Although it allows for an explicit form of the estimate it has the disadvantage that it can not be used for continuous predictors. With the computational facilities available today ML estimates are widely preferred.

*R packages*. GLMs can be fitted by use of the model fitting functions *glm* from the *MASS* package. For the selection and smoothing of ordinal predictors one can use the package *ordPens*. Quasi-variances can be computed with *qvcalc*.

## 4.8 Exercises

4.1 Derive the likelihood of a binary response model $\pi(\boldsymbol{x}_i) = h(\boldsymbol{x}_i^T\boldsymbol{\beta})$ for independent observations $(y_i, \boldsymbol{x}_i), i = 1, \ldots, n$, with $y_i \in \{0, 1\}$.

(a) Find the derivatives $\partial l(\boldsymbol{\beta})/\partial\beta_j$ for the components of $\boldsymbol{\beta}$ by elementary differentiation.

(b) Show that the resulting score function $\boldsymbol{s}(\boldsymbol{\beta}) = \partial l(\boldsymbol{\beta})/\partial\boldsymbol{\beta}$ is that of a generalized linear model.

(c) Derive the entries of the matrix of second derivatives $\boldsymbol{H}(\boldsymbol{\beta}) = \partial l^2(\boldsymbol{\beta})/\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T$.

(d) Determine $\boldsymbol{F}(\boldsymbol{\beta}) = -\,\mathrm{E}\,\boldsymbol{H}(\boldsymbol{\beta})$ and show that it has the form of the Fisher matrix of a GLM.

4.2

(a) Derive the likelihood of a binary response model $P(y_{ij} = 1|\boldsymbol{x}_i) = \pi(\boldsymbol{x}_i) = h(\boldsymbol{x}_i^T\boldsymbol{\beta})$ for independent binomial observations $(y_i, \boldsymbol{x}_i), i = 1, \ldots, N$, with $y_i = y_{i1} + \cdots + y_{in_i} \sim B(n_i, \pi(\boldsymbol{x}_i))$.

(b) Show that the resulting likelihood equation for the logit model has the form $t = \mathrm{E}(T)$, where $T = \sum_i y_i\boldsymbol{x}_i$ is a sufficient statistic of $\boldsymbol{\beta}$ and $t$ is the observed value.

(c) Derive the sufficient statistic for the logit model when only two populations are distinguished, namely males ($x$ = 1) and females ($x$ = –1) and interpret the components of the statistic.

4.3 The following table shows a comparison of a new agent and an active control with the binary response 1 for much improvement and 0 for improvement (compare Example 9.8). The model to be considered is the logit model with predictor $\eta = \beta_0 + x_A\beta$, where $x_A = 1$ represents the new agent and $x_A = 0$ represents the active control.

| Drug | 1 | 0 |
|---|---|---|
| 1: New agent | 24 | 37 |
| 0: Active control | 11 | 51 |

(a) Fit the model by using the underlying single binary observations, that is, use 24 binary responses with response 1 and $x_A = 1$, etc, and examine the significance of the parameter estimates.

(b) Fit the model by using the binomial distribution, where one considers only two populations, the first with 61 observations the second with 62 observations. Compare with the results from (a)

(c) Compute the null deviance $D_0$ (deviance for the model with intercept only) and the deviance for the model $D_M$ for the single and grouped observations cases. How can the difference of deviances be used to examine the goodness-of-fit of the model?

4.4 Consider a binary logit model with two factors $A \in \{1, \ldots, I\}$, $B \in \{1, \ldots, J\}$. It can be given by $\log(\pi(A = i, B = j))/(1 - \pi(A = i, B = j)) = \beta_0 + \beta_{A(i)} + \beta_{B(j)} + \beta_{AB(ij)}$ with side constraints $\beta_{A(I)} = \beta_{B(J)} = \beta_{AB(Ij)} = \beta_{AB(iJ)} = 0$ for all $i, j$ or with linear predictor $\eta(A = i, B = j) = \beta_0 + x_{A(1)}\beta_{A(1)} + \cdots + x_{A(I-1)}\beta_{A(I-1)} + x_{B(1)}\beta_{B(1)} + \cdots + x_{B(J-1)}\beta_{B(J-1)} + x_{A(1)}x_{B(1)}\beta_{AB(11)} + \cdots + x_{A(I-1)}x_{B(J-1)}\beta_{AB(I-1,J-1)}$, where $x_{A(i)}, x_{B(j)}$ are dummy variables in 0–1 coding.

(a) Show how the parameters can be represented as functions of odds and odds ratios (one obtains, for example, $e^{\beta_{AB(ij)}} = \log(\gamma(i,j)/\gamma(i,J))/\gamma(I,j)/\gamma(I,J)$, where $\gamma(i,j) = \pi(A=i, B=j)/(1-\pi(A=i,B=j))$.

(b) Interpret the parameters in terms of odds ratios.

(c) Let the probability of being a regular reader of a specific journal depending on gender and age be given by the following table:

| | | Age | |
|---|---|---|---|
| | | Young (1) | Old (0) |
| Gender | Male (1) | 0.5 | 0.4 |
| | Female (0) | 0.8 | 0.6 |

Compute the parameters of the corresponding logit model and interpret them.

(d) Compute the parameter estimates of a saturated logit model for the data in Table 4.9 and interpret them.

4.5 The dataset *dust* that was used in Example 4.6 is available from package *catdata* (or at http://www.stat.uni-muenchen.de/sfb386/ under the name "Chronic bronchitis and dust concentration").

(a) Fit models for non-smokers and smokers separately that include quadratic terms of dust concentration and years of exposure. Decide what effects are needed. Consider alternatively models that use log-transformed predictors. Compare the fitted models.

(b) Fit an appropriate model that includes smoker status as a predictor.

4.6 Table 2.2 shows the vasoconstriction dataset.

(a) Compare the logit model with predictors volume and rate to the logit model with log-transformed predictors.

(b) Investigate if an interaction term is needed.

(c) Consider binary response models with alternative link functions and decide on an appropriate model.

4.7 Show that for binary observations the proportional reduction in variance $(D(M_0) - D(M))/D(M_0)$ is equivalent to Gini's concentration measure $G = (\sum_i \hat{\pi}_i^2 - n\bar{p}^2)/(n\bar{p}(1-\bar{p}))$ if one uses the empirical variance $D(M_0) = n\bar{p}(1-\bar{p})$, where $\bar{p}$ is the proportion of ones in the total sample, and $D(M) = \sum_i \hat{\pi}_i(1-\hat{\pi}_i)$.

# Chapter 5

# Alternative Binary Regression Models

In this chapter we will first consider alternatives to the logit link. Although the logit model has some advantages, which are discussed in Section 5.1.3, alternative link functions may be more appropriate in concrete applications. Moreover, we will consider extensions of the simple binary regression model that allow for overdispersion in the response and the conditional likelihood approach.

## 5.1 Alternative Links in Binary Regression

As in Chapter 4, we will consider models of the form $\pi(\boldsymbol{x}) = h(\boldsymbol{x}^T\boldsymbol{\beta})$, but in this section we will denote them by

$$\pi(\boldsymbol{x}) = F(\boldsymbol{x}^T\boldsymbol{\beta}).$$

The use of $F$ for the response function refers to the derivation of the models from latent variables models, where $F$ is the distribution function of the latent variable. Therefore, the response functions used here are strictly monotone distribution functions. The inverse functions $F^{-1}$ correspond to the link. In the following, common choices for link and response functions are motivated.

### 5.1.1 Binary Response Models

#### Probit Model

A widely used model, particularly in economics, is the probit model, which is based on the standard normal distribution $\phi(\eta) = (2\pi)^{-1/2}\int_{-\infty}^{\eta} e^{-x^2/2}dx$.

**Probit Model**

$$\pi(\boldsymbol{x}) = \phi(\boldsymbol{x}^T\boldsymbol{\beta}), \quad \phi^{-1}(\pi(\boldsymbol{x})) = \boldsymbol{x}^T\boldsymbol{\beta}$$

In applications the probit model usually yields approximately the same results as the logit model. The goodness-of-fit is comparable, the same variables turn out to be relevant, and $p-$values are about the same, although the values of the estimates should not be compared directly (see Section 5.1.2). Very large sample sizes are needed to distinguish between the logit

and the probit model. One may consider it as a drawback that the response function has no explicit form and that parameters do not have the same simple interpretation in terms of log-odds as in the logit model; nevertheless, the results are similar.

### Complementary Log-Log Model and Log-Log Model

A distribution function that is distinctly different from the logistic distribution function is the *minimum extreme value* (or *Gompertz*) distribution $F(\eta) = 1 - \exp(-\exp(\eta))$. While the logistic distribution function is symmetric, the Gompertz distribution is asymmetric (see Figure 5.1 for the density distribution function). The model has the following representations.

**Complementary Log-Log Model**

$$\pi(\boldsymbol{x}) = 1 - \exp(-\exp(\boldsymbol{x}^T\boldsymbol{\beta})) \qquad \log(-\log(1-\pi(\boldsymbol{x}))) = \boldsymbol{x}^T\boldsymbol{\beta} \quad (5.1)$$

The name complementary log-log model derives from the second form, where one sees that the link is log-log effecting the complementary probability $1 - \pi(\boldsymbol{x})$.

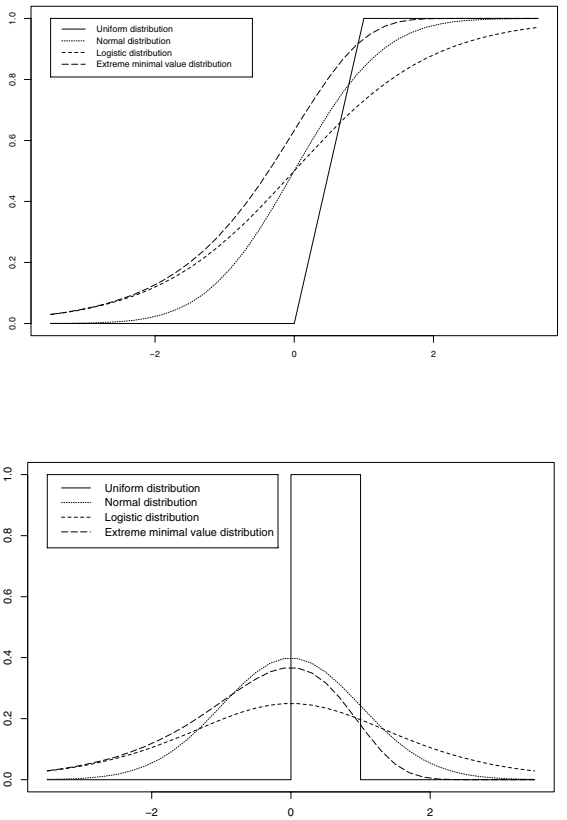


FIGURE 5.1: Response functions that correspond to distribution functions. Upper panel shows distribution functions of uniform distribution, normal distribution, logistic distribution, and minimum extreme value distribution; lower panel shows the corresponding densities.

A closely related model is the log-log model. The complementary log-log model (5.1) models $\pi(\boldsymbol{x}) = P(y = 1|\boldsymbol{x})$. Since the two possible values of $y, y = 1$ and $y = 0$, may be interchanged, one might use model (5.1) as well for the response $y = 0$, assuming $1 - \pi(\boldsymbol{x}) = 1 - \exp(-\exp(\boldsymbol{x}^T\boldsymbol{\beta}))$, which yields the log-log model.

**Log-Log Model**

$$\pi(\boldsymbol{x}) = \exp(-\exp(-\boldsymbol{x}^T\boldsymbol{\beta})) \qquad \log(-\log(\pi(\boldsymbol{x}))) = -\boldsymbol{x}^T\boldsymbol{\beta}$$

The use of $-\boldsymbol{x}^T\boldsymbol{\beta}$ rather than $\boldsymbol{x}^T\boldsymbol{\beta}$ (which is only important for the interpretation of parameters) has the advantage that the model has the form $\pi(\boldsymbol{x}) = F(\boldsymbol{x}^T\boldsymbol{\beta}))$, where $F$ is the *maximum value* (or *Gumbel*) distribution $F(\eta) = \exp(-\exp(-\eta))$.

For symmetrical distributions like the logistic distribution it is just a matter of taste if one uses response $y$ or the transformed response $\tilde{y} = 1 - y$. If $y = 1$ corresponds to success, $\tilde{y} = 1$ corresponds to failure. If $\boldsymbol{\beta}$ is the parameter vector of the logistic model when modeling success, $-\boldsymbol{\beta}$ is the parameter vector when modeling failure. Gomertz and Gumbel distributions are not symmetrical. They are connected in the following way: If a random variable $\varepsilon$ follows the Gompertz distribution $F(\eta)$, the variable $-\varepsilon$ follows the Gumbel distribution $1 - F(-\eta)$, and vice versa. If $\boldsymbol{\beta}$ is the parameter vector of the Gompertz model for response $y$, $-\boldsymbol{\beta}$ is the parameter vector of the Gumbel model for response $\tilde{y}$. However, the Gumbel and the Gompertz models for fixed response $y$ are not equivalent. Goodness-of-fit as well as parameter vectors will differ.

### Exponential Model

Suppose that $Z$ is a count variable taking values $0, 1, 2 \ldots$, which may refer to the number of cars in a household or in medicine to the number of symptoms. An often useful approximation to the distribution of $Z$ is the Poisson distribution that has probability function $P(Z = z) = e^{-\lambda}\lambda^z/z!, z = 0, 1, 2, \ldots$, where $\lambda$ represents the expectation of $Z$. If one is interested only in the dichotomization $Z = 0$ or $Z > 0$ (no car in household versus at least one car in household), one obtains for the dichotomous variable

$$y = \begin{cases} 1 & Z > 0 \\ 0 & Z = 0 \end{cases}$$

that

$$P(y = 1) = P(Z > 0) = 1 - P(Z = 0) = 1 - e^{-\lambda}.$$

By assuming that the expectation $\lambda$ depends on the covariates in a linear way, $\lambda = \boldsymbol{x}^T\boldsymbol{\beta}$, one obtains the exponential model. Of course the complementary log-log model also can be motivated in this way, because the specification $\lambda = \exp(\boldsymbol{x}^T\boldsymbol{\beta})$ yields the complementary log-log model.

**Exponential Distribution or Complementary Log Model**

$$\pi(\boldsymbol{x}) = 1 - \exp(-\boldsymbol{x}^T\boldsymbol{\beta}) \qquad -\log(1-\pi(\boldsymbol{x})) = \boldsymbol{x}^T\boldsymbol{\beta}$$

Since the Poisson distribution is strongly connected to the exponential distribution, which is the waiting time until the next event in a Poisson process, it is not surprising that the exponential distribution model may also be motivated by a waiting time distribution model. Let a duration time (survival or sojourn time) $T$ have exponential distribution $T \sim E(\lambda)$ with distribution function $F(x) = 1 - \exp(-\lambda x), \lambda > 0$. At time point $\tau$ one considers the dichotomization

$$y = \begin{cases} 1 & T < \tau \\ 0 & T \geq \tau, \end{cases}$$

which determines if the process $T$ has ended or not until then; one obtains with parameterization $\lambda = \boldsymbol{x}^T\boldsymbol{\gamma}$

$$P(y=1) = P(T < \tau) = 1 - \exp(-\tau\boldsymbol{x}^T\boldsymbol{\gamma}) = 1 - \exp(-\boldsymbol{x}^T\boldsymbol{\beta}),$$

where $\boldsymbol{\beta} = \tau\boldsymbol{\gamma}$.

The motivation by dichotomizations of counts or waiting times shows that the model might be appropriate in many applications. Sometimes a problem is the non-convergence of estimates. It may arise because the link function is not differentiable everywhere. Since $F(x) = 1 - \exp(-\lambda x), x \geq 0$, the model actually has the form $\pi(\boldsymbol{x}) = 1 - \exp(-\boldsymbol{x}^T\boldsymbol{\beta})$ if $\boldsymbol{x}^T\boldsymbol{\beta} \geq 0$ and $\pi(\boldsymbol{x}) = 0$ if $\boldsymbol{x}^T\boldsymbol{\beta} < 0$. The simple form $\pi(\boldsymbol{x}) = 1 - \exp(-\boldsymbol{x}^T\boldsymbol{\beta})$ holds only if $\boldsymbol{x}^T\boldsymbol{\beta} \geq 0$, which implies severe restrictions on the parameter space. For discussions and applications see Wacholder (1986), Baumgarten et al. (1989), Guess and Crump (1978), Whittemore (1983), and Cornell and Speckman (1967). A nice overview on modeling with this link function is found in Piegorsch (1992).

If $\pi(x)$ is replaced by $1-\pi(x)$, one obtains the simple log-link or exponential model, which may be seen as a model that uses the distribution function $F(\eta) = \exp(\eta)$ for $\eta \leq 0$.

**Exponential or Log-Link Model**

$$\pi(\boldsymbol{x}) = \exp(\boldsymbol{x}^T\boldsymbol{\beta}) \qquad \log(\pi(\boldsymbol{x})) = \boldsymbol{x}^T\boldsymbol{\beta}$$

## Cauchy Model

A model that is offered by some program packages is based on the Cauchy distribution. The "cauchit" link uses the (standard) Cauchy distribution function $F(\eta) = \tan^{-1}(\eta)/\pi + 1/2$, where $\tan^{-1} = \arctan$ is the inverse of the tangens and $\pi = 3.14159\ldots$. The Cauchy distribution is somewhat peculiar because it has no mean, variance, or higher moments defined, although the mode and median are well defined and are both equal to zero. It coincides with the Student's $t$-distribution with one degree of freedom. The cauchit link function $g(u) = \tan(\pi(u - 1/2))$ yields the following model.

**Cauchy Model**

$$\pi(\boldsymbol{x}) = \tan^{-1}(\boldsymbol{x}^T\boldsymbol{\beta})/\pi + 1/2 \qquad \tan(\pi(\pi(\boldsymbol{x}) - \frac{1}{2})) = \boldsymbol{x}^T\boldsymbol{\beta} \quad (5.2)$$

It should be noted that $\pi(\boldsymbol{x})$ denotes the probability whereas $\pi$ in (5.2) it denotes the fixed and well-known number $\pi = 3.14159\ldots$. When compared to the normal distribution, the Cauchy distribution has heavier tails, thus allowing more extreme values than the normal distribution. For the modeling of binary responses it is attractive when observations occur for which the linear predictor is large in absolute value, indicating that the outcome is rather certain and yet the outcome is different. The model is more tolerant to these "outliers" than the logit or probit model. An early reference to the cauchit link model is Morgan and Smith (1993), where an example is given in which cauchit performs better than the probit link.

### Identity Link Model

In a normal regression model the most widely used link is the identity link yielding $\mathrm{E}(y_i) = \boldsymbol{x}_i^T\boldsymbol{\beta}$. With some care it can also be used in binary regressions. However, the assumption $\pi(\boldsymbol{x}_i) = \boldsymbol{x}_i^T\boldsymbol{\beta}$ distinctly ignores that $\pi(\boldsymbol{x})$ is restricted to the unit interval. Nevertheless, it may be applied in cases where the covariate space is strongly limited in a way that the restriction to the unit interval holds. In Example 4.1 the logistic model yields an almost straight line and the logistic model and the linear model yield similar results (see Figure 4.3).

**Identity Link Model**

$$\pi(\boldsymbol{x}) = \boldsymbol{x}^T\boldsymbol{\beta}$$

## 5.1.2 Comparing Link Functions

The models considered in this section have the basic form $\pi(\boldsymbol{x}_i) = F(\boldsymbol{x}_i^T\boldsymbol{\beta})$, where $F$ is a distribution function. Figure 5.1 shows the response functions for several of these models. At first sight the response functions seem to differ rather strongly and therefore should yield quite different discrepancies between data and fits. However, one should be aware that the distribution functions used as links are not comparable because they refer to different means and variances. For example, the standard normal distribution that is used in the normal model has mean zero and variance one while the logistic distribution (underlying the logistic regression model) has mean zero and variance $\pi^2$ (where $\pi = 3.14159\ldots$). Thus, it is no wonder that parameter estimates for the probit model and the logit model usually are quite different (although having about the same exploratory value). It is useful to consider again the derivation of binary regression models from latent regression models. In Section 2.2.2, it has been shown that all models of the form $\pi(\boldsymbol{x}_i) = F(\beta_0 + \boldsymbol{x}_i^T\boldsymbol{\beta})$ (with separated intercept) may be derived from an underlying continuous response model, $\tilde{y}_i = \gamma_0 + \boldsymbol{x}_i^T\boldsymbol{\beta} - \varepsilon_i$, where $\varepsilon_i$ has the distribution function $F$ and $y_i = 1$ if $\tilde{y}_i$ is above some threshold $\theta$. One obtains

$$y_i = 1 \quad \Leftrightarrow \quad \tilde{y}_i = \gamma_0 + \boldsymbol{x}_i^T\boldsymbol{\beta} - \varepsilon_i \geq \theta \quad \Leftrightarrow \quad \varepsilon_i \leq \beta_0 + \boldsymbol{x}_i^T\boldsymbol{\beta},$$

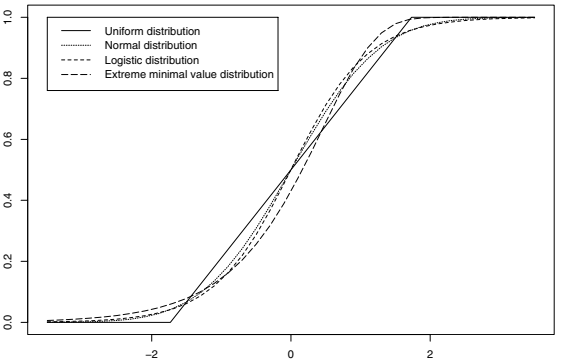


FIGURE 5.2: Response functions, standardized to mean zero and variances one, for several models.

where $\beta_0 = \gamma_0 - \theta$. If one wants to compare parameters of models with different link functions, one should at least assume that the distribution functions $\varepsilon_i$ have the same mean and variance. Standardization of $\varepsilon_i$ yields

$$y_i = 1 \Leftrightarrow \quad \frac{\varepsilon_i - E(\varepsilon_i)}{\sqrt{var(\varepsilon_i)}} \leq \frac{\gamma_0 - \theta - E(\varepsilon_i)}{\sqrt{var(\varepsilon_i)}} + x_{i1}\frac{\beta_1}{\sqrt{var(\varepsilon_i)}} + ... + x_{ip}\frac{\beta_p}{\sqrt{var(\varepsilon_i)}}$$

with the "standardized" parameters

$$\tilde{\beta}_0 = \frac{\beta_0 - E(\varepsilon_i)}{\sqrt{var(\varepsilon_i)}}, \quad \tilde{\beta}_i = \frac{\beta_i}{\sqrt{var(\varepsilon_i)}}.$$

With $F_{\text{stand}}(\eta)$ denoting the standardized distribution function (centered around zero with variance one), the models

$$\pi(\boldsymbol{x}) = F(\beta_0 + \boldsymbol{x}^T\boldsymbol{\beta}) \quad \text{and} \quad \pi(\boldsymbol{x}) = F_{\text{stand}}(\tilde{\beta}_0 + \boldsymbol{x}^T\tilde{\boldsymbol{\beta}})$$

are equivalent. It is seen from Figure 5.2 that the standardized response (distribution) functions are not so far apart. In particular, the normal and the logistic distribution functions are quite close. Therefore, it is quite natural that they yield similar goodness-of-fit, although parameter estimates for the unstandardized versions will differ.

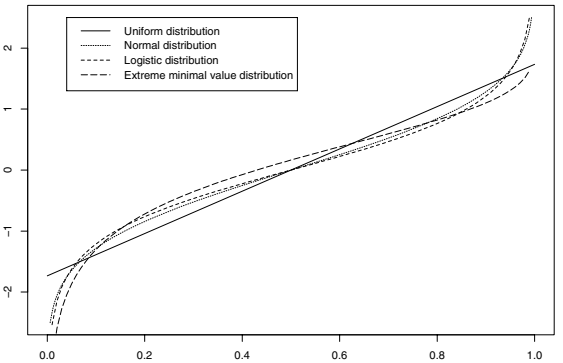


FIGURE 5.3: Link functions $F^{-1}(\pi)$ plotted against $\pi$.

Table 5.2 shows the fit of several binary response models for the response "car in household" with net income in the linear predictor. It is seen that the original estimates $\hat{\beta}$ are quite different but the standardized values are similar. In this dataset the Cauchy model showed the best fit, followed by the logit model.

TABLE 5.1: Means and variances of $\varepsilon$ for several models.

| Distrib. | Logit Logistic | Probit Normal | Complementary Log-Log Min. Extr. Value | Log-Log Max. Extr. Value | Compl. Exp Exponential |
|---|---|---|---|---|---|
| Mean | $0$ | $0$ | $-0.577$ | $0.577$ | $1$ |
| Variance | $\pi^2/3$ | $1$ | $\pi^2/6$ | $\pi^2/6$ | $1$ |

TABLE 5.2: Estimates and standardized estimates for several link functions modeling the dependence of car ownership on net income.

| | $\hat{\beta}_0$ | $\tilde{\beta}_0$(standardized) | $\hat{\beta}$ | $\tilde{\beta}$(standardized) | Deviance |
|---|---|---|---|---|---|
| Logit | $-2.424$ | $-1.334$ | $0.0020$ | $0.0011$ | $1497.7$ |
| Probit | $-1.399$ | $-1.399$ | $0.0011$ | $0.0011$ | $1505.0$ |
| C.log-log | $-1.743$ | $-0.909$ | $0.0010$ | $0.0008$ | $1538.2$ |
| Cauchy | $-2.655$ | | $0.0023$ | | $1479.7$ |

### 5.1.3 Choice between Models and Advantages of Logit Models

For the choice of the link function, in particular, two aspects are relevant: goodness-of-fit and ease of interpretation. If one focusses on goodness-of-fit, one simply chooses the model that shows the best fit. Nevertheless, one always should be aware that only a selection of possible fits has been considered, so the best fit is only the best fit among the considered ones. In addition, differences of deviances have no standard distribution since there is no hierarchical order between the models. If responses are binomially distributed with not too small sample sizes $n_i$, the deviance can be compared to the $\chi^2$-distribution showing if the models have a distinct lack-of-fit. An alternative approach to the selection of link functions is to estimate the fit in a non-parametric way. One can also fit parametric families of link functions that include the classical link functions as members. Then parameter estimates decide on the specific link function (for both approaches see Section 5.2).

If the deviances are not too different, a result that is usually found when comparing the logit model with the probit model, issues of interpretation become dominant. The logit model has several advantages that make it the most widely used model:

(1) The parameters are easily interpretable in terms of log-odds (for $\boldsymbol{\beta}$) or odds (for $\exp(\boldsymbol{\beta})$).

(2) If all of the covariates are categorical, the hypothesis $H_0 : \boldsymbol{\beta} = \mathbf{0}$ that none of the covariates has any exploratory value is equivalent to the statement that the response variable and the covariates are independent. For single predictors the hypothesis $H_0 : \beta_j = 0$ corresponds to the conditional independence of the response given the other variables (compare Chapter 12).

(3) The model is linked to the normal distribution, which plays a central role in statistics, by the derivation from the assumption $\boldsymbol{x}|y = i \sim N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$; see Section 2.2.2.

(4) The logit link is the canonical link function, linking the natural parameter $\log(\pi/(1-\pi))$ directly to the linear predictor. One consequence is that the conditions for the existence of maximum likelihood estimates are weaker because the log-likelihood function is concave under weak conditions (see Fahrmeir and Kaufmann, 1985).

(5) The effects of variables may also be estimated if observations are drawn from the conditional distribution of $\boldsymbol{x}$ given $y$ instead of the usual form of observing responses $y$ given $x$. In econometrics this is called choice-based sampling (see Section 4.1).

Some motivation for the logit model may also be obtained from looking at the origins of the logistic distribution function (see Section 2.4).

## 5.2 The Missing Link

In Section 5.1 binary regression models with known link function were considered. Although various link functions may be used, there is always the danger that the true link function might not be among them or, more realistically, that none of the link functions provides a sufficiently good approximation to the data. This may be important since it has been demonstrated that misspecification of the link function can lead to substantial bias in the regression parameters (see Czado and Santner, 1992 , for binomial responses). Similar results have been demonstrated for single-index models by Horowitz and Härdle (1996).

### Parametric Families of Link Functions

One approach to obtain more flexible models is to fit the parametric families of link functions. Several families have been proposed, including the link functions, that are in common use; see Prentice (1976), Pregibon (1980), Aranda-Ordaz (1983), Morgan (1985), Stukel (1988), Czado (1992), and Czado (1997). When using families of response functions, the common response function $F(\eta)$ is replaced by $F(\eta, \psi)$ with additional parameter $\psi$. As an example, let us consider a useful family allowing for right-tail modification of the logistic link (see Czado, 1997):

$$F(\eta, \psi) = \frac{\exp(h(\eta, \psi))}{1 + \exp(h(\eta, \psi))},$$

where

$$h(\eta, \psi) = \begin{cases} \frac{(\eta+1)^{\psi}-1}{\psi} & \eta > 0 \\ \eta & \text{otherwise.} \end{cases}$$

If $\psi = 1$, the logistic link results; for $\psi < 1$ ($\psi > 1$) the right tail is heavier (lighter) than for the logistic distribution. Families of this type have the advantage that the parameterization is orthogonal in a neighbourhood around $\boldsymbol{\beta} = \mathbf{0}$. When using families of response functions there is usually a parameter $\psi^*$ that corresponds to the canonical link. In the case of the right-tail modification family one has $\psi^* = 1$.

A common approach to decide on the link is to treat it as a testing problem:

$$H_0 : \psi = \psi^* \quad \text{against} \quad H_1 : \psi \neq \psi^*.$$

If $H_0$ is not rejected, one keeps the canonical link; if $H_0$ is rejected, $\psi$ and the regression parameters are estimated jointly to obtain the MLE $\hat{\boldsymbol{\delta}} = (\hat{\boldsymbol{\beta}}, \hat{\psi})$. Asymptotic theory, including strong

consistency and asymptotic normal distributions was derived by Czado and Munk (2000). The use of the testing problem as a tool of model selection has been critized by various authors. In particular, Czado and Munk (2000) point out that for large sample sizes $H_0$ is frequently rejected in favor of $H_1$, although the mean space of both link functions is almost indistinguishable and therefore not scientifically relevant. They propose an alternative testing strategy that takes a measure of discrepancy between the response functions into account.

## Non-Parametric Fitting of Link Functions

Families of link functions have the advantage that the link function is estimated by using parameters, and by fitting a larger family one avoids having to estimate separately non-nested models. A disadvantage is that the functions are still restricted to belong to the specified family. More flexible models are obtained by estimating link functions non-parametrically.

Especially in the economic literature, models of this type are known under the name single-index models. A *single-index model* has the form

$$\mu_i = h(\boldsymbol{x}_i^T \boldsymbol{\beta}),$$

where $h$ is a smooth but *unspecified* function and $\mu_i = E(y_i|\boldsymbol{x}_i)$ denotes the conditional mean given covariates $\boldsymbol{x}_i$. The linear predictor $\boldsymbol{x}_i^T \boldsymbol{\beta}$ is known as the *single index*. The model may be seen as a special case of a *projection pursuit regression*, which assumes that $\mu_i$ has the additive form $h_1(\boldsymbol{x}_i^T \boldsymbol{\beta}_1) + \cdots + h_m(\boldsymbol{x}_i^T \boldsymbol{\beta}_m)$ with unknown functions $h_1, \ldots, h_m$, which transform the indices; see Friedman and Stützle (1981). Several approaches to the estimation of single-index models have been proposed in the literature (for references see the end of the chapter). However, in single-index models typically the response is assumed to be metrically scaled, and often a normal distribution is assumed. Moreover, single-index models do not assume that the function $h(.)$ is monotone. Therefore, the approaches are less helpful when one wants to estimate the unknown link function in generalized models. Although monotonicity is not needed when searching for a single index $\boldsymbol{x}_i^T \boldsymbol{\beta}$, it is useful for interpreting the parameters. If the link function $h(.)$ is not monotone, it is hard to interpret the parameters because a positive coefficient might increase or decrease the mean depending on the value of the other predictors. Thus, what might be helpful for dimension reduction is less helpful for fitting models that have easy interpretations.

Estimation of the unknown link function when the underlying distribution is from a simple exponential family was considered, for example, by Weisberg and Welsh (1994), Ruckstuhl and Welsh (1999), and Muggeo and Ferrara (2008). Weisberg and Welsh (1994) proposed estimating regression coefficients using the canonical link and then estimating the link via kernel smoothers given the estimated parameters. Then the parameters are re-estimated. Alternating between estimation of link and parameters yields consistent estimates.

The basic principle of alternating between these two estimates was also used by Yu and Ruppert (2002), but instead of kernel smoothers the unknown function is approximated by an expansion in basis functions. The approach may be outlined briefly as follows. For the unknown link function one uses the expansion

$$h(\eta_i) = \sum_{j=1}^{m} \alpha_j \phi_j(\eta_i),$$

where $\eta_i = \boldsymbol{x}_i^T \boldsymbol{\beta}$. To make the problem identifiable, $||\boldsymbol{\beta}|| = 1$ is postulated, where $||.||$ denotes the Euclidean norm. Moreover, $\boldsymbol{\beta}$ contains no intercept; it is included in $h(.)$. Splines are obtained by using the truncated power series basis functions $B_j(.)$ of degree $q$, which have

also been used, for example, by Ruppert (2002). Thus the functions have the form $B_1(\eta) = 1, B_2(\eta) = \eta, B_{q+1}(\eta) = \eta^q$, $B_{q+j}(\eta) = |\eta - \tau_j|_+, j > 1$, where $\tau_1, \tau_2, \ldots$ are fixed knots. In a P-spline regression (see Section 10.1.3), usually a rather high number of equidistant knots is used (say $m = 20$ or $40$) and the smoothness of the function estimate is controlled by an appropriate penalization. Ruppert (2002) suggested penalizing the squared coefficients that belong to the truncated powers, that is, $\sum_{j=q+2}^{m} \alpha_j^2$.

Let the response vector be given by $\mathbf{y}^T = (y_1, \ldots, y_n)$ and the design matrix by $\mathbf{X} = (\mathbf{x}_{(1)}, \ldots, \mathbf{x}_{(p)})$, where $\mathbf{x}_{(j)} = (x_{1j}, \ldots, x_{nj})^T$ denotes the observations of the $j$th covariate, $j = 1, \ldots, p$. Then, in the simplest case of normally distributed responses, an estimator of the single-index model is formulated as a minimizer of the penalized least-squares criterion:

$$Q(\boldsymbol{\alpha}, \boldsymbol{\beta}) = (\mathbf{y} - \boldsymbol{\phi}(\boldsymbol{\eta})\boldsymbol{\alpha})^T(\mathbf{y} - \boldsymbol{\phi}(\boldsymbol{\eta})\boldsymbol{\alpha}) + \lambda_P \boldsymbol{\alpha}^T \mathbf{P} \boldsymbol{\alpha}, \tag{5.3}$$

where $\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta}$, $\boldsymbol{\phi}(\boldsymbol{\eta}) = (\phi_1(\boldsymbol{\eta}), \ldots, \phi_m(\boldsymbol{\eta})) = (\mathbf{1}, \boldsymbol{\eta}, \ldots, \boldsymbol{\eta}^q, (\boldsymbol{\eta} - \mathbf{1}\tau_1)_+^q, \ldots, (\boldsymbol{\eta} - \mathbf{1}\tau_m)_+^q)$, $\boldsymbol{\phi}_j(\boldsymbol{\eta}) = (\phi_j(\eta_1), \ldots, \phi_j(\eta_n))^T$, $\mathbf{P} = \text{diag}\{\mathbf{0}_{q+1}, \mathbf{1}_m\}$, and $\lambda_P$ is a penalization parameter. Yu and Ruppert (2002) suggest solving (5.3) by using common non-linear least-squares routines while Leitenstorfer and Tutz (2011) and Tutz and Petry (2011) used boosting techniques.

In the case of GLMs, it is more appropriate to fit the model $\mu_i = h_0(h(\eta_i))$, where $h_0(.)$ is a fixed transformation function, which has to be chosen, and the inner function $h(.)$ is considered as unknown and has to be estimated. Typically, the choice of $h_0(.)$ depends on the distribution of the response. When the response is binary, a canonical choice is the logistic distribution function. The main advantage of specifying a fixed link function is that it may be selected such that the predictor is automatically mapped into the admissible range of the mean response. The expansion in basis functions is applied to the inner function $h(.)$.

Estimates are obtained by iteratively estimating the regression coefficients $\boldsymbol{\beta}$ and the parameters of the link function $\boldsymbol{\alpha}$, where $h(\eta_i) = \sum_{j=1}^{m} \alpha_j \phi_j(\eta_i) = \boldsymbol{\alpha}^T \boldsymbol{\Phi}_i$. In matrix notation, let $\hat{\boldsymbol{\beta}}^{(l)}$ and $\hat{\boldsymbol{\eta}}^{(l)} = \boldsymbol{X}\hat{\boldsymbol{\beta}}^{(l)}$ denote the parameter estimate and the fitted predictor in the $l$th step. Moreover, $\boldsymbol{\Phi}^{(l)} = (\boldsymbol{\Phi}_1^{(l)}, \ldots, \boldsymbol{\Phi}_n^{(l)})^T$ with $\boldsymbol{\Phi}_i^{(l)} = (\phi_1(\hat{\eta}_i^{(l)}), \ldots, \phi_m(\hat{\eta}_i^{(l)}))^T$ is the current design matrix for the basis functions. Then two steps are iterated:

*Estimation of basis coefficients for a fixed predictor.* For a fixed predictor $\hat{\boldsymbol{\eta}}^{(l-1)} = \mathbf{X}\hat{\boldsymbol{\beta}}^{(l-1)}$ the model $\boldsymbol{\mu} = h_0((\Phi^{(l-1)})^T \boldsymbol{\alpha}^{(l)})$ is fitted by one step of penalized Fisher scoring that uses the matrix of derivations $\hat{\mathbf{D}}^{(l-1)} = \text{diag}(\partial h_0(\hat{h}^{(l-1)}(\hat{\eta}_i^{(l-1)}))/\partial h^{(l-1)}(\eta))$ evaluated at the estimate of the previous step, the diagonal matrix of variances evaluated at $h_0(\hat{h}^{(l-1)}(\eta_i^{(l-1)}))$, and a penalty matrix $\mathbf{P}_h$ that penalizes the second derivation of the estimated (approximated) response function (for penalties see Section 10.1.3).

*Estimation of regression coefficients for a fixed response function.* For $h(.)$ fixed, one fits the model $\boldsymbol{\mu} = h_0(h(\mathbf{X}\boldsymbol{\beta}^{(l)})$. Fisher scoring has the form

$$\hat{\boldsymbol{\beta}}^{(l)} = (\boldsymbol{X}^T \hat{\boldsymbol{D}}_\eta^{(l-1)} (\hat{\boldsymbol{\Sigma}}^{(l-1)})^{-1} \hat{\boldsymbol{D}}_\eta^{(l-1)} \boldsymbol{X})^{-1} \boldsymbol{X}^T \hat{\boldsymbol{D}}_\eta^{(l-1)} (\hat{\boldsymbol{\Sigma}}^{(l-1)})^{-1} (\boldsymbol{y} - \hat{\boldsymbol{\mu}}^{(l-1)}), \tag{5.4}$$

where $\hat{\boldsymbol{D}}_\eta^{(l-1)} = \text{diag}(\partial h_0(\hat{h}^{(l-1)}(\hat{\eta}_i^{(l-1)}))/\partial\eta)$ is the matrix of derivatives evaluated at the values of the previous iteration and $\hat{\boldsymbol{\Sigma}}^{(l-1)}$ is the variance from the previous step.

The second step can be modified to include a selection step that includes the most relevant predictor within a boosting procedure; see Section 6.4, where more details are given.

## 5.3 Overdispersion

In practice it is not too rarely found that models have large deviance although there seems to be no systematic lack-of-fit. Additional noise that is not accounted for may make the responses

more variable than is to be expected under the assumed distribution model. The data show *overdispersion*. Although underdispersion, which signals lower variability than expected, is also found, it is much rarer. There are several strategies for dealing with overdispersion, but first we consider potential sources for the phenomenon in binary data.

## 5.3.1 Sources of Overdispersion

### Correlated Observations

When considering binomial data $y_i = y_{i1} + \cdots + y_{in_i} \sim B(n_i, \pi_i)$ in prevous sections it was assumed that $y_{11}, \ldots, y_{1n_1}, y_{21}, \ldots, y_{Nn_N}$ are independent given the covariates. In particular, if measurements $y_{i1}, \ldots, y_{in_i}$ are collected at one unit, this assumption may be violated.

If one assumes that $y_{i1}, \ldots, y_{in_i}, y_{ij} \sim B(1, \pi_i)$ are correlated, one obtains

$$\text{var}(y_i) = var(\sum_{j=1}^{n_i} y_{ij}) = \sum_{j=1}^{n_i} \text{var}(y_{ij}) + \sum_{r \neq s} \text{cov}(y_{ir}, y_{is}).$$

By using $\text{var}(y_{ij}) = \pi_i(1-\pi_i), \text{cov}(y_{ir}, y_{is}) = \rho(\text{var}(y_{ir})\,\text{var}(y_{is}))^{1/2}$ with $\rho$ denoting the correlation coefficient, one has

$$\text{var}(y_i) = n_i\pi_i(1-\pi_i)[1 + (n_i - 1)\rho] = n_i\pi_i(1-\pi_i)\phi_i$$

with dispersion parameter $\phi_i = 1 + (n_i - 1)\rho$. The resulting effects are

- for $n_i = 1$ one has no overdispersion, since $\phi_i = 1$;
- for a positive correlation, $\rho > 0$, the variability is larger than expected under the binomial probability model (provided $n_i > 1$);
- for a negative correlation, $\rho < 0$, underdispersion is found (assuming $n_i > 1$); however, since $\phi_i \geq 0$ has to hold, the negative correlation is restricted by $\rho \geq -1/(n_i - 1)$.

### Unobserved Heterogeneity

A possible source of heterogeneity is that unobserved or unobservable variables induce extra variability in $y$. Let us assume that there is a latent variable $D_i$ that selects a probability and the response $y_i$ given the selected value (and observed covariates) has the usually assumed binomial distribution (see also Williams, 1982). To be more specific, the following is assumed.

(1) A latent variable $D_i \in [0, 1]$ with

$$\text{E}(D_i) = \pi_i, \quad \text{var}(D_i) = \delta\pi_i(1-\pi_i), \delta \geq 0.$$

selects a value $\vartheta_i$ from $[0, 1]$.

(2) Given $D_i = \vartheta_i$, the response has the binomial distribution

$$y_i | D_i = \vartheta_i \sim B(n_i, \vartheta_i).$$

By using $\text{E}(\text{E}(Y|X)) = \text{E}(Y)$ and $\text{var}(Y) = \text{var}\,\text{E}(Y|X) + E(\text{var}(Y|X))$, one obtains for the marginal distribution of $y_i$

$$\text{E}(y_i) = n_i\pi_i, \quad \text{var}(y_i) = n_i\pi_i(1-\pi_i)\phi_i,$$

where $\phi_i = 1 + (n_i - 1)\delta$. Thus the variable $y_i$ has the usual mean to be expected under the binomial model, but the variances are inflated. If $\delta > 0$ and $n_i > 1$, the variability is larger than in the binomial model. The limiting case $\delta = 0$ means that the latent variable has zero variance. Then the latent variable has no effect and the binomial model holds.

It is noteworthy that the overdispersion resulting from correlated responses and the assumption of an underlying latent variable yields almost the same model for overdispersion. A difference is that the latent variable approach allows only overdispersion whereas the correlation model also allows for a restricted form of underdispersion.

There are several strategies for dealing with overdispersion data. One is the explicit modeling of heterogeneity and an example is the beta-binomial model, which is considered in the following. Alternatively, one can assume a normal distribution for the unobserved heterogeneity or a finite mixture; both modeling approaches will be treated in Chapter 14. A second strategy is to use generalized estimation functions, which are also considered in the next section.

### 5.3.2 Beta-Binomial Model

A candidate for the distribution of the latent variable $D_i \in [0, 1]$ is the beta distribution (see Appendix A). If one assumes $D_i \sim Beta(a_i, b_i)$ with parameters $a_i, b_i > 0$, the means and variances are given by

$$\mathrm{E}(D_i) = \pi_i = \frac{a_i}{a_i + b_i}, \quad \mathrm{var}(D_i) = \pi_i(1 - \pi_i)/(a_i + b_i + 1).$$

Assuming the parametric model $\pi_i = h(\boldsymbol{x}_i^T \beta)$, one obtains $\mathrm{var}(D_i) = \delta_i \pi_i (1 - \pi_i)$ with the dispersion parameter $\delta_i = 1/(a_i + b_i + 1)$.

The marginal distribution of $y_i$ is called the *beta-binomial distribution*; the explicit form is obtained by integrating with respect to the density of the beta distribution:

$$\begin{aligned} P(y_i; n_i, a_i, b_i) &= \int P(y_i | D_i = \vartheta_i) p(\vartheta_i) d\,\vartheta_i \\ &= \int \binom{n_i}{y_i} \vartheta_i^{y_i} (1 - \vartheta_i)^{n_i - y_i} \frac{\Gamma(a_i + b_i)}{\Gamma(a_i)\Gamma(b_i)} \vartheta_i^{a_i - 1} (1 - \vartheta_i)^{b_i - 1} d\,\vartheta_i \\ &= \binom{n_i}{y_i} \frac{(a_i + y_i - 1)_{y_i} (b_i + n_i - y_i - 1)_{n_i - y_i}}{(a_i + b_i + n_i - 1)_{n_i}}, \end{aligned}$$

where $(k)_r = k(k-1)\ldots(k-r+1)$. An alternative form is

$$P(y_i; n_i, a_i, b_i) = \frac{B(a_i + y_i, b_i + n_i - y_i)}{B(a_i, b_i)},$$

$y_i \in \{0, 1 \ldots, n_i\}$, where $B(p, q) = \Gamma(p)\Gamma(q)/\Gamma(p + q)$ is the beta function.

The model contains the parameters $\boldsymbol{\beta}$ implicitly. Instead of $a_i, b_i$ one may use the parameters $\pi_i = a_i/(a_i + b_i), \delta_i = 1/(a_i + b_i + 1)$, which yields $a_i = \pi_i(1 - \delta_i)/\delta_i, b_i = (1 - \pi_i)(1 - \delta_i)/\delta_i$ and turns the density into $B(\pi_i(1 - \delta_i)/\delta_i + y_i, (1 - \pi_i)(1 - \delta_i)/\delta_i + n_i - y_i)/B(\pi_i(1 - \delta_i)/\delta_i, (1 - \pi_i)(1 - \delta_i)/\delta_i)$. For simplicity it is often assumed that $\delta_i = \delta$ is the same for all observations. Then, assuming that the beta-binomial distribution holds, one specifies the mean by $\pi_i = h(\boldsymbol{x}_i^T \boldsymbol{\beta})$ with a fixed response function $h$. Thus the mean is given by $\mu_i = n_i \pi$ and the variance by $\mathrm{var}(y_i = n_i \pi(1 - \pi)[1 + (n_i - 1)\delta]$, with $\delta = 0$ representing the limiting case of a binomial distribution.

Beta-binomial models cannot be treated within the framework of generalized linear models. Crowder (1987) and Hinde and Démetrio (1998) gave algorithms for solving the maximum

likelihood equations. The latter obtained the fit by iterating between estimates of $\beta$ for fixed $\delta$ and estimates of $\delta$ for fixed $\boldsymbol{\beta}$. Prentice (1986) considered a more general model, where $\delta_i$ could also depend on covariates.

### 5.3.3 Generalized Estimation Functions and Quasi-Likelihood

Explicit parametric modeling of heterogeneity as used in the derivation of the beta-binomial model has several drawbacks. Numerical integration can be avoided only for specific distributions. More seriously, the assumption of a specific distribution function determines inferences, although it is hard to validate. To avoid the restrictive assumption of a specific latent variable one can use generalized estimation equations that are based on quasi-likelihood approaches (see also Section 3.11).

In *quasi-likelihood approaches* it is not necessary to specify a distribution for the responses. Much weaker, one only specifies the first two moments. For count data $y_i$ with $y_i \in \{0, 1 \ldots, n_i\}$ one might assume that the means and variances are given by

$$\mathrm{E}(y_i) = n_i\pi_i = n_i h(\boldsymbol{x}_i^T\boldsymbol{\beta}), \quad \mathrm{var}(y_i) = n_i\pi_i(1-\pi_i)\phi,$$

which corresponds to an overdispersed binomial distribution. For proportions $\bar{y}_i = p_i = y_i/n_i$ one has $\mathrm{E}(\bar{y}_i) = \pi_i = h(\boldsymbol{x}_i^T\beta)$, $\mathrm{var}(\bar{y}_i) = \pi_i(1-\pi_i)\phi/n_i$.

The essential point in these equations is that the mean and variance are specified, with the variance having the simple form of an inflated binomial variance:

$$\mathrm{var}(y_i) = v(\pi_i)\phi,$$

where $v(\pi_i)$ is a known (or fully specified) variance function. This means that the functional form of the variance (depending on covariates) is assumed to be known. Only the scaling factor $\phi$ is unknown and has to be estimated. An estimate of $\phi$ is

$$\hat{\phi} = \frac{1}{N-p}\sum_{i=1}^{N}\frac{(y_i-\hat{\mu}_i)^2}{n_i v(\pi_i)} = \frac{1}{N-p}\sum_{i=1}^{N}\frac{(\bar{y}_i-\hat{\pi}_i)^2}{v(\pi_i)/n_i}.$$

For $v(\pi_i) = \pi_i(1-\pi_i)$ one obtains $\hat{\phi} = \chi^2_P/(N-p)$, where $\chi^2_P$ is Pearson's goodness-of-fit statistic.

The variance function $v(\pi_i) = \pi_i(1-\pi_i)$ is very easy to handle. One fits the ordinary binary regression model and uses the ML estimate. To obtain the correct covariances matrix of $\hat{\beta}$ one multiplies the maximum likelihood covariance by $\hat{\phi}$ since the covariance is approximated by $\mathrm{cov}(\boldsymbol{\beta}) \approx \hat{\phi}\boldsymbol{F}(\boldsymbol{\beta})^{-1}$, where $\boldsymbol{F}(\hat{\boldsymbol{\beta}})$ is the Fisher matrix of the ordinary model. Maximum likelihood standard errors are multiplied by $\sqrt{\hat{\phi}}$ and $t$ statistics are divided by $\sqrt{\hat{\phi}}$.

An alternative estimate of $\phi$, based on the deviance, is $\tilde{\phi} = D/(N-p)$. It is comparable to $\hat{\phi}$ if all $n_i$'s are of similar size. While $\hat{\phi}$ is also consistent for small local samples, this does not hold for $\tilde{\phi}$ (compare McCullagh and Nelder, 1989).

The estimation equation that is used to obtain an estimate of $\boldsymbol{\beta}$ has the form

$$\sum_{i=1}^{N}\boldsymbol{x}_i\frac{h'(\boldsymbol{x}_i^T\boldsymbol{\beta})}{\mathrm{var}(p_i)}(p_i - h(\boldsymbol{x}_i^T\boldsymbol{\beta})) = \mathbf{0}. \tag{5.5}$$

When the variance of $p_i$ is the inflated binomial variance, specified by $\mathrm{var}(p_i) = \phi h(\boldsymbol{x}_i'\boldsymbol{\beta})(1 - h(\boldsymbol{x}_i'\boldsymbol{\beta}))/n_i$, the solution of equation (5.5) does not depend on $\phi$. Of course, when $\phi = 1$ and

$v(\pi_i) = \pi_i(1-\pi_i)$ is assumed, equation (5.5) is equivalent to the ML estimation equation for the corresponding binomial model.

When using these estimates it is assumed that the mean and variance are correctly specified. More generally, one may consider equation (5.5) as a *generalized estimation function* under the assumption that only the mean is correctly specified, whereas the variance is considered a working covariance that does not have to be the variance of the data-generating model (e.g., Gourieroux et al., 1984). It may be shown that under regularity conditions one obtains an asymptotically normally distributed estimate $\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \hat{\boldsymbol{F}}^{-1}\hat{\boldsymbol{V}}\hat{\boldsymbol{F}}^{-1})$, where the sandwich matrix $\hat{\boldsymbol{F}}^{-1}\hat{\boldsymbol{V}}\hat{\boldsymbol{F}}^{-1}$ is determined by

$$\hat{\boldsymbol{F}} = \sum_{i=1}^{N} \boldsymbol{x}_i \boldsymbol{x}_i^T \frac{h'(\boldsymbol{x}_i^T\hat{\boldsymbol{\beta}})^2}{\hat{\text{var}}(p_i)} \quad \hat{\boldsymbol{V}} = \sum_{i=1}^{g} \boldsymbol{x}_i \boldsymbol{x}_i^T \frac{h'(\boldsymbol{x}_i^T\hat{\boldsymbol{\beta}})^2}{\hat{\text{var}}(p_i)^2}(p_i - h(\boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}))^2$$

with $\hat{\text{var}}(p_i) = \hat{\phi} v(h(\boldsymbol{x}_i'\hat{\boldsymbol{\beta}}))$.

Although the inflated binomial variance is easy to handle, the assumption of correlated responses as well as the modeling by latent variables suggest that for count data $y_i \in \{0, 1 \ldots, n_i\}$ variances are given by $var(y_i) = n_i\pi_i(1-\pi_i)[1 + (n_i - 1)\delta]$, where $\delta$ is a dispersion parameter. The corresponding quasi-likelihood approach, which assumes that the mean and variance are correctly specified, is less easy to handle because the dispersion parameter does not cancel out. Williams (1982) proposed an algorithm that iterates between estimates of $\beta$ for fixed $\delta$ and estimates of $\delta$ for fixed $\beta$.

Liang and McCullagh (1993) compared various approaches to the modeling of overdispersion; an overview can be found in Poortema (1999). Lambert and Roeder (1995) introduced a convexity plot that detects overdispersion, relative variance curves, and tests that help to understand the nature of the overdispersion. Relative variance curves and tests sometimes distinguish the source of the overdispersion better than score tests.

### Example 5.1: Teratology

In a teratology experiment considered by Moore and Tsiatis (1991) and Liang and McCullagh (1993), 58 rats on iron-deficient diets were assigned to four groups (see Table 5.3). In the first group only placebo injections were given, and in the other groups iron supplements were given. The animals were impregnated and sacrificed after three weeks. The response was whether the fetus was dead $(y_{ij} = 1)$ for each fetus in each rat's litter.

TABLE 5.3: Response counts of (litter size, number dead) for 58 litters of rats in low-iron teratology study.

| |
|---|
| Group 1: Untreated (low iron) |
| (10,1)(11,4)(12,9)(4,4)(10,10)(11,9)(9,9)(11,11)(10,10)(10,7)(12,12) |
| (10,9)(8,8)(11,9)(6,4)(9,7)(14,14)(12,7)(11,9)(13,8)(14,5)(10,10) |
| (12,10)(13,8)(10,10)(14,3)(13,13)(4,3)(8,8)(13,5)(12,12) |
| Group 2: Injections days 7 and 10 |
| (10,1)(3,1)(13,1)(12,0)(14,4)(9,2)(13,2)(16,1)(11,0)(4,0)(1,0)(12,0) |
| Group 3: Injections days 0 and 7 |
| (8,0)(11,1)(14,0)(14,1)(11,0) |
| Group 4: Injections weekly |
| (3,0)(13,0)(9,2)(17,2)(15,0)(2,0)(14,1)(8,0)(6,0)(17,0) |

Source: Moore and Tsiatis (1991)

Since the observations for the $i$th litter $y_{i1}, \ldots y_{in_i}$ were measured on one female rat, one might suspect overdispersion. Let $y_{ij} \sim B(1, \pi_i)$ denote the response of the $j$th fetus in litter $i$ and

$y_i = y_{i1} + \cdots + y_{in_i} \sim B(n_i, \pi_i)$ denote the number dead out of the $n_i$ fetuses in litter $i$. As a structural component one assumes for $\pi_{ij} = \mathrm{E}(y_{ij})$ a logit model

$$\mathrm{logit}(\pi_{ij}) = \beta_0 + x_{G(2)}\beta_2 + x_{G(3)}\beta_3 + x_{G(4)}\beta_4,$$

where $x_{G(i)}$ is a (0-1)-dummy variable with $x_{G(i)} = 1$ if the observation is from group $i$ and 0 otherwise. The naive approach assumes that all observations $y_{11}, y_{12}, \ldots$ are independent binary variables. One obtains a deviance of 173.45 on 54 degrees of freedom, which, however, should not be interpreted as a goodness-of-fit statistic. In Table 5.4 results are given for the approaches that yield the same estimates but differing standard errors. The independence model assumes that all binary observations are independent. The quasi-likelihood approach with an inflated binomial variance uses $var(y_i) = \phi n_i \pi_i (1 - \pi_i)$. In the weaker generalized estimation approach, independence was used as a working covariance. It is seen that standard errors are larger when overdispersion is taken into account. The simple independence model is certainly not appropriate. In Table 5.5 estimates are given for the beta-binomial model and for the mixed model approach, estimated by penalized quasi-likelihood and Gauss-Hermite quadrature with 14 quadrature points. The mixed model assumes that each unit (rat) has its own intercept, which follows a normal distribution. Therefore, heterogeneity across units is modeled quite flexibly (for details see Chapter 14). It is seen that the estimates for the beta-binomial model are slightly smaller than the estimated values in Table 5.4. For the mixed model approach estimates are distinctly larger, an effect that is frequently found in subject-specific models (Chapter 14). In addition, a discrete mixture model with two components was fitted. The model assumed that the response was a mixture of two components that have distinct intercepts (see Chapter 14). The two intercepts of the components, which are not given in the table, were $-0.211$ and $2.458$. The estimated effects are quite close to the estimates for the mixed model with normally distributed random effects. □

TABLE 5.4: Estimates and standard errors for independence model, quasi-likelihood model, and generalized estimation functions fitted to teratology data with logit link.

| | Estimates | Standard Errors | | |
|---|---|---|---|---|
| | | Independence Model | Quasi-Likelihood $var(y_i) = \phi n_i \pi_i (1 - \pi_i)$ | GEE Independence |
| $\beta_0$ | 1.144 | 0.129 | 0.219 | 0.275 |
| $\beta_2$ | −3.323 | 0.331 | 0.560 | 0.440 |
| $\beta_3$ | −4.476 | 0.731 | 1.237 | 0.610 |
| $\beta_4$ | −4.130 | 0.476 | 0.806 | 0.576 |
| $\hat{\phi}$ | | − | 2.865 | 1.007 |

TABLE 5.5: Estimates for beta-binomial model and several mixture logit models fitted to teratology data

| | | Mixed Models | | |
|---|---|---|---|---|
| | Beta-Binomial | Pen Quasi-Likelihood | Gauss-Hermite | Discrete Mixture |
| $\beta_0$ | 1.345 (0.244) | 1.687 (0.306) | 1.802 (0.362) | - |
| $\beta_2$ | −3.087 (0.521) | −4.130 (0.614) | −4.515 (0.736) | −4.309(0.481) |
| $\beta_3$ | −3.865 (0.863) | −5.274 (0.981) | −5.855 (1.189) | −5.509(0.824) |
| $\beta_4$ | −3.919 (0.684) | −5.109 (0.747) | −5.594 (0.919) | −5.082(0.595) |
| | | $\hat{\sigma} = 1.456$ | $\hat{\sigma} = 1.533$ | |

## 5.4 Conditional Likelihood

In some applications the model involves parameters that are of minor interest to the investigator. These parameters are often called nuisance or incidental parameters. For example, in a clinical trial that compares the effect of a treatment (new drug or new therapy) with a current standard, let the response be 'success' or 'failure'. Suppose that data on the effect of treatment are available from $g$ sources or strata. In a multi-center clinical trial the strata are the medical centers in which the trials are taken. One obtains for each stratum a $2 \times 2$ table, the table for the $i$th stratum has the form

| | Success | Failure | | Proportions |
|---|---|---|---|---|
| Treatment | $y_{i1}$ | $n_{i1} - y_{i1}$ | $n_{i1}$ | $p_{i1} = y_{i1}/n_{i1}$ |
| Control | $y_{i2}$ | $n_{i2} - y_{i2}$ | $n_{i2}$ | $p_{i2} = y_{i2}/n_{i2}$ |
| | $y_{i\cdot}$ | $n_i - y_{i\cdot}$ | $n_i$ | |

The main effect logit model, which models the effect of the stratum $i$ and the treatment on the binary response ($y = 1$ for success and $y = 0$ for failure), has the form

$$\log(\frac{\pi(i, x_T)}{1 - \pi(i, x_T)}) = \beta_i + x_T\beta, \tag{5.6}$$

where $\pi(i, x_T)$ is the probability of success given the stratum $i$ and the treatment group specified by $x_T$ ($x_T = 1$ for treatment and $x_T = 0$ for control). The treatment effect is given by $\beta$ while $\beta_i$ represents the stratum effect.

The problem with the linear logistic model is that it contains $g$ parameters, which have to be estimated on the basis of $2g$ observed binomial proportions. Thus for large $g$ maximum likelihood estimates will hardly be efficient. Moreover, the parameter of interest is $\beta$, and the parameters $\beta_1, \ldots, \beta_{g-1}$ are nuisance parameters. When testing the hypothesis "no treatment effect" ($H_0 : \beta = 0$) one alternative is to condition on the success totals $y_{i1} + y_{i2} = y_{i\cdot}$, obtaining a hypergeometric distribution that does not depend on the nuisance parameters.

### The Hypergeometric Distribution

Suppose that independently two random samples of size $n_1, n_2$ are drawn and it is observed whether attribute A occurs or not. One obtains a $2 \times 2$ table with fixed marginals $n_1, n_2$. The following table gives numbers of subjects who possess the attribute in question.

| | $A$ | $\bar{A}$ | |
|---|---|---|---|
| Population 1 | $y = y_{11}$ | $y_{12}$ | $n_1$ |
| Population 2 | $y_{21}$ | $y_{22}$ | $n_2$ |
| | $n_A$ | $n_{\bar{A}}$ | $n$ |

From the marginals, $n_1, n_2$ and therefore $n = n_1 + n_2$ are fixed, while $n_A$ and $n_{\bar{A}}$ are random. If also the marginals $n_A, n_{\bar{A}}$ are fixed, the observations in the table no longer follow a binomial distribution. The selection is the same as when drawing $n_1$ marbles from a box which contains $n$ marbles, $n_A$ of them possessing attribute $A$. Thus the distribution of $y = y_{11}$ is given by the *hypergeometric distribution*

$$P(y|\boldsymbol{n}, \boldsymbol{n}_A) = \frac{\begin{pmatrix} n_A \\ y \end{pmatrix}\begin{pmatrix} n_{\bar{A}} \\ n_1 - y \end{pmatrix}}{\begin{pmatrix} n \\ n_1 \end{pmatrix}} = \frac{\begin{pmatrix} n_1 \\ y \end{pmatrix}\begin{pmatrix} n_2 \\ n_A - y \end{pmatrix}}{\begin{pmatrix} n \\ n_A \end{pmatrix}}, \tag{5.7}$$

where the range of possible values for $y$ is given by the integers, satisfying $l = \max\{0, n_1 - (n - n_A)\} \leq y \leq \min\{n_1, n_A\} = u$. The notation $P(y|\boldsymbol{n}, \boldsymbol{n}_A)$ shows that the distribution of $y$ is conditionally on the marginal counts $\boldsymbol{n}_A^T = (n_A, n_{\bar{A}})$ and $\boldsymbol{n}^T = (n_1, n_2)$. The second form in (5.7) results from the symmetry of the problem. Distribution (5.7) is abbreviated by $H(\boldsymbol{n}, \boldsymbol{n}_A)$. The hypergeometric distribution may also be derived directly from the independent binomial distributions $y_{11} \sim B(n_1, \pi)$, $y_{21} \sim B(n_2, \pi)$. Then the conditional distribution of $y = y_{11}$ conditionally on $y_{11} + y_{21} = n_A$ is given by (5.7).

### The Non-Central Hypergeometric Distribution

In the more general case the response probability will be different for the two populations. By assuming independent binomial distributions

$$y_{11} \sim B(n_1, \pi_1),\ y_{21} \sim B(n_2, \pi_2),$$

one obtains for the conditional distribution of $y = y_{11}$, conditionally on $y_{11} + y_{21} = n_A$, the non-central hypergeometric distribution $H(\boldsymbol{n}, \boldsymbol{n}_A, \gamma)$, given by

$$P(y|\boldsymbol{n}, \boldsymbol{n}_A, \gamma) = \frac{\left(\begin{array}{c} n_1 \\ y \end{array}\right)\left(\begin{array}{c} n_2 \\ n_A - y \end{array}\right)\gamma^y}{P_o(\gamma)},$$

where $P_o(\gamma)$ is the polynomial in $\gamma$,

$$P_o(\gamma) = \sum_{j=l}^{u} \left(\begin{array}{c} n_1 \\ j \end{array}\right)\left(\begin{array}{c} n_2 \\ n_A - j \end{array}\right)\gamma^j,$$

and $\gamma = \{\pi_1/(1-\pi_1)\}/\{\pi_2/(1-\pi_2)\}$ is the odds ratio. The non-central hypergeometric distribution follows a simple exponential family with the (conditional) log likelihood given by

$$y\log(\gamma) - \log(P_o(\gamma)) + \log\left\{\left(\begin{array}{c} n_1 \\ y \end{array}\right)\left(\begin{array}{c} n_2 \\ n_A - y \end{array}\right)\right\}$$

This has the form $y\theta - b(\theta) + c(y)$ with $\theta = \log(\gamma)$ and $b(\theta) = \log(P_o(e^\theta))$, yielding

$$\begin{aligned} E(y) &= b'(\theta) = P_1(e^\theta)/P_o(e^\theta), \\ var(y) &= b''(\theta) = P_2(e^\theta)/P_o(e^\theta) - \{P_1(e^\theta)/P_o(e^\theta)\}^2, \end{aligned}$$

where $P_1(e^\theta) = \partial P_o(\theta)/\partial\theta, P_2(e^\theta) = \partial^2 P_o(\theta)/\partial\theta^2$ given by

$$P_r(\gamma) = \sum_{j=l}^{n} \left(\begin{array}{c} n_1 \\ j \end{array}\right)\left(\begin{array}{c} n_2 \\ n_A - j \end{array}\right) j^r\gamma^j$$

(see McCullagh and Nelder, 1989).

When testing the nullhypothesis $H_0 : \beta = 0$ in model (5.6) conditioning on the success totals $y_{i1} + y_{i2} = y_{i\cdot}$ for each stratum yields the hypergeometric distribution

$$y_{i1}|\boldsymbol{y}_{i\cdot} \sim H(\boldsymbol{n}_{i\cdot}, \boldsymbol{y}_{i\cdot}, e^\beta),$$

where $\boldsymbol{n}_{i\cdot}^T = (n_{i1}, n_{i2}), \boldsymbol{y}_{i\cdot}^T = (y_{i1} + y_{i2}, n_i - y_{i1} - y_{i2})$ are the marginals under $H_0$. Then the conditional likelihood involves only one parameter, $e^\beta$. One obtains for the mean and variance of the hypergeometric distribution

$$\mathrm{E}(y_{i1}) = n_{i1}\frac{y_{i\cdot}}{n_i}, \quad \mathrm{var}(y_{i1}) = n_{i1}\frac{y_{i\cdot}}{n_i}\left(1 - \frac{y_{i\cdot}}{n_i}\right)\frac{n_i - n_{i1}}{n_i - 1}. \tag{5.8}$$

A statistic for the nullhypothesis may be based on the difference between observed counts and the expected values of the nullhypothesis. A test statistic that has been proposed by Mantel and Haenszel (1959), and in a similar form by Cochran (1954), is

$$T = \frac{\{\sum_{i=1}^{g}(y_{i1} - \mathrm{E}(y_{i1}))\}^2}{\sum_{i=1}^{g}\mathrm{var}(y_{i1})} \tag{5.9}$$

In the Mantel-Haenszel statistic a continuity correction is included, yielding

$$\chi^2_{MH} = \frac{\{|\sum_{i=1}^{g}(y_{i1} - \mathrm{E}(y_{i1}))| - 0.5\}^2}{\sum_{i=1}^{g}\mathrm{var}(y_{i1})}$$

with $\mathrm{E}(y_{i1})$ and $\mathrm{var}(y_{i1})$ derived from the hypergeometric distribution under $H_0(5.8)$. By using the observed proportions $p_{i1} = y_{i1}/n_{i1}, p_{i2} = y_{i2}/n_{i2}$ it may be rewritten as

$$\chi^2_{MH} = \frac{\left\{|\sum_{i=1}^{g}\frac{n_{i1}n_{i2}}{n_i}(p_{i1} - p_{i2})| - 0.5\right\}^2}{\sum\frac{n_{i1}n_{i2}}{n_i - 1}\bar{p}_i(1 - \bar{p}_i)}$$

where $\bar{p}_i = (n_{i1}p_{i1} + n_{i2}p_{i2})/n_i$. Cochran (1954) proposed the statistic (5.9) with the variances replaced by variances derived from the two binomials for treatment and control. Under the nullhypothesis the tests have a large-sample $\chi^2$-distribution. For more details on these tests see Agresti (2002). A general treatment of conditional likelihood approaches is given in McCullagh and Nelder (1989). An alternative approach to reduce the number of parameters is to assume that the stratum-specific parameters are random effects (see Chapter 14).

## 5.5 Further Reading

*Single-index model*. Several approaches to the estimation of single-index models have been proposed. One popular technique is based on an average derivative estimation, which exploits the fact that the average gradient of $h(\mathbf{x}_i'\boldsymbol{\beta})$ is proportional to $\boldsymbol{\beta}$ (see Powell et al., 1989; Hristache et al., 2001). An $M$-estimation that considers the unknown link function as an infinite-dimensional nuisance parameter was considered by Klein and Spady (1993) and Härdle et al. (1993). Weisberg and Welsh (1994) proposed an algorithm that alternates between the estimation of $\boldsymbol{\beta}$ and $h(.)$. Yu and Ruppert (2002) suggested using penalized regression splines. They reported more stable estimates compared to earlier approaches based on local regression (e.g., Carroll et al., 1997). Xia et al. (2002) proposed a symbiosis of sliced inverse regression average derivative estimation and local linear smoothing. Naik and Tsai (2001) proposed a model selection criterion for single-index models that selects variables and also smoothing parameters for the unknown link function.

*R packages*. Binary response models including quasi-likelihood models can be fitted with the function *glm*. Various link function can be specified in the family function. Generalized estimation functions are available in the library *gee*, function *gee*. The beta-binomial model can be fitted by the function *vglm*in the library *VGAM*. For the fitting procedures of mixed and finite mixture models see Chapter 14.

## 5.6 Exercises

5.1 The dataset *dust* is available from package *catdata* (or at http://www.stat.uni-muenchen.de/sfb386/ under the name "Chronic bronchitis and dust concentration").

(a) Fit models with different link functions for non-smokers and smokers separately.

(b) Compare the fitted models and discuss model selection.

5.2 The dataset *birth* is available from package *catdata*. Consider the binary response "Did a perineal tear occur" and explanatory variables weight, height, head circumference of child, and month of birth.

(a) Fit models with different link functions.

(b) Select an appropriate model.

5.3 The package *flexmix* contains the data set *betablocker*, which is from a 22-center clinical trial of beta-blockers for reducing mortality after myocardial infarction (see also Aitkin, 1999). In addition to centers, there is only one explanatory variable, treatment, coded as 0 for control and 1 for beta-blocker treatment.

(a) For the 44 binomial observations, fit the simple logit model, an appropriate quasi-likelihood model, and by using generalized estimation equations. Compare the effects and standard errors.

(b) Test if the treatment by beta-blockers has an effect by using test statistics based on conditional likelihood.

# Chapter 6

# Regularization and Variable Selection for Parametric Models

In several chapters we discussed parametric regression modeling for a moderate number of explanatory variables based on maximum likelihood methods. In some areas of application, however, the number of explanatory variables may be very high. For example, in genetics, where binary regression is a frequently used tool, the number of predictors may be even larger than the number of predictors. In this "$p > n$ problem" maximum likelihood and similar estimators are bound to fail. Typical data of this type are microarray data, where the expressions of thousands of predictors (genes) are observed and only some hundred samples are available. For example, the dataset considered by Golub et al. (1999a), which constitutes a milestone in the classification of cancer, consists of gene expression intensities for 7129 genes of 38 leukemia patients, from which 27 were diagnosed with acute lymphoblastic leukemia and the remaining patients acute myeloid leukemia.

In high-dimensional problems the reduction of the predictor space is the most important issue. A reduction technique with a long history is *stepwise variable selection*. However, stepwise variable selection as a discrete process is extremely variable. The results of a variable selection procedure may be determined by small changes in the data. The effect is often poor performance (see, e.g., Frank and Friedman, 1993). Moreover, it is challenging to investigate the sampling properties of stepwise variable selection procedures.

An alternative to stepwise subset selection is *regularization methods*. Ridge regression is a familiar regularization method that adds a simple penalty term to the log-likelihood and thereby shrinks estimates toward zero. In recent years several alternative regularization techniques based on penalties have been proposed, including methods that perform "smooth" variable selection. These methods select variables simultaneously via optimizing a penalized likelihood, and hence allow one to estimate standard errors. In the following we consider several penalty methods as well as boosting techniques, which are ensemble methods but also serve as regularization methods in structured regression.

Important aspects for regression modeling by regularization techniques are

- *existence of unique estimates* – this is where maximum likelihood estimates often fail;
- *prediction accuracy* – a model should be able to yield a decent prediction of the outcome;
- *sparseness and interpretation* – thea parsimonious model that contains the strongest effects is easier to interpret than a big model with hardly any structure.

We start with the conventional stepwise selection procedures and then consider regularized estimates. Most of the methods focus on variable selection, but regularization can also be helpful when one wants to know which categories of a categorical predictor should be distinguished.

## 6.1 Classical Subset Selection

In subset selection the predictor space is reduced by retaining only a subset of the variables. The main strategies are best subset selection, forward selection, backward selection, and a combination of the latter two methods.

*Best subset selection* aims at finding the best subset of predictors among all subsets of the variables $x_1, \ldots, x_p$. "Best" may be defined by minimizing some criterion like $AIC$ or $BIC$. For binary or Poisson-distribution models, where estimates have to be computed iteratively, the full subset selection is extremely demanding if the number of variables is large.

*Forward selection* seeks a path through all possible subsets by sequentially adding one predictor into the model. The decision for a predictor may be based on test statistics. Let $M$ denote the current model and $M_r$ the model $M$ with the additional variable $x_r$. Then, a test on the significance of variable $x_r$ within model $M_r$ is given by the difference of the deviances:

$$D(M|M_r) = D(M) - D(M_r).$$

One selects that variable $x_{r_0}$ for which the corresponding $p$-value $p_r$ is minimized, $r_0 = \arg\min_r p_r$, provided that the $p$-value is below some prechosen inclusion level $\alpha_{\text{in}}$. The procedure aims to select variables rather than single terms or parameters. This means that, in a mixture of variables, some of them metric and some of them categorical, the differences of deviances $D(M|M_r)$ compare models of different sizes. While a metric covariate typically contributes only one term (or parameter) to the model, a categorical predictor, unless it is binary, will contribute more than one term (parameter) to the predictor. In cases where each variable corresponds to one term in the linear predictor, for example, in a main effect model with only metric predictors, minimization of the $p$-values is equivalent to maximizing $D(M|M_r)$. Then one implicitly chooses the variable that most improves the fit, since the deviance $D(M_{r_0})$ as a measure for the discrepancy between data and model is minimized. The procedure stops when no additional variable contributes significantly to the model $M$. One should be aware that the significance level $\alpha_{\text{in}}$ is a threshold rather than a significance level. Since many tests are performed, one has a multiple test problem and control of the multiple significance level is difficult.

Alternative test statistics that might be computationally easier to handle are the score test and the Wald test. Fahrmeir and Frost (1992) suggested the score test and computed the test statistic by efficient sweeps of the inverse information matrix. Since deviances use maximum likelihood for both models $M$ and $M_r$, one might also run into problems with the existence of ML estimates. In contrast, the score test uses only the estimates of the restricted model; computation of the parameter estimates for the larger model is not required.

*Backward selection* starts with the full model and sequentially deletes predictors. The choice of the variable to delete is again typically based on test statistics. Let $M$ denote the current model and $M_{\setminus r}$ the model $M$ without predictor $r$. Then the deviance

$$D(M_{\setminus r}|M) = D(M_{\setminus r}) - D(M)$$

tests the significance of variable $x_r$ within model $M$. One selects the variable $r_0$ for which the corresponding $p$-value $p_r$ is maximal provided it is above some pre-chosen exclusive level $\alpha_{\text{out}}$. The procedure stops when each predictor in the model has a $p$-value below the level $\alpha_{\text{out}}$.

Computationally more efficient procedures may be obtained by using the Wald test (Fahrmeir and Frost, 1992). Backward selection strategies are restricted to cases where an estimate for the full model exists. This is often a problem when the number of predictors is large.

Forward and backward strategies may also be combined. After a new variable has been taken into the model (forward step) one investigates if one of the other variables in the model may now be deleted by performing a backward step. Both steps have to be controlled by inclusion and exclusion thresholds $\alpha_{\text{in}}$ and $\alpha_{\text{out}}$, which, however, provide only local control of the model search.

Best subset selection and forward/backward strategies have several disadvantages. As already noted, subset selection is a discrete process, either a variable is in or out of the model, and therefore extremely variable. The instability of stepwise regression models was demonstrated for example by Breiman (1996b). Moreover, one should be very cautious with the interpretation of the found effects. Standard errors computed for the final model are not trustworthy because they simply ignore the model search. Taking the model search into account would yield much larger standard errors.

Subset selection has been studied extensively for normal distribution models; see, for example, Seeber (1977), Miller (1989), and Furnival and Wilson (1974). The latter gave an efficient algorithm that performs best subset selection up to $30$ or $40$ predictors. Lawless and Singhal (1978, 1987) developed efficient screening and all-subsets procedures for generalized linear models by use of likelihood ratio statistics. These methods were discussed within a more general framework by Fahrmeir and Frost (1992).

## 6.2 Regularization by Penalization

Regularization methods that are derived from maximum likelihood estimates are based on the *penalized log-likelihood*:

$$l_p(\boldsymbol{\beta}) = \sum_{i=1}^{n} l_i(\boldsymbol{\beta}) - \frac{\lambda}{2} J(\boldsymbol{\beta}),$$

where $l_i(\boldsymbol{\beta})$ is the usual log-likelihood contribution of the $i$th observation, $\lambda$ is a tuning parameter, and $J(\boldsymbol{\beta})$ is a functional that penalizes the size of the parameters. By maximizing the penalized log-likelihood $l_p(\boldsymbol{\beta})$ one seeks estimates that are close to usual ML estimate but with regularized parameters. For example, the ridge penalty, which is one of the oldest penalization methods, uses the penalty $J(\boldsymbol{\beta}) = \sum_{j=1}^{p} \beta_j^2$. It penalizes the length of the parameter $\boldsymbol{\beta}$ and yields estimates that are shrunk toward zero.

There is a good reason for penalizing the length of the parameter. Segerstedt (1992) showed that under regularity assumptions the mean of the squared length of the ML estimate, $\mathrm{E}(\|\hat{\boldsymbol{\beta}}\|^2)$, is asymptotically $\|\boldsymbol{\beta}\|^2 + \mathrm{tr}(\boldsymbol{F}^{-1}(\boldsymbol{\beta}))$, where $\boldsymbol{F}^{-1}(\boldsymbol{\beta})$ denotes the Fisher matrix at the true value $\boldsymbol{\beta}$, which is an approximation to the covariance of $\hat{\boldsymbol{\beta}}$. Therefore, most common regularization techniques impose a penalty on the size of the regression coefficients, yielding shrunk estimates, which in particular have reduced variance. Shrinkage methods are in particular useful for obtaining estimates in applications where the use of the ML estimator involves problems. In a simple Gaussian linear regression the ML estimate is obtained by solving the estimation equation $(\boldsymbol{X}^T\boldsymbol{X})\boldsymbol{\beta} = \boldsymbol{X}^T\boldsymbol{y}$, which is easily solved if $\boldsymbol{X}^T\boldsymbol{X}$ is of full rank and the inverse $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ exists. In the case of collinearity, $\boldsymbol{X}^T\boldsymbol{X}$ is not of full rank, the ML estimate is not unique, and one has to determine which parts of $\boldsymbol{\beta}$ still may be estimated. Moreover, it has been shown that collinearity leads to poor performance of the estimators. Figure 6.1 illustrates the instability of ML estimates for strongly correlated data. It shows the estimates for two datasets that were drawn from the same underlying linear structure (same non-zero coefficients, $\beta_j = 5$,

for the first six variables, zero coefficients for the other variables). It is seen that estimates take quite extreme values, since for strongly correlated predictors high positive values of estimated coefficients balance negative values. The effect is a high variability of estimates. For each drawing one obtains quite different estimates that are far from the true values. Shrinkage estimators like ridge regression estimators, however, yield much smaller values (open circles in Figure 6.1), which are distinctly closer to the true values. Shrinkage methods become important especially when many predictors are available and therefore the corresponding design matrix $\boldsymbol{X}$ is very large and usually contains redundant columns. Similar effects occur in a binary regression, where an additional problem occurs, since ML estimates do not exist if the data may be separated (see Section 4.1). When many predictors are available, the tendency that data structures occur in which the responses $y = 0$ and $y = 1$ are separated increases strongly.

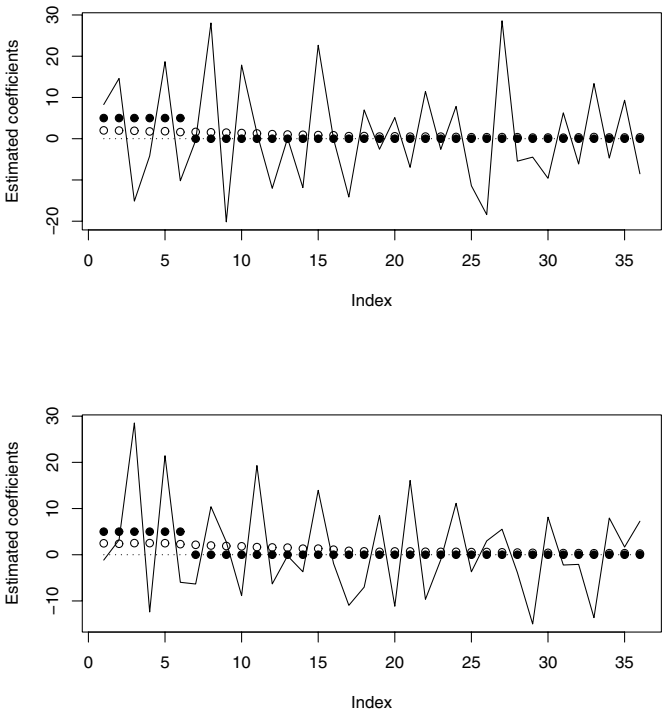


FIGURE 6.1: Maximum likelihood estimates for two datasets with correlated data; the first six variables are influential, and all other variables have zero coefficient.

Some shrinkage methods may also be seen as a continuous alternative to the selection of predictors. Since variables are retained or discarded, variable selection is a strictly discrete process that often exhibits high variance in prediction error. Shrinkage methods reduce the influence of variables in a much smoother way and therefore show less variability. Although shrinkage methods may be seen as providing alternative (more stable) methods to estimate the "true" regression coefficients, a more pragmatic view is that shrinkage methods yield regression models, true or not, that in particular in high-dimensional problems show better prediction error than usual maximum likelihood estimates. Therefore, the selection of the tuning parameter is often based on an estimate of the prediction error.

In the following let the predictor be given by $\eta_i = \boldsymbol{x}_i^T\boldsymbol{\beta}$, where $\boldsymbol{\beta}^T = (\beta_0, \beta_1, \ldots, \beta_p)$ and $\boldsymbol{x}_i$ contains a constant term. Frequently used penalties are of the *bridge penalty* type (Frank and Friedman, 1993):

$$J(\boldsymbol{\beta}) = \sum_{j=1}^{p} |\beta_j|^\gamma, \quad \gamma > 0. \tag{6.1}$$

For $\gamma = 2$ one obtains a *ridge regression* (Hoerl and Kennard, 1970) for $\gamma = 1$, the so-called *lasso* (Tibshirani, 1996 ). Alternatively, estimates are found by maximizing the log-likelihood

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{n} l_i(\boldsymbol{\beta})$$

subject to the constraint

$$\sum_{j=1}^{p} |\beta_j|^\gamma \leq t. \tag{6.2}$$

For $\gamma \geq 1$, this approach is equivalent to maximizing the penalized likelihood with $\lambda \geq 0$ since the constraint area is convex (Fu, 1998). Maximization of $l_p$ is usually referred to as a *penalized regression* whereas maximization of $l$ subject to (6.2) is called a *constrained regression*. In the following we will consider penalties of the bridge penalty type and others.

### 6.2.1 Ridge Regression

Ridge regression as introduced by Hoerl and Kennard (1970) for linear models and extended to GLM type models by Nyquist (1991) is based on the penalty $J(\boldsymbol{\beta}) = \sum_{j=1}^{p} \beta_j^2$ yielding the penalized log-likelihood

$$l_p(\boldsymbol{\beta}) = \sum_{i=1}^{p} l_i(\boldsymbol{\beta}) - \frac{\lambda}{2} \sum_{j=1}^{p} \beta_j^2.$$

For deriving estimates it is useful to rewrite the penalty in the form

$$J(\boldsymbol{\beta}) = \sum_{j=1}^{p} \beta_j^2 = \boldsymbol{\beta}^T \boldsymbol{P} \boldsymbol{\beta},$$

where $\boldsymbol{P} = (p_{ij})$ differs from the $(p+1) \times (p+1)$ identity matrix only by having $p_{11} = 0$ instead of $p_{11} = 1$. The corresponding penalized score function $\boldsymbol{s}_p(\boldsymbol{\beta})$ is given by

$$\boldsymbol{s}_p(\boldsymbol{\beta}) = \sum_{i=1}^{n} \boldsymbol{x}_i \frac{\partial h(\eta_i)}{\partial \eta} (y_i - \mu_i)/\sigma_i^2 - \lambda \boldsymbol{P} \boldsymbol{\beta},$$

yielding the estimation equation

$$\boldsymbol{X}^T \boldsymbol{D}(\boldsymbol{\beta}) \boldsymbol{\Sigma}^{-1}(\boldsymbol{\beta})(\boldsymbol{y} - \boldsymbol{\mu}) - \lambda \boldsymbol{P} \boldsymbol{\beta} = \boldsymbol{0},$$

where $\boldsymbol{y}^T = (y_1, \ldots, y_n)$, $\boldsymbol{\mu}^T = (\mu_1, \ldots, \mu_n)$, $\boldsymbol{X}^T = (\boldsymbol{x}_1 \ldots \boldsymbol{x}_n)$, $\boldsymbol{D}(\boldsymbol{\beta}) = \mathrm{diag}(\partial h(\eta_1)/\partial \eta, \ldots \partial h(\eta_r)/\partial \eta)$, $\boldsymbol{\Sigma}(\boldsymbol{\beta}) = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_n^2)$, $\sigma_i^2 = \mathrm{var}(y_i)$. For the normal distribution model one obtains with $\sigma_i^2 = \sigma^2$, $\boldsymbol{D} = \boldsymbol{I}$, $\tilde{\lambda} = \lambda\sigma^2$ the explicit solution

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T \boldsymbol{X} + \tilde{\lambda} \boldsymbol{P})^{-1} \boldsymbol{X}^T \boldsymbol{y},$$

which is the maximum likelihood estimate except for the term $\tilde{\lambda}\boldsymbol{P}$. For the covariance one obtains $\mathrm{cov}(\hat{\boldsymbol{\beta}}) = \sigma^2(\boldsymbol{X}^T\boldsymbol{X} + \lambda\sigma^2\boldsymbol{P})^{-1}\boldsymbol{X}^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X} + \lambda\sigma^2\boldsymbol{P})^{-1}$.

For generalized linear models iterative procedures, for example, Fisher scoring, have to be used. Fisher scoring for solving $\boldsymbol{s}_p(\hat{\boldsymbol{\beta}}) = \mathbf{0}$ has the form

$$\hat{\boldsymbol{\beta}}^{(k+1)} = \hat{\boldsymbol{\beta}}^{(k)} + \boldsymbol{F}_p(\hat{\boldsymbol{\beta}}^{(k)})^{-1}\boldsymbol{s}_p(\hat{\boldsymbol{\beta}}^{(k)}),$$

where $\boldsymbol{F}_p(\boldsymbol{\beta}) = \mathrm{E}(-\partial l_p/\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T) = \boldsymbol{F}(\boldsymbol{\beta}) + \lambda\boldsymbol{P}$ with $\boldsymbol{F}(\boldsymbol{\beta})$ being the usual Fisher matrix and $\boldsymbol{W}(\boldsymbol{\beta}) = \mathrm{diag}((\partial h(\hat{\eta}_1)/\partial\eta)^2/\sigma_1^2, \dots))$. By adding $\lambda$ to the diagonal of $\boldsymbol{F}(\boldsymbol{\beta})$, the matrix $\boldsymbol{F}_p(\boldsymbol{\beta})$ becomes invertible even if $\boldsymbol{F}(\boldsymbol{\beta})$ is not.

The penalty term $(\lambda/2)\sum_j \beta_j^2$ contains only one tuning parameter $\lambda$, which determines the amount of shrinkage for all $\beta_j$. Since the parameter $\beta_j$ depends on the scaling of the corresponding covariate $x_j$, solutions of $\boldsymbol{s}_p(\boldsymbol{\beta}) = \mathbf{0}$ are not equivariant under scaling of the covariates. Therefore, usually covariates are standardized before solving the estimation equation.

Ridge estimates have nice properties. Under weak conditions that hold for generalized linear models, estimates exist and are unique for $\lambda > 0$ (Fu, 1998). For small $\lambda$, the resulting estimate will be mildly biased and the covariance may be approximated by

$$\hat{\mathrm{cov}}(\hat{\boldsymbol{\beta}}) = (\boldsymbol{X}^T\boldsymbol{W}(\hat{\boldsymbol{\beta}})\boldsymbol{X} + \boldsymbol{\lambda}\boldsymbol{P})^{-1}(\boldsymbol{X}^T\boldsymbol{W}(\hat{\boldsymbol{\beta}})\boldsymbol{X})(\boldsymbol{X}^T\boldsymbol{W}(\hat{\boldsymbol{\beta}})\boldsymbol{X} + \lambda\boldsymbol{P})^{-1}.$$

Early attempts to generalized ridge regression were restricted to logistic regression; see Anderson and Blair (1982), Schaefer et al. (1984), and Duffy and Santner (1989). Ridge regression in generalized linear models has been investigated by Nyquist (1991), Segerstedt (1992), and LeCessie (1992).

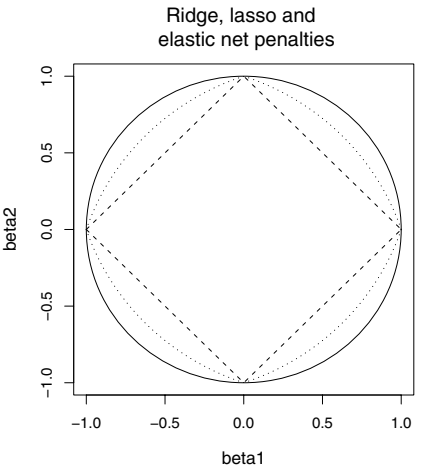


FIGURE 6.2: Constraint regions for the ridge penalty (circle), the lasso (diamond), and the elastic net (in between) for a two-dimensional predictor space.

### Example 6.1: Heart Disease

The selection of variables by regularization will be illustrated by the use of the heart disease data that are available from the R package *glmpath* (see Park and Hastie, 2007). The data contain 462 observations on 9 variables and the binary response coronary heart disease. The explanatory variables are sbp (systolic blood pressure), tobacco (cumulative tobacco), ldl (low density lipoprotein cholesterol), adiposity, famhist (family history of heart disease), typea (type A behavior), obesity, alcohol (current alcohol consumption), and age (age at onset). Figure 6.3 shows the coefficient buildups for the ridge estimate based on 10-fold

cross-validation (standardized explanatory variables, package *lqa*). Parameter estimates are not plotted against $\lambda$ but against $\|\boldsymbol{\beta}\|/\max\|\boldsymbol{\beta}\|$, where $\max\|\boldsymbol{\beta}\|$ denotes the maximum value that $\|\boldsymbol{\beta}\|$ can take. Therefore, small values of $\|\boldsymbol{\beta}\|$ correspond to large values of $\lambda$ and large values of $\|\boldsymbol{\beta}\|$ to small values of $\lambda$. For $\lambda = 0$ one obtains the ML estimates on the right-hand side. It is seen that, depending on the value of the smoothing parameter, the estimates are shrunk toward zero. □

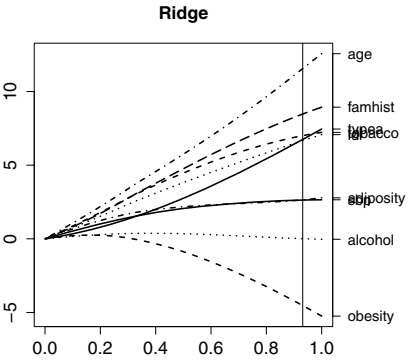


FIGURE 6.3: Coefficient paths for heart disease data when using ridge (package *lqa*; vertical line shows estimate selected by 10-fold cross-validation).

## 6.2.2 $L_1$-Penalty: The Lasso

Ridge regression often achieves better prediction performance than maximum likelihood based regression. However, ridge regression does not produce a parsimonious model, since all variables are retained. With a large number of predictors one often wants to determine a smaller subset that contains the strongest variables. Tibshirani (1996) proposed a new technique, called the lasso, for "least absolute shrinkage and selection operator", that shrinks some coefficients and sets others to 0. It tends to avoid the high variability of subset selection while producing a sparse model that shows good prediction performance. The lasso uses the $L_1$-penalty

$$J(\boldsymbol{\beta}) = \sum_{j=1}^{p} |\beta_j|, \tag{6.3}$$

which is a member of the bridge penalty family. In signal regression the $L_1$-penalization approach has also been called *basis pursuit* (Chen et al., 2001). It was originally proposed for the linear model in the constrained regression version, which means that the log-likelihood is maximized subject to the constraint

$$\sum_{j=1}^{p} |\beta_j| \leq t$$

for some $t$. The constrained regression version is helpful for illustrating why lasso often produces coefficients that are exactly zero.

Figure 6.2 shows the constraint regions for the ridge penalty, the lasso, and the elastic net (the latter will be introduced in the next section). The constraint region for the ridge penalty is

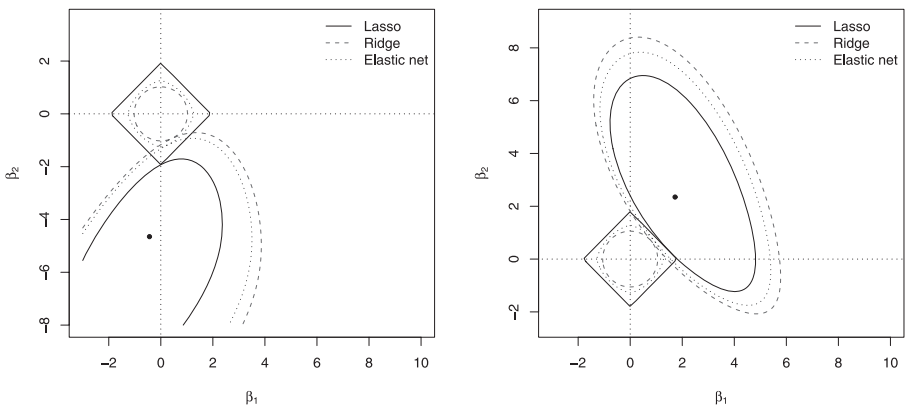


FIGURE 6.4: Constraint regions for the ridge penalty (circle), the lasso (diamond), and the elastic net (in between) together with the log-likelihood functions for negatively correlated predictors (left) and positively correlated predictors (right) for binary regression models.

the disk $\beta_1^2 + \beta_2^2 \leq t$; for the lasso one obtains the diamond $|\beta_1^2| + |\beta_2^2| \leq t$. Figure 6.4 shows the constraint regions for the ridge penalty and the lasso penalty in the two-dimensional case together with the contours of the likelihood function for a binary regression model. The left picture shows the contours of the log-likelihood function for a binary logit model with $n = 30$ and negatively correlated predictors ($\varrho = -0.87$), and the picture on the right shows the log-likelihood for $n = 30$ and positively correlated predictors ($\varrho = 0.87$). The point within the contours is the ML estimate for the specific dataset, and the penalized estimate is represented by the point where the contours touch the penalty region. Maximization of the likelihood subject to the constraint region yields an estimate that is closer to zero than the ML estimate. Of course, the amount of shrinkage depends on the size of the constraint region, which is determined by $t$ or, equivalently, by $\lambda$. The advantage of the lasso over ridge regression is that the constraint region is not smooth. Since the diamond has distinct corners, if a solution occurs at a corner, then one parameter is set to zero. The same happens in higher dimensions, but the constraint regions are harder to visualize. For three dimensions (see Figure 6.5), if the penalty region touches at a corner, two parameters are set to zero; if it touches at a connection between two corners, one parameter is set to zero.

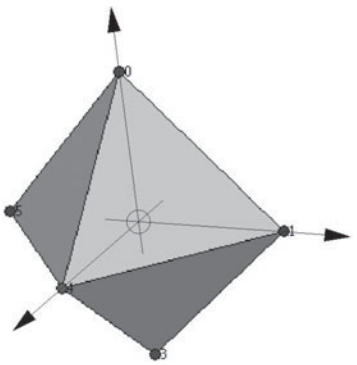

FIGURE 6.5: Constraint regions for the lasso (diamond) for a three-dimensional predictor space.

In contrast to linear regression, in binary regression the contours for a finite sample size are only approximately elliptical. An elliptical approximation is obtained by using a second-order Taylor approximation of the log-likelihood at the maximum likelihood estimate $\hat{\boldsymbol{\beta}}_{ML}$, obtaining

$$l(\boldsymbol{\beta}) \approx l(\hat{\boldsymbol{\beta}}_{ML}) + \frac{1}{2}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_{ML})^T \boldsymbol{F}(\hat{\boldsymbol{\beta}}_{ML})(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_{ML}),$$

where $\boldsymbol{F}(\hat{\boldsymbol{\beta}}_{ML})$ is the Fisher matrix. For normal regression models the approximation is exact and one obtains (apart from constants)

$$l(\boldsymbol{\beta}) = l(\hat{\boldsymbol{\beta}}_{ML}) - \frac{1}{2\sigma^2}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_{ML})^T \boldsymbol{X}^T \boldsymbol{X}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_{ML}),$$

and therefore a quadratic function centered at $\hat{\boldsymbol{\beta}}_{ML}$. For a large sample size, the contours of the likelihood also show an almost elliptical form for binary models.

By using the $L_1$-penalty (6.3) the lasso does both, continuous shrinkage and automatic variable selection, simultaneously. Concerning prediction error, it has been shown that the performance is not better than ridge regression in any case. In a comparison of the lasso, ridge, and bridge regression it has been shown that neither of them uniformly dominates the other two (see Tibshirani, 1996; Fu, 1998). The big advantage of the lasso is its sparse representation, which makes it attractive for practitioners.

The implicit shrinkage of the lasso can be illustrated by the idealized case of orthonormal columns in the design matrix of a linear model. Then the lasso penalty $\tilde{\lambda} J(\boldsymbol{\beta})$ yields the *soft thresholding* rule

$$\hat{\beta}_j = S(\beta_j^{ML}, \tilde{\lambda}) = \text{sign}(\hat{\beta}_j^{ML})(|\hat{\beta}_j^{ML}| - \tilde{\lambda})_+,$$

where $\beta_j^{ML}$ denotes the ML estimate of the $j$th component, $\tilde{\lambda} = \lambda\sigma^2$, and $(z)_+ = 1$, if $z \geq 0$, $(z)_+ = 0$, if $z < 0$. Figure 6.6 shows the estimated values as functions of the ML estimate. It is seen that estimates are set to zero if the ML estimate is below some threshold and are shrunk if the ML estimate is above the threshold. The term "soft thresholding" was built to distinguish it from hard thresholding, where estimates are set to zero if the ML estimate is below some threshold and retained if the ML estimate is above the threshold (also given in Figure 6.6). Moreover, estimates for SCAD are included, which will be considered in Section 6.2.5.

For the lasso in linear models, various computational procedures were proposed. Tibshirani (1996) used a combined quadratic programming method, Fu (1998) gave a modified Newton-Raphson algorithm and introduced the shooting algorithm, and Osborne et al. (2000) considered the lasso and its dual. Fan and Li (2001) proposed an alternative algorithm based on quadratic approximations. A fast implementation for large-scale logistic regression with the lasso has been presented by Genkin et al. (2004). Alternative approaches aim at the estimation of the entire path of the coefficient estimates as $\lambda$ varies, to find estimates $\hat{\boldsymbol{\beta}}(\lambda), 0 < \lambda < \infty$. Efron et al. (2004) proposed *LARS*, which determines the exact piecewise linear coefficient paths for lasso in the linear case. However, in the case of GLMs, the paths are not piecewise linear. Park and Hastie (2007) proposed an efficient path algorithm for generalized linear models that uses the predictor-corrector method. Rosset (2004) suggested a general path-following algorithm that can be used for any loss function. Zhao and Yu (2004) proposed a boosted lasso, which includes backward steps. An extremely fast algorithm based on pathwise coordinate optimization was given by Friedman et al. (2007), Friedman et al. (2010). It uses coordinate descent methods, which have been proposed earlier (for example Fu, 1998) but were not fully appreciated at the time.

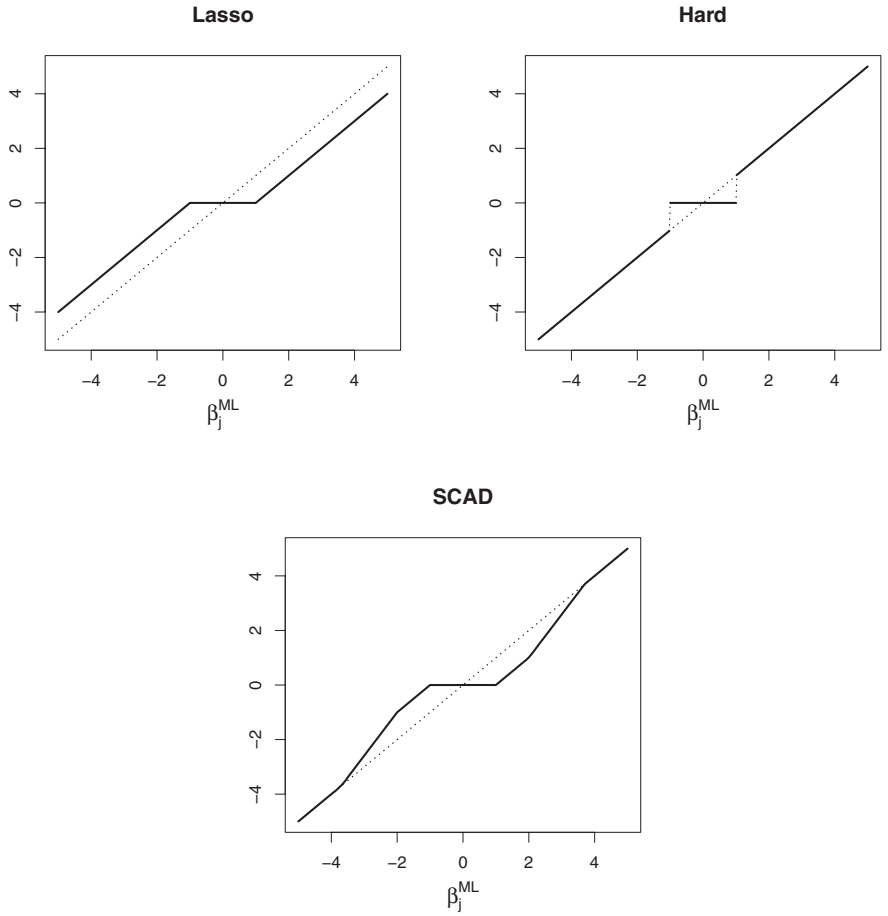


FIGURE 6.6: Estimates for lasso, hard thresholding, and SCAD when columns in the design matrix are orthonormal.

## The Adaptive Lasso

Variable selection procedures are often discussed in terms of *oracle* properties, which refer to the identification of the right subset model. For parameter vector $\boldsymbol{\beta}$ let $A = \{j : \beta_j \neq 0\}$ denote the active set, where $|A| = p_0 < p$. For simplicity, let $\boldsymbol{\beta}$ be partitioned into $\boldsymbol{\beta}^T = (\boldsymbol{\beta}_1^T, \boldsymbol{\beta}_2^T)$, where $\boldsymbol{\beta}_1$ represents the active set and $\boldsymbol{\beta}_2 = \mathbf{0}$. Then oracle properties means that (1) estimates $\hat{\boldsymbol{\beta}}^T = (\hat{\boldsymbol{\beta}}_1^T, \hat{\boldsymbol{\beta}}_2^T)$ must asymptotically satisfy $\hat{\boldsymbol{\beta}}_1 \neq \mathbf{0}$ and $\hat{\boldsymbol{\beta}}_2 = \mathbf{0}$, and (2) the optimal estimation rate is obtained, so that the estimator performs as well as if the underlying model were known. A *selection procedure* is *consistent* if asymptotically the right subset model is found, $\lim_n P(A_n = A) = 1$, where $A_n$ is the active set for $n$ observations. Zou (2006) showed that lasso variable selection can be inconsistent and gave necessary conditions for consistency. He proposed an extended version of lasso, for which the penalty has the form

$$J(\boldsymbol{\beta}) = \sum_{j=1}^{p} w_j |\beta_j|, \tag{6.4}$$

where $w_j$ are known weights. By using weights on coefficients the variables are not equally penalized, which adds some flexibility. He showed that for cleverly chosen data-dependent

weights the adaptive lasso has oracle properties. One choice of weights is based on a root-$n$ consistent estimator $\tilde{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$, for example, the ML estimate. Then weights are fixed by $w_j = 1/|\tilde{\beta}_j|^\gamma$, for fixed chosen $\gamma > 0$. The oracle properties that Zou (2006) derived for the adaptive lasso use that for growing sample size the weights for zero-coefficients get inflated, whereas the weights on non-zero-coefficients converge to a finite constant. Moreover, Zou (2006) showed that the adaptive lasso leads to near-minimax-optimal estimators.

### Example 6.2: Heart Disease

Figure 6.7 shows the coefficient buildups for the lasso and the adaptive lasso (standardized explanatory variables; package *lqa*) plotted against $\|\boldsymbol{\beta}\|/\max\|\boldsymbol{\beta}\|$. The vertical line shows the regularization obtained when using 10-fold cross-validation. It is seen that, in contrast to the ridge, not all variables are found to be influential. Based on 10-fold cross-validation one concludes that the variables alcohol and adiposity can be omitted. Moreover, it is seen that for this dataset the adaptive lasso is very close to the simple lasso. □

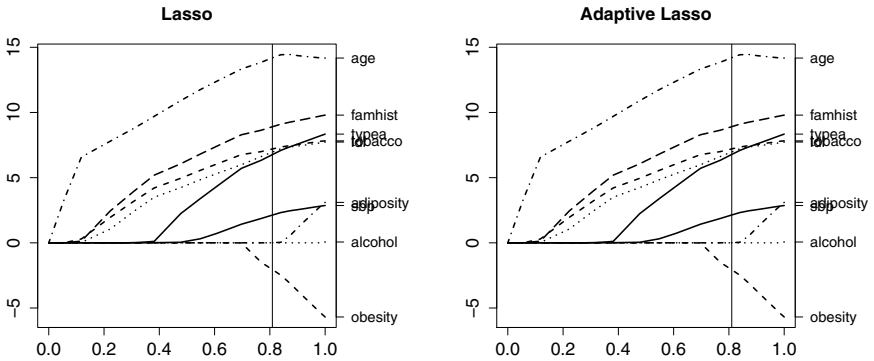


FIGURE 6.7: Lasso coefficient paths for heart disease data (package *lqa*; vertical line shows estimate selected by 10-fold cross-validation).

## Categorical Predictors and the Group Lasso

The lasso as considered in the previous section selects individual predictors. That approach is sensible when all variables are of the same type, for example, if all the variables are continuous, or all are binary. For a mixture of predictors, some of them categorical and some of them binary, the penalty is unsatisfactory. If the categorical predictor (factor) is represented by dummy variables, the lasso penalty selects individual dummy variables instead of whole factors, and the solution depends on how the dummy variables are encoded. A sensible procedure should select whole factors or continuous variables. The group lasso proposed by Yuan and Lin (2006) can overcome these problems.

Let the $p$-dimensional predictor be structured as $\boldsymbol{x}_i^T = (\boldsymbol{x}_{i1}^T, \ldots, \boldsymbol{x}_{i,G}^T)$, where $\boldsymbol{x}_{ij}$ corresponds to the $j$th *group of variables*. A group of variables may refer to the dummy variables of one factor, with $df_j$ denoting the number of the variables in the $j$th group. A continuous variable that has a linear form within the predictor obviously has $df_j = 1$. A group of variables may also refer to interactions between factors or between factors and continuous variables, where $df_j$ is

the number of individual interaction terms. Correspondingly the parameter vector is partitioned into subvectors, $\boldsymbol{\beta}^T = (\boldsymbol{\beta}_1^T, \ldots, \boldsymbol{\beta}_G^T)$. The *group lasso* uses the penalty

$$J(\boldsymbol{\beta}) = \sum_{j=1}^{G} \sqrt{df_j} \|\boldsymbol{\beta}_j\|_2,$$

where $\|\boldsymbol{\beta}_j\|_2 = (\beta_{j1}^2 + \cdots + \beta_{j,df_j}^2)^{1/2}$ is the $L_2$-norm of the parameters of the $j$th group. The penalty encourages sparsity in the sense that either $\hat{\boldsymbol{\beta}}_j = \mathbf{0}$ or $\beta_{js} \neq 0$ for $s = 1, \ldots, df_j$. For a geometrical interpretation of the penalty, see Yuan and Lin (2006). Meier et al. (2008) showed that under sparsity the resulting estimates are consistent even when the number of predictors is larger than the sample size. The penalty may be seen as a special case of the composite absolute penalty family proposed by Zhao et al. (2009).

### 6.2.3 The Elastic Net

Although the lasso has several advantages, it has some severe limitations, pointed out by Zou and Hastie (2005). If there are high correlations between predictors, it has been observed that ridge regression dominates the lasso (Tibshirani, 1996). As a variable selection method it is restricted to $n$ variables. In the $p > n$ case, the lasso selects at most $n$ variables before it saturates. Moreover, the lasso does not necessarily have a unique solution, since the penalty term is not strictly convex. A point of special interest concerns the variables that are selected by the lasso. If there is a group of variables among which the correlations are very high, the lasso tends to select only one of the variables as a representative. In particular, this way of selecting variables is different from the *elastic net*, proposed by Zou and Hastie (2005). The elastic net does automatic variables selection, and, rather than selecting one representative, it can select *groups* of correlated variables. According to Zou and Hastie it works "like a stretchable fishing net that retains all the big fish." The elastic net uses the elastic net criterion

$$J(\boldsymbol{\beta}) = \lambda_1 \sum_{j=1}^{p} |\beta_j| + \lambda_2 \sum_{j=1}^{p} \beta_j^2, \tag{6.5}$$

which depends on two tuning parameters, $\lambda_1, \lambda_2 > 0$. The elastic net penalty is a convex combination of the lasso and the ridge penalty. In constraint form it may be written as $(1 - \alpha) \sum_j |\beta_j| + \alpha \sum_j \beta_j^2 \leq t$ for some $t$ and tuning parameter $\alpha = \lambda_2/(\lambda_1 + \lambda_2)$. With $\alpha \in [0, 1]$ the lasso and ridge are limiting cases. For illustration, the contour plots of the elastic net penalty, the lasso, and the ridge are shown in Figure 6.2. Zou and Hastie (2005) called (6.5) the naïve elastic net criterion and proposed a rescaled solution $\hat{\boldsymbol{\beta}} = (1 + \lambda_2)\hat{\boldsymbol{\beta}}_{net}$, where $\hat{\boldsymbol{\beta}}_{net}$ is the penalized least-squares solution of the naïve elastic net criterion. They give several reasons for choosing $(1 + \lambda_2)$ as scaling factor.

The interesting property of the elastic net is the *grouping effect*. A regression method exhibits the grouping effect if the regression coefficients of a group of highly correlated variables tend to be equal, up to a change of sign if negatively correlated. Zou and Hastie (2005) show that for penalized least-squares problems the coefficient paths of predictors $x_i$ and $x_j$ with sample correlation $\varrho_{ij}$ are confined by

$$|\hat{\beta}_i - \hat{\beta}_j| / \sum_i |y_i| \leq \frac{1}{\lambda_2} \sqrt{2(1 - \varrho_{ij})},$$

where $\hat{\beta}_i$, $\hat{\beta}_j$ are naïve net solutions with parameters $\lambda_1$, $\lambda_2$. If $x_i$ and $x_j$ are highly correlated, $(\varrho_{ij} \to 1)$, the coefficient paths of $x_i$ and $x_j$ are very close. Thus the elastic net shows the

grouping the effect, which is important, for example, in genetics, where groups of genes that are relevant are to be selected ("grouped selection").

To illustrate why the grouping effect is useful we use the idealized example given by Zou and Hastie (2005). With $Z_1$ and $Z_2$ being two independent $\mathrm{U}(0, 20)$ variables, the response is generated as $\mathrm{N}(Z_1 + 0.1Z_2, 1)$. It is assumed that one observes only noisy versions of $Z_1$ and $Z_2$:

$$x_1 = Z_1 + \epsilon_1, \quad x_2 = -Z_1 + \epsilon_2, \quad x_3 = Z_1 + \epsilon_3,$$
$$x_4 = Z_2 + \epsilon_4, \quad x_5 = -Z_2 + \epsilon_5, \quad x_6 = Z_2 + \epsilon_6,$$

where $\epsilon_i$ are independent identically distributed $\mathrm{N}(0, 1/16)$. The variables $x_1, x_2$, and $x_3$ may be considered as forming one group and $x_4, x_5$, and $x_6$ as forming a second group. Figure 6.8 shows the coefficient buildups for the lasso and a method that shows the grouping effect for sample size $n = 100$. The method used is BlockBoost, which is described in the next section. It is seen that BlockBoost selects the variables $x_1, x_2$, and $x_3$, and the corresponding estimates are (up to sign) identical. The strong group consistency of $x_1, x_2$, and $x_3$ is distinctly identified. Lasso shows quite different coefficient buildups, selecting as strongly influential the variables $x_1$ and $x_3$ and, with rather weak effect, $x_2$. While the coefficient paths for BlockBoost reflect the high correlation of $x_1, x_2$ and $x_3$, the paths of the lasso are rather irregular. The elastic net behaves quite similar to BlockBoost (compare Zou and Hastie, 2005).

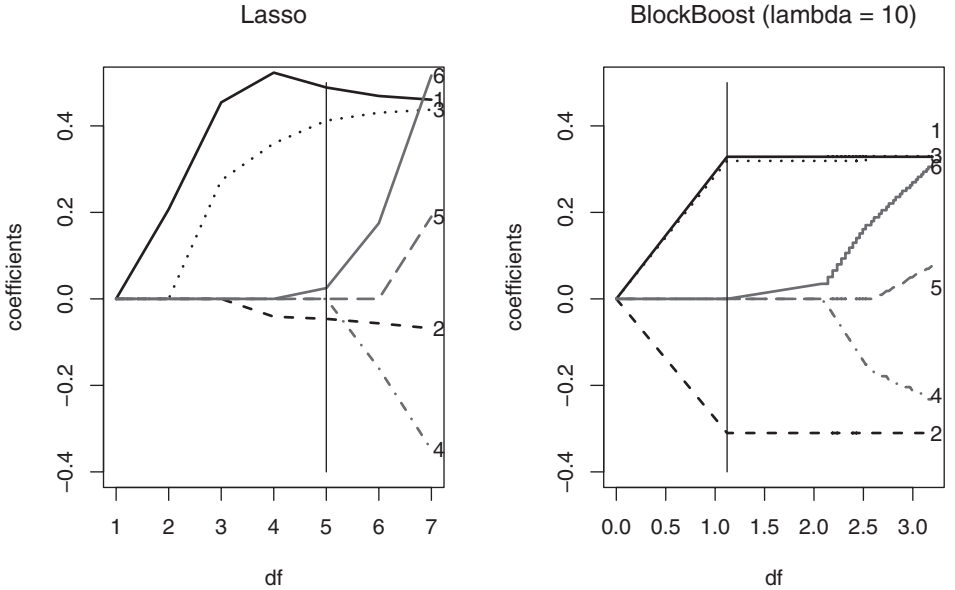


FIGURE 6.8: Coefficient buildups for lasso (left) and BlockBoost (right) of the hidden factors example. Vertical lines indicate the degrees of freedom corresponding to the tuning parameter(s) chosen by 10-fold cross-validation.

## 6.2.4 Alternative Estimators with Grouping Effect

The grouping effect of the elastic net gives similar coefficients to highly correlated variables. More recently, alternative penalties were proposed that aim at the grouping effect.

### OSCAR

Bondell and Reich (2008) proposed a method called OSCAR, for Octagonal Shrinkage and Clustering Algorithm for Regression. The method shrinks, like the lasso, some coefficients to zero, but in addition yields the exact equality of some of the coefficients. The predictors with equal coefficients form clusters that are represented by a single coefficient. OSCAR in constrained form uses the restriction

$$\sum_j |\beta_j| + c \sum_{j<k} \max\{|\beta_j|, |\beta_k|\} \leq t$$

for some tuning parameters $t > 0, c \geq 0$. The parameter $c$ controls the relative weighting of the $L_1$-norm and the pairwise $L_\infty$-norm. With $c = 0$ the lasso is a special case. For two predictors, the constraint region forms an octagon. The vertices on the diagonals and on the axis encourage equality of coefficients (when a vertice on the diagonal is hit) and sparsity (when a vertice on the axis is hit). Varying $c$ changes the angle formed in the octagon, yielding a diamond if $c = 0$ and a square if $c \to \infty$.

The exact grouping property derived by Bondell and Reich (2008) uses parameters $\lambda$ and $c$ from the penalized version $J(\boldsymbol{\beta}) = \lambda\{\sum_j |\beta_j| + c\sum_{j<k} \max\{|\beta_j|, |\beta_k|\}$. For signed coefficients, so that $\hat{\beta}_j \geq 0$ for all $j$, they show that for the linear model there exists a $c_0$ such that

$$0 \leq c_0 \leq 2\lambda^{-1}|\boldsymbol{y}|\sqrt{2(1-\varrho_{jk})}$$

and $\hat{\beta}_j = \hat{\beta}_k$ for all $c \geq c_0$. Here $\varrho_{jk}$ denotes the correlation between the variables $x_j$ and $x_k$, and the response vector $\boldsymbol{y}$ is centered. Therefore, there is a threshold on $c$, which can be very small when $\varrho_{jk}$ is close to one, such that the coefficients are equal and therefore form a cluster. A representation of OSCAR's penalty region as a polytope is found in Petry and Tutz (2012). The OSCAR for GLMs was considered by Petry and Tutz (2011).

#### Example 6.3: Heart Disease

For elastic net and OSCAR, which contain the lasso as a special case, the coefficient paths based on 10-fold cross-validation for the heart disease data are very close to the paths found for lasso. To illustrate the grouping effect we show the coefficient paths for $c = 0.2$ and $c = 0.5$, which enforce the grouping property. Figure 6.9 shows the resulting coefficient buildups. It is seen that for strong smoothing some effects are set equal. □

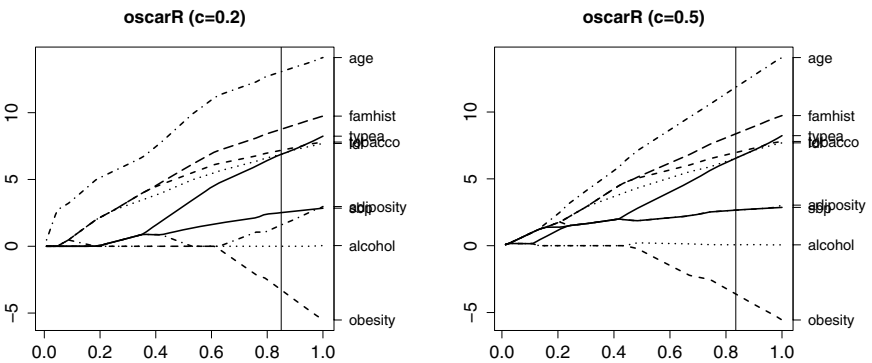


FIGURE 6.9: OSCAR coefficient paths for heart disease data (two values of c).

### Correlation-Based Penalties

An alternative approach that aims at the grouping effect uses explicitly the correlation between pairs of predictors. The correlation-based penalty has the form

$$J_c(\boldsymbol{\beta}) = \sum_{i=1}^{p-1} \sum_{j>i} \left\{ \frac{(\beta_i - \beta_j)^2}{1 - \varrho_{ij}} + \frac{(\beta_i + \beta_j)^2}{1 + \varrho_{ij}} \right\} = 2 \sum_{i=1}^{p-1} \sum_{j>i} \frac{\beta_i^2 - 2\varrho_{ij}\beta_i\beta_j + \beta_j^2}{1 - \varrho_{ij}^2}, \qquad (6.6)$$

where $\varrho_{ij}$ denotes the (empirical) correlation between the $i$th and the $j$th predictors. It is designed to focus on the grouping effect, that is, its highly correlated effects show comparable values of estimates ($|\hat{\beta}_i| \approx |\hat{\beta}_j|$) with the sign being determined by positive or negative correlation. For strong positive correlation ($\varrho_{ij} \to 1$) the first term becomes dominant, having the effect that estimates for $\beta_i, \beta_j$ are similar ($\hat{\beta}_i \approx \hat{\beta}_j$). For strong negative correlation ($\varrho_{ij} \to -1$) the second term becomes dominant and $\hat{\beta}_i$ will be close to $-\hat{\beta}_j$. Consequently, for weakly correlated data the performance is quite close to the ridge penalty.

Figure 6.10 shows the two-dimensional contour plots for selected values of $\varrho$ together with the constraint region for the ridge penalty and the lasso. It is seen that contours for the ridge and lasso are highly symmetric; $x_1 = 0$ is an axis of symmetry as well as $x_2 = 0$. In contrast, the constrained region for the correlation-based estimator is an ellipsoid that becomes narrower with increasing correlation. Spectral decomposition of $J_c(\beta)$ yields eigenvectors $(1, 1)$ and $(1, -1)$ with corresponding eigenvalues $\lambda/(1 - \varrho)$ and $\lambda/(1 + \varrho)$. Thus, for $\varrho > 0$, the first eigenvalue becomes dominant while for $\varrho < 0$ it is the second eigenvalue that determines the orientation of the ellipsoid. When computing the penalized least-squares criterion, the effect is that, for $\varrho > 0$, estimates are preferred for which the components $\hat{\beta}_1, \hat{\beta}_2$ are similar; for $\varrho < 0$, similarity of $\hat{\beta}_1$ and $-\hat{\beta}_2$ is preferred. This may be seen from the contour plots, since for $\varrho > 0$ the increase in $P_c(\beta)$ is slower when moving in the direction of the first eigenvector $(1, 1)$ than in the orthogonal direction $(1, -1)$. For $\varrho < 0$, the eigenvalue corresponding to $(1, -1)$ is larger, and therefore parameter values where $\beta_1$ is close to $-\beta_2$ are preferred. Thus the use of penalty $P_c$ implies shrinkage, with the strength of shrinkage being determined by $\lambda$, but shrinkage differs from ridge shrinkage, which occurs for the special case $\varrho_{ij} = 0$.

Assume that $\lambda > 0$ and $\varrho_{ij}^2 \neq 1$ for $i \neq j$. Then $J_c(\boldsymbol{\beta})$ is strictly convex and the estimate exists and is unique. For linear models an explicit solution to the penalized least-squares problem is obtained, called the *correlation-based estimator*:

$$\hat{\boldsymbol{\beta}}_c = (\boldsymbol{X}^T\mathbf{X} + \lambda \boldsymbol{M})^{-1}\boldsymbol{X}^T\boldsymbol{y}, \qquad (6.7)$$

where $\mathbf{X}^T = (\mathbf{x_1} \ldots \mathbf{x_n})$ is the design matrix, $\mathbf{y}$ collects the responses, $\mathbf{y}^T = (y_1, \ldots, y_n)$, and $\boldsymbol{M}$ is a matrix that is determined by the correlations $\varrho_{ij}, i, j = 1, \ldots, p$. The explicit form exploits that the correlation-based penalty (9.7) can be written as a quadratic form:

$$J_c(\boldsymbol{\beta}) = \boldsymbol{\beta}^T \boldsymbol{M} \boldsymbol{\beta}, \qquad (6.8)$$

where $\boldsymbol{M} = (m_{ij})$ has entries $m_{ij} = 2\sum_{s \neq i} 1/(1 - \varrho_{is}^2)$ if $i = j$, and $m_{ij} = -2\varrho_{ij}/(1 - \varrho_{ij}^2)$ if $i \neq j$. For GLM-type models the ML estimate is obtained by penalized Fisher scoring (see Section 6.2.1). Although the correlation-based penalty enforces the grouping effect with good results in simulations, it does not enforce sparsity. Therefore, Tutz and Ulbricht (2009) proposed a specific form of blockwise boosting, called BlockBoost. To obtain the grouping effect of the correlation-based estimator combined with variable selection, a boosting procedure is used that updates in each step the coefficients of more than one variable. The procedure differs from common componentwise boosting, where just one variable is selected and the corresponding coefficient is adjusted. The algorithm is able to handle high-dimensional data and,

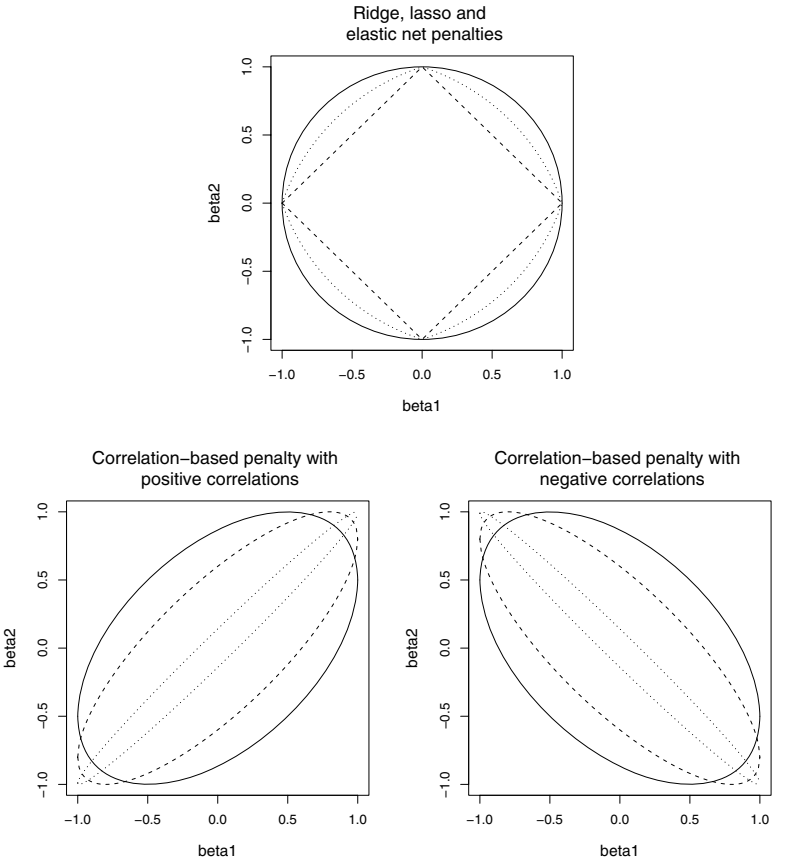


FIGURE 6.10: Top panel: Two-dimensional contour plots for ridge, lasso (dashed line), and elastic net with $\alpha = 0.5$ (dotted line). Lower panel left: Contour plots of correlation-based penalty for positive correlations: $\varrho = 0.5$ (solid line), $\varrho = 0.8$ (dashed line), and $\varrho = 0.99$ (dotted line). Lower panel right: Contour plots of correlation-based penalty for negative correlations: $\varrho = -0.5$ (solid line), $\varrho = -0.8$ (dashed line), and $\varrho = -0.99$ (dotted line).

depending on the tuning parameter, enforces sparsity. The grouping effect is demonstrated in Figure 6.8, where coefficients are plotted against degrees of freedom for the simulation described in Section 6.2.3. It is seen that lasso fails to recognize the grouping structure in contrast to BlockBoost. For more details on the correlation-based approach in linear models, see Tutz and Ulbricht (2009). GLM-type models were considered by Ulbricht and Tutz (2008). An alternative way is to combine correlation-based penalties and the $L_1$-penalty into the form

$$J_c(\boldsymbol{\beta}) = \lambda_1 \sum_{j=1}^{p} |\beta_j| + \lambda_2 \boldsymbol{\beta}^T \boldsymbol{M} \boldsymbol{\beta}. \tag{6.9}$$

Anbari and Mkhadri (2008) demonstrated that the penalty shows good performance in many applications.

### Fusion-Type Estimators

Daye and Jeng (2009) proposed the penalty

$$J(\boldsymbol{\beta}) = \lambda_1 \sum_{j=1}^{p} |\beta_j| + \lambda_2 \sum_{i<j} w_{ij}(\beta_i - \text{sign}(\varrho_{ij})\beta_j)^2,$$

where $\text{sign}(\varrho_{ij})$ denotes the sign of the correlation coefficient $\varrho_{ij}$ taking values 1 or –1. The penalty combines the lasso with a term that enforces the fusion of variables. When $\lambda_2$ is large and weights are positive, the second term enforces $\hat{\beta}_i \approx \hat{\beta}_j$ for positive correlation between $x_i$ and $x_j$ and $\hat{\beta}_i \approx -\hat{\beta}_j$ for negative correlation. Because of the tendency to fuse predictors, Daye and Jeng (2009) call the resulting estimator a *weighted fusion estimator*. They use the correlation-driven weights $w_{ij} = |\varrho_{ij}|^\gamma/(1 - |\varrho_{ij}|^\gamma)$, where $\gamma$ is an additional tuning parameter, and derive conditions that make the estimator sign consistent for the linear model. Sign consistency is somewhat stronger than variable selection consistency and postulates that asymptotically the sign of the estimate is the same as the sign of the true parameter.

The penalty belongs to the general family of combination penalties

$$J(\boldsymbol{\beta}) = \lambda_1 \|\boldsymbol{\beta}\|_1 + \lambda_2 \|\boldsymbol{\beta}\|_{\boldsymbol{R}}{}^2, \tag{6.10}$$

where $\|\boldsymbol{\beta}\|_1 = (|\beta_1| + \cdots + |\beta_p|)$ is the $L_1$-norm and $\|\boldsymbol{\beta}\|^2_{\boldsymbol{R}} = \boldsymbol{\beta}^T \boldsymbol{R}\boldsymbol{\beta}$ is the squared norm built with matrix $\boldsymbol{R}$. For $\boldsymbol{R} = \boldsymbol{I}$ one obtains the elastic net, for $\boldsymbol{R} = \boldsymbol{M}$ one obtains (6.9), and for specific $\boldsymbol{R}$ one obtains the weighted fusion estimator. It should be noted that all combination penalties (6.10) can be reformulated as lasso problems by simple data augmentation, and therefore algorithms that compute lasso solutions can be used (see, for example, Zou and Hastie, 2005).

In their derivation of the penalty Daye and Jeng (2009) refer to the fused lasso, which was proposed by Tibshirani et al. (2005). However, the latter enforces a fusion of predictors by using a lasso-type estimator instead of a ridge-type estimator for the differences. The corresponding *pairwise fused lasso* estimator has the form

$$J(\boldsymbol{\beta}) = \lambda_1 \sum_{j=1}^{p} |\beta_j| + \lambda_2 \sum_{i<j} |\beta_i - \beta_j|,$$

or with weights and correlation-based penalty,

$$J(\boldsymbol{\beta}) = \lambda_1 \sum_{j=1}^{p} |\beta_j| + \lambda_2 \sum_{i<j} w_{ij}|\beta_i - \text{sign}(\varrho_{ij})\beta_j|.$$

For applications see Petry et al. (2011). The fused lasso itself is considered in more detail in Section 10.4.4.

## 6.2.5 SCAD

An alternative penalty that yields simultaneous estimation and selection has been proposed by Fan and Li (2001). They identified three properties that a penalized estimator should have and derived an appropriate penalty. The resulting estimator should be nearly unbiased for large unknown coefficients (unbiasedness), it should automatically set small estimated coefficients to

zero (sparsity), and it should be continuous in the data to avoid instability in model prediction (continuity). The penalty function considered by Fan and Li (2001) has the additive form

$$\lambda J(\boldsymbol{\beta}) = \lambda \sum_{j=1}^{p} p(|\beta_j|) = \sum_{j=1}^{p} p_\lambda(|\beta_j|),$$

where $p(.)$ are penalty functions that, in the most general form, can also depend on the variable. The functions $p_\lambda(|\beta|) = \lambda p(|\beta|)$ are introduced to allow that the penalty may depend on $\lambda$. Obviously, ridge and lasso are special cases with functions $p(|\beta_j|) = |\beta_j|^2$ and $p(|\beta_j|) = |\beta_j|$, respectively. The continuously differentiable penalty function proposed by Fan and Li is the *smoothly clipped absolute derivation (SCAD) penalty*, defined by its derivative:

$$p'_\lambda(\beta) = \lambda\{I(\beta \leq \lambda) + \frac{(a\lambda - \beta)_+}{(a-1)\lambda} I(\beta > \lambda)\},$$

for some $a > 2$ and $\beta > 0$, where $(x)_+ = x$ if $x > 0$ and 0 otherwise. The penalty corresponds to a quadratic spline function with knots at $\lambda$ and $a\lambda$. Figure 6.11 shows the penalty and its derivative. It is seen that for small values the penalty is similar to the lasso penalty whereas for larger values the penalty levels off.

To gain some insight about the effect of penalty functions, it is common to study the linear regression model with orthonormal columns in the design matrix. Then it may be shown that the lasso penalty $p(|\beta|) = \lambda|\beta|$ yields the soft thresholding rule $\hat{\beta}_j = \text{sign}(\hat{\beta}_j^{ML})(|\hat{\beta}_j^{ML}| - \lambda)_+$, where $\beta_j^{ML}$ denotes the ML estimate and SCAD yields

$$\hat{\beta}_j = \begin{cases} \text{sign}(\hat{\beta}_j^{ML})(|\hat{\beta}_j^{ML}| - \lambda)_+ & \text{if} \quad |\hat{\beta}_j^{ML}| < 2\lambda \\ \{(a-1)\hat{\beta}_j^{ML} - \text{sign}(\beta_j^{ML})a\lambda\}/a - 2 & \text{if} \quad 2\lambda < |\hat{\beta}_j^{ML}| \leq a\lambda \\ \hat{\beta}_j^{ML} & \text{if} \quad |\hat{\beta}_j^{ML}| > a\lambda \end{cases}$$

(see Fan and Li, 2001). Usually these penalties are compared to the hard thresholding penalty function $p_\lambda(|\beta|) = \lambda^2 - (|\beta| - \lambda)^2 I(|\beta| < \lambda)$, which sets the estimate to zero if the ML estimate is below some threshold and retains the ML estimate if it is above the threshold, that is, $\hat{\beta}_j = \hat{\beta}_j^{ML} I(|\hat{\beta}_j^{ML}| > \lambda)$. It is seen from Figure 6.6 that SCAD and hard thresholding avoid bias for large coefficients, in contrast to lasso. SCAD shares with the lasso the continuity of the resulting function (for details see Fan and Li, 2001).

For generalized linear models, the penalized log-likelihood has to be minimized. Since the SCAD penalty functions are singular at the origin and do not have continuous second-order derivations for the computation, Fan and Li use that they can be locally approximated by a quadratic function. The proposed local quadratic approximations also applies to the lasso and hard thresholding penalties.

An advantage of the SCAD penalty is its oracle property. Let the parameter vector $\boldsymbol{\beta}$ be partitioned into $\boldsymbol{\beta}^T = (\boldsymbol{\beta}_1^T, \boldsymbol{\beta}_2^T)$ and assume $\boldsymbol{\beta}_2 = \mathbf{0}$. With $I(\boldsymbol{\beta})$ denoting the full Fisher matrix and $J_1(\boldsymbol{\beta}_1)$ the Fisher matrix and knowing $\boldsymbol{\beta}_2 = \mathbf{0}$, it may be shown that $\hat{\boldsymbol{\beta}}^T = (\hat{\boldsymbol{\beta}}_1^T, \hat{\boldsymbol{\beta}}_2^T)$ must asymptotically satisfy $\hat{\boldsymbol{\beta}}_2 = \mathbf{0}$ and $\hat{\boldsymbol{\beta}}_1$ is asymptotic normal with covariance matrix $J_1(\boldsymbol{\beta}_1)^{-1}$ if $n^{1/2}\lambda_n \to \infty$. This means that asymptotically the estimator performs as well as if $\boldsymbol{\beta}_2 = \mathbf{0}$ were known. The penalized estimate is root-$n$ consistent if $\lambda_n \to 0$ and converges at the rate $O_p(n^{-1/2} + a_n)$, where $a_n = \max\{p'_{\lambda_n}(|\beta_j|), \beta_j \neq 0\}$.

The penalized score function has the form $s_p(\boldsymbol{\beta}) = s(\boldsymbol{\beta}) - \Sigma_j p'_\lambda(|\beta_j|)$, where $s(\boldsymbol{\beta})$ is the usual score function of a GLM. The derivative on the first term on the right-hand side yields the usual negative information matrix. The derivative of the second term has to be approximated

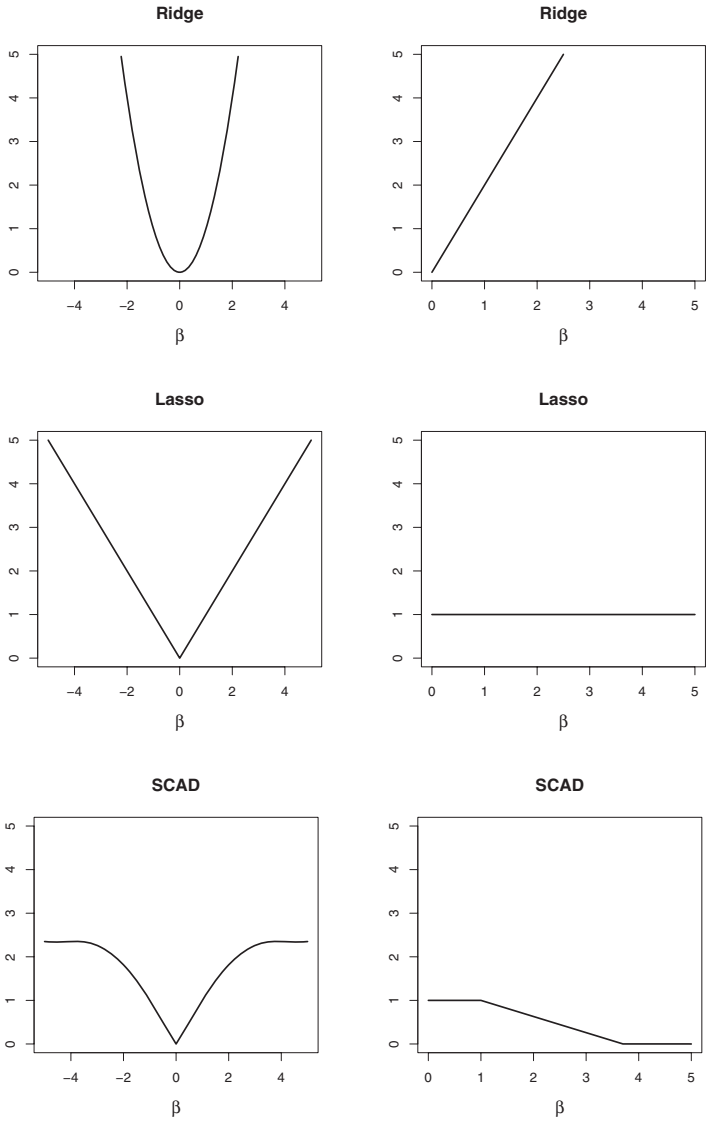


FIGURE 6.11: Components $p_\lambda(\beta)$ (left) and derivations $p'_\lambda(\beta)$ (right) of penalty functions for ridge, lasso, and SCAD.

because derivatives do not exist. Fan and Li (2001) used a quadratic approximation based on an initial estimate $\boldsymbol{\beta}_0$, which is given by $p_\lambda(|\beta_j|) = p'_\lambda(|\boldsymbol{\beta}_j|)\,\mathrm{sign}(\beta_j) \approx \{p'_\lambda(|\beta_{0j}|)/|\beta_{0j}|\}\beta_j$ when $\beta_j \neq 0$. Let $\hat{\boldsymbol{\beta}}_1^T = (\hat{\beta}_{11}, \ldots \hat{\beta}_{1p})$ denote the non-vanishing components of $\hat{\boldsymbol{\beta}}$, obtained from $s_p(\hat{\boldsymbol{\beta}}) = 0$. Then the corresponding sandwich formula that approximates the covariance of $\hat{\boldsymbol{\beta}}_1$ is

$$\hat{\mathrm{cov}}(\hat{\boldsymbol{\beta}}) \approx [I(\boldsymbol{\beta}_1) + P_\lambda(\hat{\boldsymbol{\beta}}_1)]^{-1}\hat{\mathrm{cov}}(sp(\hat{\beta}_1))[I(\boldsymbol{\beta}_1) + P_\lambda(\hat{\boldsymbol{\beta}}_1)]^{-1},$$

where $I(\boldsymbol{\beta}_1) = -\partial l(\hat{\boldsymbol{\beta}}_1)/\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T$ and $P_\lambda(\hat{\boldsymbol{\beta}}_1) = \mathrm{diag}(p'_\lambda(|\beta_{01}|)/|\beta_{01}|, \ldots, p'_\lambda(|\beta_{0p}|)/|\beta_{0p}|)$. According to Fan and Li (2001), the formula has good accuracy for moderate sample sizes.

### 6.2.6 The Dantzig Selector

The Dantzig selector was proposed by Candes and Tao (2007) and generalized to GLMs by James and Radchenko (2008). Like the lasso, it obtains variable selection by using an $L_1$ penalty to shrink the coefficients toward zero. However, the penalty is used in a different way. For simplicity, let the response function $h$ of the model $\mu = h(\boldsymbol{x}^T\boldsymbol{\beta})$ be the canonical response function. Then the generalized Dantzig selector criterion is

$$\min\|\tilde{\boldsymbol{\beta}}\|_1 \quad \text{subject to} \quad |\boldsymbol{x}_{\cdot j}^T(\boldsymbol{y} - \tilde{\boldsymbol{\mu}})| \leq \lambda, j = 1, \ldots, p,$$

where $\|\tilde{\boldsymbol{\beta}}\|_1 = |\tilde{\boldsymbol{\beta}}_1| + \cdots + |\tilde{\boldsymbol{\beta}}_p|$ represents, the $L_1$-norm, $\boldsymbol{x}_{\cdot j}^T = (x_{1j}, \ldots, x_{nj})$ is the $j$th column of the design matrix, and $\boldsymbol{y}$, $\tilde{\boldsymbol{\mu}}$ denote the vector of observations and fitted values, respectively. The constraint region defined by $|\boldsymbol{x}_{\cdot j}^T(\boldsymbol{y} - \tilde{\boldsymbol{\mu}})| \leq \lambda$, where $\lambda$ is the tuning parameter, is based on the score function, which with a canonical link has the form $\boldsymbol{s}(\boldsymbol{\beta}) = \boldsymbol{X}^T(\boldsymbol{y} - \boldsymbol{\mu})$. Therefore, $\boldsymbol{x}_{\cdot j}^T(\boldsymbol{y} - \boldsymbol{\mu})$ represents the $j$th component of the score function. While ML estimates are obtained when $\boldsymbol{x}_{\cdot j}^T(\boldsymbol{y} - \hat{\boldsymbol{\mu}}) = 0$ for all $j$, the constraint $|\boldsymbol{x}_{\cdot j}^T(\boldsymbol{y} - \tilde{\boldsymbol{\mu}})| \leq \lambda$ represents a weaker condition depending on the value of $\lambda$.

One of the strengths of the Dantzig selector is that for linear models it can be formulated as a linear programming problem and therefore also can be efficiently computed for high-dimensional problems. For linear problems, the constraint region has the simple form $|\boldsymbol{x}_{\cdot j}^T(\boldsymbol{y} - \tilde{\boldsymbol{X}}\tilde{\boldsymbol{\beta}})| \leq \lambda$. For GLMs, one can use that ML estimates are obtained iteratively by a weighted least-squares algorithm. For a general link function, the score function has the form $\boldsymbol{s}(\boldsymbol{\beta}) = \boldsymbol{X}^T\boldsymbol{D}\,\boldsymbol{\Sigma}^{-1}(\boldsymbol{y} - \boldsymbol{\mu})$, where $\boldsymbol{D} = \text{Diag}\,(\partial h(\eta_1)/\partial\eta, \ldots \partial h(\eta_n)/\partial\eta)$ is the diagonal matrix of derivatives. With weight matrix $\boldsymbol{W} = \boldsymbol{D}\,\boldsymbol{\Sigma}^{-1}\boldsymbol{D}^T$, one step of the Fisher scoring iteration for obtaining ML estimates has the form

$$\hat{\boldsymbol{\beta}}^{(k+1)} = (\boldsymbol{X}^T\boldsymbol{W}(\hat{\boldsymbol{\beta}}^{(k)})\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{W}(\hat{\boldsymbol{\beta}}^{(k)})\tilde{\boldsymbol{\eta}}(\hat{\boldsymbol{\beta}}^{(k)})$$

with a vector of pseudo-observations $\tilde{\boldsymbol{\eta}}(\hat{\boldsymbol{\beta}}) = \boldsymbol{X}\hat{\boldsymbol{\beta}} + \boldsymbol{D}(\hat{\boldsymbol{\beta}})^{-1}(\boldsymbol{y} - \hat{\boldsymbol{\mu}})$. The implicitly used weighted least-squares estimate is equivalent to a least-squares estimate for a design matrix $\boldsymbol{W}(\hat{\boldsymbol{\beta}}^{(k)})^{1/2}\boldsymbol{X}$ and pseudo-response $\boldsymbol{W}(\hat{\boldsymbol{\beta}}^{(k)})^{1/2}\tilde{\boldsymbol{\eta}}(\hat{\boldsymbol{\beta}}^{(k)})$. Therefore, instead of solving the score equations, one uses an iterative procedure that, for given parameter $\hat{\boldsymbol{\beta}}^{(k)}$, computes the linear model Dantzig selector:

$$\min\|\tilde{\boldsymbol{\beta}}\|_1 \quad \text{subject to} \quad |\boldsymbol{x}_{\cdot j}^T\boldsymbol{W}(\hat{\boldsymbol{\beta}}^{(k)})(\tilde{\boldsymbol{\eta}}(\hat{\boldsymbol{\beta}}^{(k)}) - \boldsymbol{X}\tilde{\boldsymbol{\beta}})| \leq \lambda, j = 1, \ldots, p,$$

yielding the new estimate $\hat{\boldsymbol{\beta}}^{(k+1)}$. James and Radchenko (2008) also gave an algorithm for fitting the generalized Dantzig selector path. By computing the whole path efficiently, cross-validation on a fine grid can be performed. Efficient computing is needed since James and Radchenko (2008) use two shrinkage parameters. The reason is that the Dantzig selector tends to overshrink the coefficients. If strong shrinkage (large $\lambda$) is applied so that noisy variables are excluded, the estimates of coefficients are too small. If small $\lambda$ is selected, noisy variables tend to be included.

#### Example 6.4: Heart Disease

Figure 6.12 shows the coefficient buildups for SCAD ($a = 3$; package *lqa*) and the Dantzig selector (standardized explanatory variables ) plotted against $\|\boldsymbol{\beta}\|/\max\|\boldsymbol{\beta}\|$. Although the paths are differing, selection based on 10-fold cross-validation yields similar coefficients. □

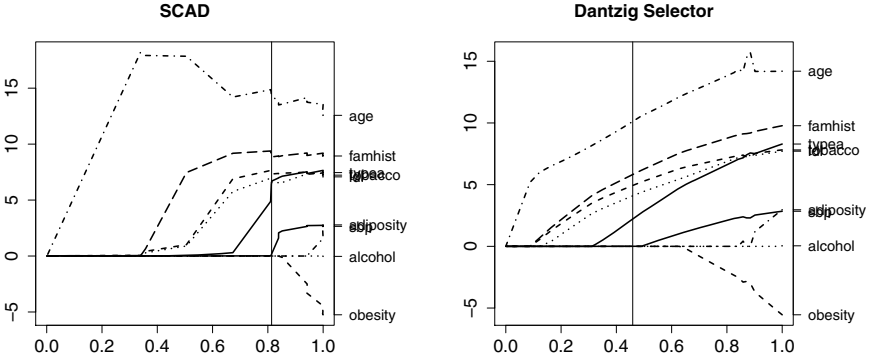


FIGURE 6.12: SCAD and Dantzig coefficient paths for heart disease data (package *lqa*; vertical line shows estimate selected by 10-fold cross-validation).

## 6.3 Boosting Methods

Boosting methods were originally developed in the machine learning community as a means to improve classification (e.g., Shapire, 1990). They have been proposed as ensemble methods, which rely on generating multiple predictions and averaging across the individual predictions. Later it was shown that boosting can be seen as the fitting of an additive structure by minimizing specific loss functions (see Friedman, 2001; Friedman et al., 2000). Bühlmann and Yu (2003) and Bühlmann (2006) proposed and investigated boosted estimators in the context of a linear regression with the focus on $L_2$ loss. In regressions, boosting may be seen as a regularization technique that also allows one to select predictors. For more background on boosting see also Section 15.5.3.

### 6.3.1 Boosting for Linear Models

Before considering the boosting of generalized linear models, we consider briefly the familiar case of normal regression models. Let the underlying regression structure be given by $\mathrm{E}(y|\boldsymbol{x}) = \eta(\boldsymbol{x})$ and data be given by $(y_i, \boldsymbol{x}_i)$, $i = 1, \ldots, n$.

Boosting is based on fitting a structured function that is supposed to approximate $\eta(\boldsymbol{x})$. The fitting of a structured function (a learner in machine learning terminology) is considered as a base procedure. Ensemble methods are based on averaging across several such procedures. Let $\hat{g}(\boldsymbol{x}, \{u_i, \boldsymbol{x}_i\})$ denote the base procedure at value $\boldsymbol{x}$ based on input data $\{u_i, \boldsymbol{x}_i\}$, which are not necessarily the original data $\{y_i, \boldsymbol{x}_i\}$. When fitting linear models, the base procedure uses a linear function $g(\boldsymbol{x}, \{u_i, \boldsymbol{x}_i\}) = \tilde{\boldsymbol{x}}^T\boldsymbol{\gamma}$, where $\tilde{\boldsymbol{x}}$ is usually a subvector of $\boldsymbol{x}$. In one step of the procedure one does not aim at estimating the whole vector $\boldsymbol{\beta}$ (from predictor $\eta = \boldsymbol{x}^T\boldsymbol{\beta}$) but at improving the estimate of the parameters $\boldsymbol{\gamma}$ that correspond to a subset of $\boldsymbol{\beta}$. A basic boosting algorithm then is given by:

*Step 1 (Initialization)*

Given data $\{y_i, \boldsymbol{x}_i\}$, fit the base procedure to yield the function estimate $\eta^{(0)}(.) = \hat{g}(., \{y_i, \boldsymbol{x}_i\})$.

*Step 2 (Iteration)*

For $l = 0, 1, 2, \ldots$, compute the residuals $u_i = y_i - \hat{\eta}^{(l)}(\boldsymbol{x}_i)$ and fit the base procedure to the current data $\{u_i, \boldsymbol{x}_i\}$. The fit $\hat{g}(., \{u_i, \boldsymbol{x}_i\})$ is an estimate based on the

original predictor variables and the current residuals. The improved fit is obtained by the update

$$\hat{\eta}^{(l+1)}(.) = \hat{\eta}^{(l)}(.) + \hat{g}(., \{u_i, \boldsymbol{x}_i\}).$$

The iteration is stopped by applying a stopping criterion, for example, $AIC$ or a cross-validation measure. Boosting in this form iteratively improves the fit by adding the fit of a base learner, which reduces the discrepancy between the current residual and the fit. It may be seen as forward stepwise additive modeling.

Bühlmann and Yu (2003) thoroughly investigated $L_2$ boosting, which utilizes least-squares estimates as a fitting procedure. Thereby, in one step one minimizes $\Sigma_i(u_i - \hat{g}(\boldsymbol{x}_i, \{u_i, \boldsymbol{x}_i\})^2$. $L_2$ boosting may also be derived as a gradient descent algorithm (see Section 15.5.3) and, as mentioned by Bühlmann and Yu, is nothing else more than repeated least-squares fitting of residuals, which for one boosting step has already been proposed by Tukey (1977) under the name "twicing." For linear models, which assume that the conditional mean $\eta(\boldsymbol{x}) = \mathrm{E}(y|\boldsymbol{x})$ has the form $\eta(\boldsymbol{x}) = \boldsymbol{x}^T\boldsymbol{\beta}$, a simple least squares-fitting of a linear predictor $g(\boldsymbol{x}, \{u_i, \boldsymbol{x}_i\}) = \boldsymbol{x}^T\boldsymbol{\beta}$ within one boosting step would simply yield the usual least-squares estimate after one step. Therefore, to obtain a regularized estimate, alternative base procedures have to be used.

An approach that implicitly selects variables is *componentwise boosting*. Componentwise boosting means that each part of the linear predictor is refitted separately, and among the fitted parts one is selected to be used in the update. In its simplest form, componentwise boosting refits only one coefficient, which is selected by some optimality criterion. Bühlmann (2006) proposed a componentwise linear least-squares algorithm for linear models by using the base learner $g(\boldsymbol{x}, \{u_i, x_i\}) = \gamma_{\hat{s}} x_{\hat{s}}$, where $\hat{\gamma}_j$ is the usual least-squares estimate resulting from using only the $j$th variable, $\hat{\gamma}_j = \Sigma_i u_i x_{ij} / \Sigma_i x_{ij}^2$ (centered predictors), and

$$\hat{s} = \arg \min_{1 \le j \le p} \sum_{i=1}^{n} (u_i - \hat{\gamma}_j x_{ij})^2$$

determines which variable is selected. Thus, the base procedure in componentwise linear least-squares boosting performs a linear least-squares regression against the one selected variable that reduces the residual sum of squares the most. The actual refit typically uses $\hat{g}(\boldsymbol{x}, \{u_i, x_i\}) = \nu\gamma_{\hat{s}} x_{\hat{s}}$, where the parameter $\nu$ is a fixed shrinkage parameter, in order to obtain a weak learner (see next section). The corresponding refit of parameters is given by $\hat{\beta}_j^{(l+1)} = \hat{\beta}_j^{(l)}$, $j \neq \hat{s}$, $\hat{\beta}_{\hat{s}}^{(l+1)} = \hat{\beta}_{\hat{s}}^{(l)} + \nu\gamma_{\hat{s}}$. Since in high-dimensional settings usually not all of the predictors are selected before the stopping criterion is reached, the procedure selects variables automatically. Bühlmann (2006) showed that the procedure is consistent for underlying regression functions, which are sparse in terms of the $L_1$-norm.

Boosting procedures are based on "weak" learners, a concept that has been derived in the machine learning community (Freund and Schapire, 1997). In classification, a weak learner may be considered as an estimator that is slightly better than guessing. In regression, a weak learner refers to small step sizes within the algorithm. Therefore, the update step in linear least-squares boosting uses

$$\hat{\eta}^{(r+1)}(.) = \eta^{(r)}(.) + \nu\hat{g}(., \{u_i, x_i\}),$$

where $\nu$ is a shrinkage parameter, for example, $\nu = 0.1$. Small step sizes (small $\nu$) make the boosting algorithm slow and require a larger number of iterations, but improve the performance. Small values of $\nu$ have been shown to avoid early overfitting of the procedure.

### 6.3.2 Boosting for Generalized Linear Models

In generalized linear models least-squares estimates are not the best choice because they do not relate adequately to the underlying error structure. A better choice is likelihood-based boosting, which more generally aims at maximizing the log-likelihood rather than minimizing the squared residuals. For fixed link functions, likelihood-based approaches iteratively estimate the linear predictor, $\eta_i = \boldsymbol{x}_i^T \boldsymbol{\beta}$, which is linked to the mean $\mu_i = \mathrm{E}(y_i|\boldsymbol{x}_i)$ by $\mu_i = h(\eta_i)$.

One difference between the $L_2$ boost and a generalized linear model boosting is that in the iteration step one cannot fit a GLM to the residuals because, for example, with binary data, residuals are not from $\{0, 1\}$. The role of the residuals is taken by the offset. The basic likelihood-based boosting algorithm (GenBoost), which is also used in Chapter 15, has the following form:

---

**Likelihood Boosting (GenBoost)**

*Step 1 (Initialization)*

For given data $(y_i, \boldsymbol{x}_i), i = 1, \ldots, n$, fit the intercept model $\mu^{(0)}(\boldsymbol{x}) = h(\beta_0)$ by maximizing the likelihood, yielding $\eta^{(0)} = \hat{\beta}_0, \hat{\mu}^{(0)} = h(\hat{\beta}_0), \hat{\boldsymbol{\beta}}^{(0)} = (\hat{\beta}_0, 0, \ldots, 0)^T$.

*Step 2 (Iteration)*

For $l = 0, 1, 2, \ldots$, fit the model

$$\mu_i = h(\hat{\eta}^{(l)}(\boldsymbol{x}_i) + \eta(\boldsymbol{x}_i, \boldsymbol{\gamma}))$$

to data $(y_i, \boldsymbol{x}_i), i = 1, \ldots, n$, where $\hat{\eta}^{(l)}(\boldsymbol{x}_i)$ is treated as an offset and the predictor is estimated by fitting the parametrically structured term $\eta(\boldsymbol{x}_i, \boldsymbol{\gamma})$, obtaining $\hat{\boldsymbol{\gamma}}$. The improved fit is obtained by

$$\hat{\eta}^{(l+1)}(\boldsymbol{x}_i) = \hat{\eta}^{(l)}(\boldsymbol{x}_i) + \hat{\eta}(\boldsymbol{x}_i, \hat{\boldsymbol{\gamma}}), \quad \hat{\mu}_i^{(l+1)} = h(\hat{\eta}^{(l+1)}(\boldsymbol{x}_i)).$$

The improved parameter $\hat{\boldsymbol{\beta}}^{(l+1)}$ is obtained by adding $\hat{\boldsymbol{\gamma}}$ to the components of $\hat{\boldsymbol{\beta}}^{(l)}$.

---

One candidate for fitting is Fisher scoring, which is familiar from generalized linear model fitting. One first has to compute the pseudo-responses and weights:

$$\tilde{\eta}_i^{(l)} = \frac{y_i - \hat{\mu}_i^{(l)}}{\partial h(\hat{\eta}_i^{(l)})/\partial \eta}, \quad w_i^{(l)} = \frac{(\partial h(\hat{\eta}_i^{(l)})/\partial \eta)^2}{\sigma_i^2},$$

and then compute the weighted regression with weights $w_i^{(l)}$ and dependent variables $\tilde{\eta}_i^{(l)}$ to obtain $\hat{\gamma}$. It is noteworthy that the pseudo-responses, in contrast to their usual definition, do not include a linear term, because the previous fit is contained in the offset. For the logit model one has $\partial h(\hat{\eta}_i)/\partial \eta = h(\eta_i)/(1 - h(\eta_i))$, and therefore the pseudo-responses and weights simplify to

$$\tilde{\eta}_i = \frac{y_i - \hat{\mu}_i^{(l)}}{\hat{\pi}^{(l)}(1 - \hat{\pi}^{(l)})}, \quad w_i = \hat{\pi}^{(l)}(1 - \hat{\pi}^{(l)}).$$

More concretely, let us consider the linear predictor $\eta(\boldsymbol{x}_i, \boldsymbol{\gamma}) = \tilde{\boldsymbol{x}}_i^T \boldsymbol{\gamma}$, where $\tilde{\boldsymbol{x}}_i$ is a specified subvector of $\boldsymbol{x}_i$. For example, when using $\eta(\boldsymbol{x}_i, \boldsymbol{\gamma}) = x_{ij}\gamma_j$, $\tilde{\boldsymbol{x}}_i$ contains only the $j$th

covariate as a candidate for updating. One computes within the iteration steps one-step Fisher scoring estimates:

$$\hat{\boldsymbol{\gamma}} = (\tilde{\boldsymbol{X}}^T \boldsymbol{W}^{(l)} \tilde{\boldsymbol{X}})^{-1} \tilde{\boldsymbol{X}}^T \boldsymbol{W}^{(l)} \tilde{\boldsymbol{\eta}}^{(l)},$$

where $\tilde{\boldsymbol{X}}$ is the design matrix built from the vectors $\tilde{\boldsymbol{x}}_i$, $\boldsymbol{W}^{(l)}$ is a diagonal matrix that contains the weights, $w_i^{(l)}$, and $\tilde{\boldsymbol{\eta}}^{(l)}$ contains the pseudo-responses $\tilde{\eta}_i^{(l)}$ (for Fisher scoring see Chapter 3, Section 3.9). Since the linear predictor $\eta(\boldsymbol{x}_i, \boldsymbol{\gamma}) = \tilde{\boldsymbol{x}}_i^T \boldsymbol{\gamma}$ is fitted, refitting refers only to the components contained in $\tilde{\boldsymbol{x}}_i$, that is, $\hat{\beta}_j^{(l+1)} = \hat{\beta}_j^{(l)} + \hat{\gamma}_j$ if $x_{ij}$ is contained in $\tilde{\boldsymbol{x}}_i$ (with $\hat{\gamma}_j$ denoting the corresponding estimate) and $\hat{\beta}_j^{(l+1)} = \hat{\beta}_j^{(l)}$ if $x_{ij}$ is not contained in $\tilde{\boldsymbol{x}}_i$.

To obtain a weak learner $\hat{\boldsymbol{\gamma}}$ can be replaced by $\nu\hat{\boldsymbol{\gamma}}$ with $\nu$ denoting a shrinkage parameter. One-step Fisher scoring, starting with the zero vector, can also be given in the form $\hat{\boldsymbol{\gamma}} = (\boldsymbol{F}^{(l)})^{-1} \boldsymbol{s}^{(l)}$, where $\boldsymbol{F}^{(l)} = \tilde{\boldsymbol{X}}^T \boldsymbol{W}^{(l)} \tilde{\boldsymbol{X}}$ is the Fisher matrix and $\boldsymbol{s}^{(l)} = \tilde{\boldsymbol{X}}^T \boldsymbol{W}^{(l)} \tilde{\boldsymbol{\eta}}^{(l)}$. Weak learners may also be obtained in the spirit of ridge estimation by using a penalized Fisher matrix $\boldsymbol{F}^{(l)} + \lambda \boldsymbol{I}$, where $\lambda$ is chosen large.

In componentwise boosting within the fitting step, a selection step is included that determines which of the parameters is refitted. In the simplest case one fits all one-covariate models $\eta(\boldsymbol{x}_i, \boldsymbol{\gamma}) = \gamma_j x_{ij}$, $j = 0, \ldots, p$, as candidates, obtaining $\hat{\gamma}_j$, and then selects the variable that has the strongest impact on the improvement of the fit. A criterion is the improvement in deviance:

$$Dev(\hat{\eta}^{(l)}) - Dev(\hat{\eta}_{new(j)}),$$

where $\tilde{\eta}_{new(j)}$ is based on the parameter vector in which only the $j$th component is updated to $\hat{\beta}_j^{(l)} + \hat{\gamma}_j$. When the $s$th variable is selected, the new parameter vector is $\hat{\boldsymbol{\beta}}^{(l+1)} = (\hat{\beta}_0^{(l)}, \ldots, \hat{\beta}_s^{(l)} + \hat{\gamma}_s, \ldots)^T$. Selection of the parameter that is actually updated within one step can also be based on information criteria like AIC or BIC.

In the case of the logit model, likelihood-based boosting is equivalent to the *LogitBoost* algorithm for two classes that were proposed by Friedman et al. (2000). The advantage of GenBoost is that it applies to all kinds of link functions and exponential family responses. The first extension to the exponential family setting was given by Ridgeway (1999). He gave two similar, though slightly different algorithms for boosting exponential family models (for details of Fisher scoring-based algorithms, see also Tutz and Binder, 2007).

Example 6.5: Heart Disease

Figure 6.13 shows the coefficient buildups for likelihood-based boosting with 500 iterations. The left plot was computed with the package *GAMBoost*, and the right plot uses the quadratic approximation used in the package *mboost*. The resulting paths are quite similar; however, *mboost* is much faster. □

## Blockwise Boosting

The strategy to update just one variable is rather limited, and it is especially inadequate if categorical variables are in the predictor. A categorical predictor that takes $k$ categories will be represented by $k - 1$ dummy variables in the linear predictor. If one wants to avoid having the resulting selection depend on the coding scheme, the parameters for all the dummy variables representing one variable should be refitted simultaneously. Then the structured term to be fitted is $\eta(\boldsymbol{x}_i, \boldsymbol{\gamma}) = \boldsymbol{x}_{ir}^T \boldsymbol{\gamma}_r$, where $\boldsymbol{x}_{ir}$ is a vector of dummy variables corresponding to a categorical variable.

In general, to obtain a selection of relevant terms, the base procedures that are used typically contain only a small number of variables. In the extreme case only one coefficient is refitted; in

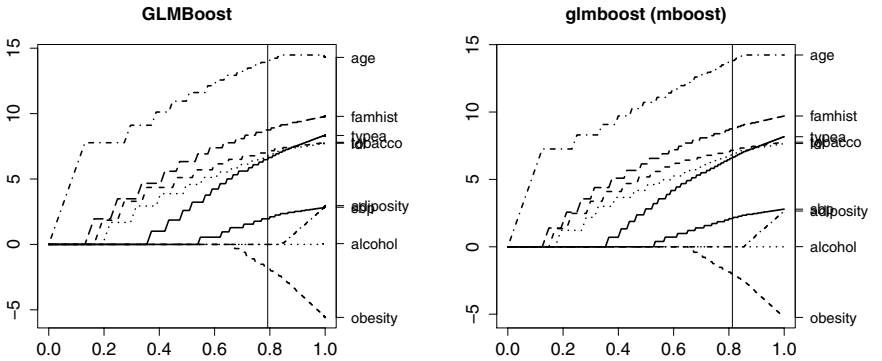


FIGURE 6.13: GlmBoost coefficient paths for heart disease data (package *mboost* (right), vertical line shows estimate selected by 10-fold cross-validation).

other cases, it can be a group of coefficients. A list of parametrically structured terms that may be fitted is

- $\eta(\boldsymbol{x}_i, \boldsymbol{\gamma}) = x_{ir}\gamma_r$, which specifies the linear effect of the $r$th covariate;
- $\eta(\boldsymbol{x}_i, \boldsymbol{\gamma}) = \gamma_0 + x_{ir}\gamma_r$, which specifies the intercept and the linear effect of the $r$th covariate;
- $\eta(\boldsymbol{x}_i, \boldsymbol{\gamma}) = \boldsymbol{x}_{ir}^T\boldsymbol{\gamma}_r$, where $\boldsymbol{x}_{ir}$ is a vector of dummy variables corresponding to a categorical variable;
- $\eta(\boldsymbol{x}_i, \boldsymbol{\gamma}) = x_{ir}x_{is}\gamma_{rs}$, representing an interaction between the $r$th and the $s$th covariates;
- $\eta(\boldsymbol{x}_i, \boldsymbol{\gamma}) = \boldsymbol{x}_{ir}\boldsymbol{x}_r^T\boldsymbol{\gamma}_{rs}$, representing an interaction between the $r$th variable and the $s$th categorical variable given by a vector of dummy variables.

These parametrically structured terms define the learner that is used. In general, the parameter $\boldsymbol{\gamma}$ is a vector that refers to a group or block of variables that define the design matrix $\tilde{\boldsymbol{X}}$. The update by adding $\hat{\boldsymbol{\gamma}}$ is done blockwise. When one fits all one-covariate models with intercepts $\eta(\boldsymbol{x}_i, \boldsymbol{\gamma}) = \gamma_0 + \gamma_r x_{ir}$, $r = 1, \ldots, p$, as candidates, after selection of the best update, $s$, the update is given by $\hat{\boldsymbol{\beta}}^{(l+1)} = (\hat{\beta}_0^{(l)} + \hat{\gamma}_0, \ldots, \hat{\beta}_j^{(l)} + \hat{\gamma}_s, \ldots)^T$.

In addition, in some cases it is not sensible to let the procedure select among all the variables. For example, in treatment studies, the treatment should be considered as a mandatory variable that is always included in the predictor. Therefore, one should distinguish between *mandatory* and *optional* predictors. The more general concept of blockwise boosting allows one to distinguish between these types of variables and to refit groups of variables.

Blockwise (or partial) boosting means that in the $l$th iteration selected components of the parameter vector are re-estimated. The selection is determined by a specific structuring of the parameters (variables). Let the parameter indices $V = \{1, \ldots, p\}$ be partitioned into disjoint sets by $V = V_c \cup V_{o1} \cup \ldots \cup V_{oq}$, where $V_c$ stands for the (mandatory) parameters (variables) that have to be included in the analysis, and $V_{o1}, \ldots, V_{oq}$ represent blocks of parameters that are optional. A block $V_{or}$ may refer to all the parameters that refer to a multicategorical variable, such that not only parameters but variables are evaluated. Candidates in the refitting process are all combinations $V_c \cup V_{or}, r = 1, \ldots, q$, representing combinations of necessary and optional variables. Componentwise boosting that refers to single coefficients is the special case where $V_c = \emptyset, V_{oj} = \{j\}$.

The base procedure that is used in refitting steps refers to the fitting of a shrinked estimator (weak learner) to the candidate sets and the selection of the "least" candidate. As usual, fitting for generalized linear models means maximizing the log-likelihood. Since a shrinked version has better performance, a shrinkage estimator, for example, the ridge estimator with large $\lambda$, is used. Moreover, since the boosting procedure itself means an iterative refitting of residuals, within one refitting step of the boosting algorithm we use one-step Fisher scoring rather than a complete fit.

More technically, let $V_m = V_c \cup V_{om}$ denote the indices of parameters to be considered for refitting and $\boldsymbol{X}_{V_m}$ denote the corresponding submatrix of the full design matrix $(\boldsymbol{x}_{.1}, \ldots, \boldsymbol{x}_{.p})$. Then the partial boosting algorithm is given by:

---

**Blockwise/Partial Likelihood Boosting (GenPartBoostR)**

*Step 1: Initialization*

Fit model $\mu_i = h(\beta_0)$ by iterative Fisher scoring to obtain $\hat{\boldsymbol{\beta}}^{(0)} = (\hat{\beta}_0, 0, \ldots, 0)^T$, $\hat{\boldsymbol{\eta}}_{(0)} = \boldsymbol{X}\hat{\boldsymbol{\beta}}^{(0)}$.

*Step 2: Iteration*

For $l = 1, 2, \ldots$

(a) Estimation: Estimation for candidate sets $V_m$, $m = 1, \ldots, q$, corresponds to fitting of the model

$$\boldsymbol{\mu} = h(\hat{\boldsymbol{\eta}}^{(l-1)} + \boldsymbol{X}_{V_m}\boldsymbol{\gamma}_{V_m}),$$

where $\hat{\boldsymbol{\eta}}^{(l-1)} = \boldsymbol{X}\hat{\boldsymbol{\beta}}^{(l-1)}$ is treated as an offset (fixed constant). Fitting is performed by one step of Fisher scoring by use of a weak learner for $\boldsymbol{\gamma}_{V_m}$.

(b) Selection: For candidate sets $V_m$, $m = 1, \ldots, q$, the set $V_{m_0}$ is selected that improves the fit maximally.

(c) Update: One sets

$$\hat{\boldsymbol{\gamma}}^{(l)} = \begin{cases} \hat{\boldsymbol{\gamma}}_{V_m, j} & j \in V_{m_0} \\ 0 & j \notin V_{m_0}, \end{cases}$$

$\hat{\boldsymbol{\beta}}^{(l)} = \hat{\boldsymbol{\beta}}^{(l-1)} + \hat{\boldsymbol{\gamma}}^{(m)}$, $\hat{\boldsymbol{\eta}}^{(l)} = \boldsymbol{X}\hat{\boldsymbol{\beta}}^{(l)}$, $\hat{\boldsymbol{\mu}}^{(l)} = h(\boldsymbol{X}\hat{\boldsymbol{\beta}}^{(l)})$, where $h$ is applied componentwise.

---

Example 6.6: Abortion of Treatment

We illustrate the application of simple ridge boosting and partial boosting with real data from 344 admissions at a psychiatric hospital. The (binary) response variable to be investigated is whether treatment is aborted by the patient against physicians' advice (about 55% for this group of patients). From a total of 101 variables available, 8 variables likely to be relevant were identified: age, number of previous admissions ("noupto"), cumulative length of stay ("losupto") and the 0/1-variables indicating a previous ambulatory treatment ("prevamb"), no second diagnosis ("nosec"), second diagnosis "personality disorder" ("persdis"), somatic problems ("somprobl"), homelessness ("homeless"), and joblessness ("jobless"). Based on subject matter considerations, the two variables that relate to secondary diagnosis ("nosec" and "persdis") are mandatory members of the response set and no penalty is applied to their estimates. This illustrates

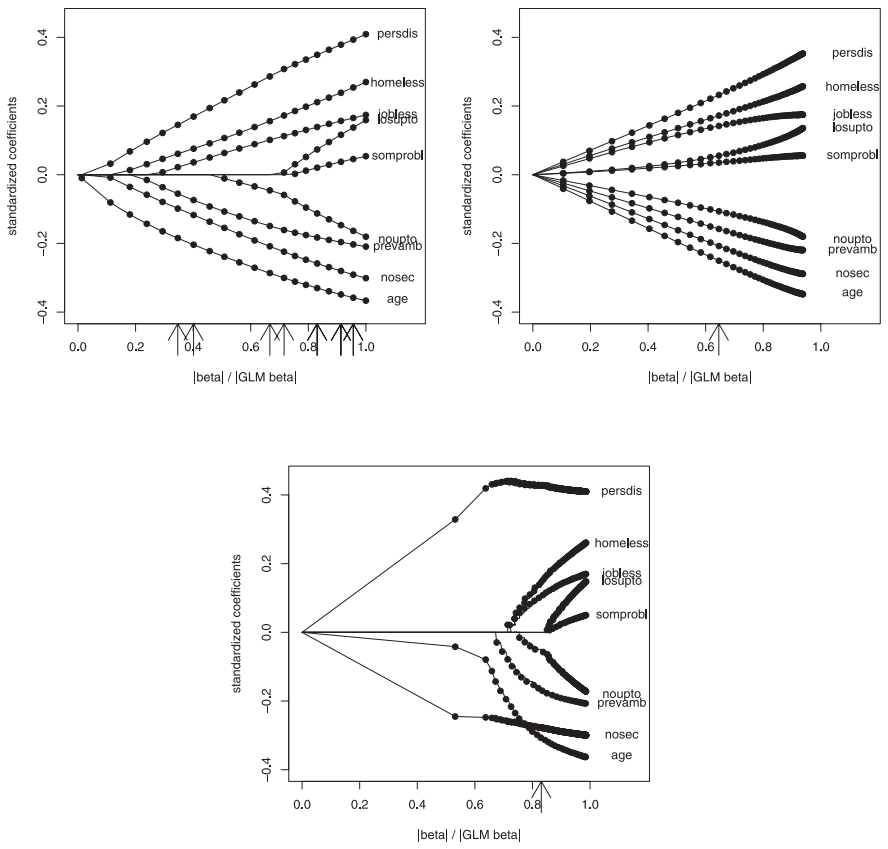


FIGURE 6.14: Coefficient buildups for abortion of treatment data when using lasso (upper left), ridge (upper right), and boosted ridge (lower panel) with mandatory variables.

the effect of augmenting an unpenalized model with a few mandatory variables with optional predictors. Figure 6.14 shows the coefficient buildup in the course of the boosting steps for partial boosting contrasted with simple ridge boosting (lower panel) and the lasso (upper left panel). The arrows indicate the number of steps chosen by $AIC$ (for partial boosting and simple ridge boosting) and 10-fold cross-validation (for the lasso repeated 10 times). It can be seen that the mandatory components introduce a very different structure in coefficient buildups. One interesting feature is the slow decrease of the estimate for "persdis" beginning with boosting step 8. This indicates some overshooting of the initial estimate that is corrected when additional predictors are included. To identify relevant variables we used all 101 predictors and divided the data into a training set of size 270 and a test set of size 74. The lasso (with cross-validation) returned 13 predictors with a prediction error of 0.392. Partial boosting (using six mandatory response set elements relating to the secondary diagnosis) with penalty varying from 500 to 10000 returned 15 to 19 predictors and a prediction error between 0.378 and 0.392. □

Variable selection by regularization is a very active research area. Modifications and improvements are proposed and properties of existing estimates are investigated in many journals. Therefore, consideration of advantages and disadvantages tends to be preliminary. Nevertheless, in Table 6.1 some properties of currently available procedures are listed.

TABLE 6.1: Properties of regularized estimators.

| | |
|---|---|
| Ridge | Estimates exist. Prediction performance better than for ML estimates.<br>Explicit solution for linear model. |
| | No selection of predictors. |
| Lasso | Selects predictors, sparse representation. Oracle properties hold for adaptive lasso. |
| | Tends to select one predictor from a group of highly correlated predictors.<br>Not necessarily consistent (but adaptive lasso is). |
| Elastic net | Selects predictors. Shows the grouping property. |
| | Two tuning parameters have to be selected. |
| Oscar | Selects predictors, exact grouping property. Clustering of predictors available. |
| | Two tuning parameters have to be selected. |
| Correlation-based | Grouping property. |
| | Does not select predictors (boosted version does) |
| SCAD | Nearly unbiased for large unknown coefficients.<br>Automatically sets small estimated coefficients to zero (sparsity).<br>Continuous in the data to avoid instability in model prediction (continuity).<br>Oracle property. |
| Dantzig selector | Variable selection included. Can be computed efficiently. |
| Componentwise boosting | Selects variables. |
| | Inference hard to obtain. |

## 6.4 Simultaneous Selection of Link Function and Predictors

When predictors are selected one typically assumes that the link function is known. However, if the assumed link function is wrong, the performance of the selection procedures can be strongly affected. For illustration, let us consider a small simulation study. Let the generating model be a Poisson model with the true response function having sigmoidal form $h_T(\eta) = 10/(1 + \exp(-5 \cdot \eta))$. Let the parameter vector of length $p = 20$ be given by $\boldsymbol{\beta}^T = (0.2, 0.4, -0.4, 0.8, 0, \ldots, 0)$ and covariates be drawn from a normal distribution $\boldsymbol{x} \sim N(\mathbf{0}_p, \boldsymbol{\Sigma})$ with $\boldsymbol{\Sigma} = \{\sigma_{ij}\}_{i,\, j \in \{1, \ldots, p\}}$, where $\sigma_{ij} = 0.5$, $i \neq j$, $\sigma_{ii} = 1$. We generate $N = 50$ datasets with $n = 200$ observations and fit the model by using the usual maximum likelihood (ML) procedure based on the canonical log-link (without variable selection). In addition, we apply three alternative fitting methods that include variable selection: a non-parametric flexible link procedure considered in the following, the lasso for generalized linear models, and a componentwise boosting procedure. While the flexible link procedure selects a link function, ML estimates as well as lasso and boosting use the canonical link. It is seen from Figure 6.15 that the best results are obtained if the link function is estimated non-parametrically. In particular, the parameters of the predictors that are not influential are estimated more stable and closer to zero. The dominance of the flexible procedure is also seen in Figure 6.16, which shows the mean squared error for the estimation of the parameter vector and the predictive deviance on an independently drawn test dataset with $n = 1000$.

The flexible procedure shown in Figure 6.15 is an extension of the non-parametric estimation procedure considered in Section 5.2. One fits the model $\mu_i = h_0(h(\eta_i))$, where $h_0(.)$ is a fixed transformation function, for example, the canonical link, and the inner function

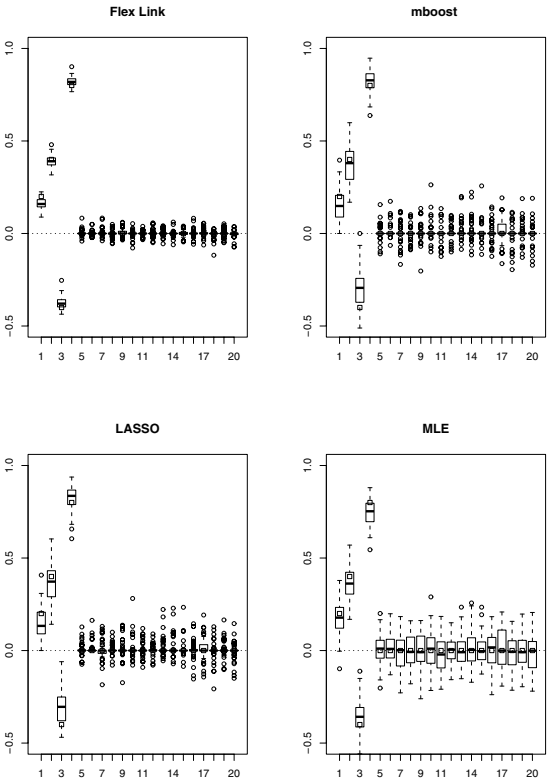


FIGURE 6.15: Resulting estimates of coefficient vector in simulation study for flexible link, boosting, lasso, and ML.

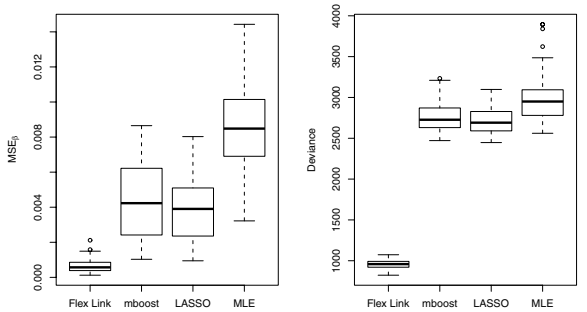


FIGURE 6.16: Mean squared error for parameter vector and predictive deviance for simulation setting.

$h(.)$ is considered as unknown and is estimated by assuming an expansion in basis functions $h(\eta_i) = \sum_{j=1}^{m} \alpha_j \phi_j(\eta_i) = \boldsymbol{\alpha}^T \boldsymbol{\Phi}_i$.

Estimates are obtained by iteratively estimating the regression coefficients $\boldsymbol{\beta}$ and the parameters of the link function $\boldsymbol{\alpha}$. In matrix notation, let $\hat{\boldsymbol{\beta}}^{(l)}$ and $\hat{\boldsymbol{\eta}}^{(l)} = \boldsymbol{X}\hat{\boldsymbol{\beta}}^{(l)}$ denote the parameter estimate and the fitted predictor in the $l$th step. Moreover, $\boldsymbol{\Phi}_i^{(l)} = (\boldsymbol{\Phi}_1^{(l)}, \ldots, \boldsymbol{\Phi}_n^{(l)})^T$ with $\boldsymbol{\Phi}^{(l)} = (\phi_1(\hat{\eta}_i^{(l)}), \ldots, \phi_m(\hat{\eta}_i^{(l)}))^T$ is the current design matrix for the basis functions. Within a boosting-type procedure two steps are iterated:

*Boosting for Fixed Predictor*. For a fixed predictor $\hat{\boldsymbol{\eta}}^{(l-1)} = \boldsymbol{X}\hat{\boldsymbol{\beta}}^{(l-1)}$, the estimation of the response function corresponds to fitting the model $\boldsymbol{\mu} = h_0((\boldsymbol{\Phi}^{(l-1)})^T\hat{\boldsymbol{\alpha}}^{(l-1)} + (\boldsymbol{\Phi}^{(l-1)})^T\hat{\boldsymbol{a}}^{(l)})$, where $\Phi^{(l-1)})^T\hat{\boldsymbol{\alpha}}^{(l-1)}$ is a fixed offset that represents the previously fitted value. One step of penalized Fisher scoring has the form

$$\begin{aligned}\hat{\boldsymbol{a}}^{(l)} &= \nu_h((\boldsymbol{\Phi}^{(l-1)})^T\hat{\boldsymbol{D}}^{(l-1)}(\hat{\boldsymbol{\Sigma}}^{(l-1)})^{-1}\hat{\boldsymbol{D}}^{(l-1)}\boldsymbol{\Phi}^{(l-1)} + \lambda_h\mathbf{P}_h)^{-1}\cdot \\ &\quad \cdot \left(\boldsymbol{\Phi}^{(l-1)}\right)^T \hat{\boldsymbol{D}}^{(l-1)}(\hat{\boldsymbol{\Sigma}}^{(l-1)})^{-1}(\boldsymbol{y} - \hat{\boldsymbol{\mu}}^{(l-1)}),\end{aligned}$$

where $\hat{\mathbf{D}}^{(l-1)} = \mathrm{diag}(\partial h_0(\hat{h}^{(l-1)}(\hat{\eta}_i^{(l-1)}))/\partial h^{(l-1)}(\eta))$ is the estimate of the derivative matrix evaluated at the estimate of the previous step and $\hat{\boldsymbol{\Sigma}}^{(l-1)}$ is the diagonal matrix of variances evaluated at $h_0(\hat{h}^{(l-1)}(\eta_i^{(l-1)}))$. $\mathbf{P}_h$ is the penalty matrix that penalizes the second derivation of the estimated (approximated) response function and the shrinkage parameter is fixed by $\nu_h = 0.1$.

*Componentwise Boosting for Fixed Response Function*. Let $h(.)$ be fixed and the design matrix have the form $\mathbf{X} = (\boldsymbol{x}_1|...|\boldsymbol{x}_p)$ with corresponding response vector $\boldsymbol{y} = (y_1, ..., y_n)^T$. Componentwise boosting means to update one parameter within one boosting step. Therefore, one fits the model $\mu = h_0(h(\boldsymbol{X}\hat{\boldsymbol{\beta}}^{(l-1)} + \boldsymbol{x}_j b_j))$, where $\boldsymbol{X}\hat{\boldsymbol{\beta}}^{(l-1)}$ is a fixed offset and only the variable $\boldsymbol{x}_j$ is included in the model. Then penalized Fisher scoring for the parameter $b_j$ has the form

$$\hat{b}_j^{(l)} = \nu_p(\boldsymbol{x}_j^T\hat{\boldsymbol{D}}_\eta^{(l-1)}(\hat{\boldsymbol{\Sigma}}^{(l-1)})^{-1}\hat{\boldsymbol{D}}_\eta^{(l-1)}\boldsymbol{x}_j)^{-1}\boldsymbol{x}_j^T\hat{\boldsymbol{D}}_\eta^{(l-1)}(\hat{\boldsymbol{\Sigma}}^{(l-1)})^{-1}(\boldsymbol{y} - \hat{\boldsymbol{\mu}}^{(l-1)}),$$

where $\nu_p = 0.1$, $\hat{\boldsymbol{D}}_\eta^{(l-1)} = \mathrm{diag}(\partial h_0(\hat{h}^{(l-1)}(\hat{\eta}_i^{(l-1)}))/\partial\eta)$ is the matrix of derivatives evaluated at the values of the previous iteration and $\hat{\boldsymbol{\Sigma}}^{(l-1)}$ is the variance from the previous step.

In each step of the boosting algorithm it is decided if the regression coefficients or the coefficients of the basis functions are updated; for details see Tutz and Petry (2011). The advantage of boosting techniques is that variable selection is included. By updating only one of the coefficients and stopping the updating procedure appropriately, one obtains the relevant predictors.

### Example 6.7: Demand for Medical Care

In Example 7.6, count data with the number of physician office visits as the response variable were considered. Here we consider a Poisson model with flexible link and the same predictors as in Example 7.6. Figure 6.17 shows the estimated response functions plotted against the linear predictor. The canonical log-link is a strictly increasing function while the non-parametrically estimated response function becomes flat for large values of the predictor. The canonical link seems not to be appropriate. This is supported by an improved prediction error in subsamples when the flexible link function is used (see Tutz and Petry, 2011). □

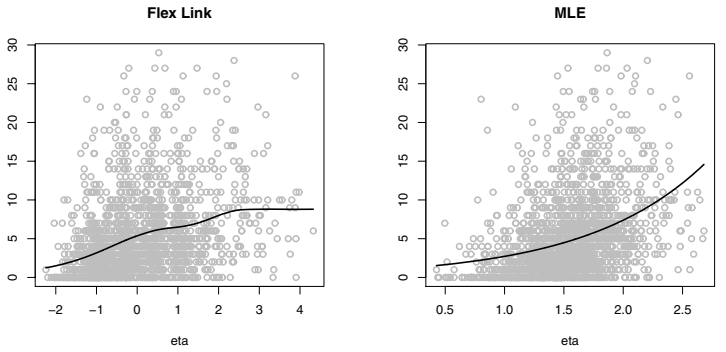


FIGURE 6.17: Response functions for medical care data against linear predictor: flexible link function (left) and canonical log link (right).

## 6.5 Categorical Predictors

The selection of categorical variables has already been briefly discussed for lasso-type penalties and boosting approaches. In the following we will consider further approaches. When predictors are categorical selection should distinguish between two cases: the selection of variables and the selection of effects within variables. If one wishes to select variables one has to select groups of variables because a categorical variable is represented by several dummy variables. If, however, one enforces selection among all terms in the linear predictor, including dummies, a selection strategy like the lasso will select single dummy variables with the effect that the coding scheme that has been used on the categorical predictors determines the result. To avoid such effects one should distinguish between the two problems:

- Which categorical predictors should be included in the model?
- Which categories within one categorical predictor should be distinguished?

The latter problem is concerned with one single variable and poses the question of which categories differ from one another with respect to the dependent variable. Or, to put it in a different way, which categories should be collapsed? The answer to that question depends on the scale level of the predictor; one should distinguish between nominal and ordered categories because of their differing information content. We will first consider selection strategies for effects within variables and then the selection of variables as groups of possibly regularized effects.

### 6.5.1 Selection within Categorical Predictors

Let us first consider just one categorical predictor $A \in \{1, \ldots, k\}$, which is included in the predictor by the use of dummy variables in the form $\eta = \beta_0 + \sum_j x_{A(j)}\beta_j$. Then, when computing a penalized estimate, for example, by use of a lasso-type penalty,

$$J(\boldsymbol{\beta}) = \sum_j |\beta_j|,$$

the shrinkage effect depends on the coding scheme that is used. For simplicity, let the categorical predictor $A$ have only three categories, $A \in \{1, 2, 3\}$, which are coded by two dummy

variables $x_{A(2)}, x_{A(3)}$. If (0–1)-coding ($x_{A(j)} = 1$ if $A = j$ and $x_{A(j)} = 0$ otherwise) is used, shrinkage refers to the difference between the first and second categories and the difference between the third and first categories, since the first category is implicitly used as a reference ($\beta_{A(1)} = 0$). If effect coding is used ($x_{A(j)} = 1$ if $A = j$, $x_{A(j)} = -1$ if $A = k$, $x_{A(j)} = 0$ otherwise), shrinkage refers to the effect of factor levels with the global level across categories as the reference point because implicitly $\beta_{A(1)} + \beta_{A(2)} + \beta_{A(3)} = 0$ is assumed. Therefore, the selection of parameters, which is enforced by the lasso penalty, yields parameters that depend on the coding scheme. If a parameter, say $|\beta_j|$, is set to zero, that means that in the case of (0–1)-coding, category $j$ and the reference category cannot be distinguished. But collapsing the categories always refers to the reference category; all other possible combinations of categories are ignored. To allow collapsing of any two categories, alternative penalties, which are considered in the following, have to be used.

### Clustering of Categories for Nominal Predictor

For *nominal* predictor variables with many categories, a useful strategy is to search for clusters of categories with similar effects. The objective is to reduce the $k$ categories to a smaller number of categories that form clusters; the effect of categories within one cluster is supposed to be the same, but responses will differ across clusters. With (0–1)-coding and reference category 1, that is, $\beta_1 = 0$, a fusion-type penalty that enforces clustering is

$$J(\boldsymbol{\beta}) = \sum_{i>j} w_{ij}|\beta_i - \beta_j| = w_{21}|\beta_2| + \cdots + w_{k1}|\beta_k| + w_{32}|\beta_3 - \beta_2| + \ldots, \tag{6.11}$$

where $w_{ij}$ is an additional weight that may depend on the sample sizes within the categories. The penalty enforces the selection among effects $\theta_{ij} = \beta_i - \beta_j$, $i = 1, \ldots, k-1, i > j$. Since the ordering of the dummy varables $x_{A(1)}, \ldots, x_{A(k)}$ is arbitrary, all differences $\beta_i - \beta_j$ are used. For large $\lambda$, the penalty $\lambda J(\boldsymbol{\beta})$ tends to form clusters of categories; for $\lambda \to \infty$, all parameter estimates become zero and the categorical predictor is excluded.

The penalty is very useful when the predictor has many categories. For small sample size as compared to the number of categories, the ML estimate becomes unstable. In contrast, regularized estimates are much more stable, and with the selection effect of the $L_1$-penalty on differences they allow one to form clusters.

### Ordered Categories

An interesting case is selection strategies for ordered predictors. Ordered categories contain more information than unordered categories, but the information has not been used in penalty (6.11). Since now the ordering of dummy coefficients is meaningful, a useful penalty for (0-1)-coding is

$$J(\boldsymbol{\beta}) = \sum_{i=2}^{k} w_i|\beta_i - \beta_{i-1}| = w_2|\beta_2| + w_3|\beta_3 - \beta_2| + \ldots, \tag{6.12}$$

with $\beta_1 = 0$. By putting the $L_1$-penalty on differences of adjacent categories, the procedure tends to fuse adjacent categories and select groups of categories that may actually be distinguished. Therefore, an ordered categorical predictor with many categories is reduced to a categorical predictor that is formed by the resulting clusters of categories. Typically the number of clusters is much smaller than the original number of categories. For $\lambda \to \infty$, all parameter estimates will become zero and the predictor is excluded since categories cannot be distinguished.

It should be noted that the penalty can also be given in the simpler form of a (weighted) lasso when split-coding of predictors is used. When using the split-coded predictors $\tilde{x}_{A(1)}, \ldots,$

$\tilde{x}_{A(k-1)}$ with $\tilde{x}_{A(i)} = 1$ if $A > i$ and $\tilde{x}_{A(i)} = 0$ otherwise, the corresponding penalty for parameters $\tilde{\beta}_1 = \beta_2$, $\tilde{\beta}_2 = \beta_3 - \beta_2$, $\tilde{\beta}_{k-1} = \beta_k - \beta_{k-1}$ used in the predictor $\eta = \beta_0 + \sum_{i=1}^{k-1} \tilde{x}_{A(i)}\tilde{\beta}_i$ is

$$\sum_{j=1}^{k-1} w_i|\tilde{\beta}_j|,$$

which is a weighted lasso-type penalty. For the transformation between (0-1)-coding and split-coding see also Section 4.4.3.

Both penalties, (6.11) for nominal predictors and (6.12) for ordinal predictors, show good clustering properties and allow one to reduce the number of categories. Fusion methodology goes back at least to Land and Friedman (1997). The penalty for ordered categories is a modification of the fused lasso penalty proposed by Tibshirani et al. (2005) (see also Section 10.4.4). The penalty for nominal predictors was considered by Bondell and Reich (2009) and Gertheiss and Tutz (2010). A further advantage of these penalties is that they have desirable asymptotic properties.

Let us consider nominal factors first. Let $\boldsymbol{\theta} = (\theta_{21}, \theta_{31}, \ldots, \theta_{k,k-1})^T$ denote the vector of pairwise differences $\theta_{ij} = \beta_i - \beta_j$. Furthermore, let $\mathcal{C} = \{(i,j) : \beta_i^* \neq \beta_j^*, i > j\}$ denote the set of indices $i > j$ corresponding to differences of (true) dummy coefficients $\beta_i^*$ that are truly non-zero, and let $\mathcal{C}_n$ denote the set corresponding to those difference that are estimated to be non-zero with sample size $n$. Let $\boldsymbol{\theta}_{\mathcal{C}}^*$ denote the true vector of pairwise differences included in $\mathcal{C}$, and $\hat{\boldsymbol{\theta}}_{\mathcal{C}}$ the corresponding estimate based on $\hat{\boldsymbol{\beta}}$. Moreover, let weights have the form

$$w_{ij} = \phi_{ij}(n)|\hat{\beta}_i^{(LS)} - \hat{\beta}_j^{(LS)}|^{-1},$$

where $\hat{\beta}_i^{(LS)}$ denotes the ordinary least-squares estimates, and for increasing $n$ one has $\phi_{ij}(n) \to q_{ij}$ $(0 < q_{ij} < \infty)$ for all $i, j$. If $\lambda = \lambda_n$ with $\lambda_n/\sqrt{n} \to 0$ and $\lambda_n \to \infty$, and all class-wise sample sizes $n_i$ satisfy $n_i/n \to c_i$, where $0 < c_i < 1$, then one obtains for the linear model $\sqrt{n}(\hat{\boldsymbol{\theta}}_{\mathcal{C}} - \boldsymbol{\theta}_{\mathcal{C}}^*) \to_d N(\mathbf{0}, \boldsymbol{\Sigma})$ (for specific matrix $\boldsymbol{\Sigma}$) and $\lim_{n\to\infty} P(\mathcal{C}_n = \mathcal{C}) = 1$. Therefore, asymptotically the right clusters are identified.

A similar property holds for ordered predictors. Now let $\mathcal{C} = \{i > 1 : \beta_i^* \neq \beta_{i-1}^*\}$ denote the set of indices corresponding to the differences of neighboring (true) dummy coefficients $\beta_i^*$ that are truly non-zero, and again let $\mathcal{C}_n$ denote the set corresponding to those differences that are estimated to be non-zero. The vector of first differences $\delta_i = \beta_i - \beta_{i-1}$, $i = 2, \ldots, k$, is now denoted as $\boldsymbol{\delta} = (\delta_2, \ldots, \delta_k)^T$. In analogy to the unordered case, let $\boldsymbol{\delta}_{\mathcal{C}}^*$ denote the true vector of (first) differences included in $\mathcal{C}$, and $\hat{\boldsymbol{\delta}}_{\mathcal{C}}$ the corresponding estimate. With weights

$$w_i = \phi_i(n)|\hat{\beta}_i^{(LS)} - \hat{\beta}_{i-1}^{(LS)}|^{-1}$$

and the same conditions for $\phi_i(n)$ and $n$ as for nominal factors, one obtains $\sqrt{n}(\hat{\boldsymbol{\delta}}_{\mathcal{C}} - \boldsymbol{\delta}_{\mathcal{C}}^*) \to_d N(\mathbf{0}, \boldsymbol{\Sigma})$ and $\lim_{n\to\infty} P(\mathcal{C}_n = \mathcal{C}) = 1$ (see Gertheiss and Tutz, 2010. ).

### 6.5.2 Selection of Variables Combined with Clustering of Categories

When several categorical predictors are available, with the $l$th variable having categories $1, \ldots, k_l$ and the corresponding parameter vector $\boldsymbol{\beta}_l^T = (\beta_{l1}, \ldots, \beta_{lk_l})$, a combination of variable selection and clustering is obtained by the penalty

$$J(\boldsymbol{\beta}) = \sum_{l=1}^{p} J_l(\boldsymbol{\beta}_l), \tag{6.13}$$

with

$$J_l(\boldsymbol{\beta}_l) = \sum_{i>j} w_{ij}^{(l)} |\beta_{li} - \beta_{lj}|, \quad \text{or} \quad J_l(\boldsymbol{\beta}_l) = \sum_i w_i^{(l)} |\beta_{li} - \beta_{l,i-1}|,$$

depending on the scale level of the predictor $x_l$. The first expression refers to nominal covariates, the second to ordinal ones. At first sight the penalty seems to enforce clustering only. However, since $\beta_{l1} = 0$ is fixed for $l = 1, \ldots, p$, a predictor is automatically excluded if all of its categories form one cluster. Due to the (additive) form of the penalty, theoretic results generalize to the case of multiple categorial inputs. For ordered categories the grouping into adjacent categories is enforced, while for unordered categories the clustering into not necessarily adjacent categories is enforced. With an appropriately chosen smoothing parameter, one automatically selects variables and the relevant information within variables.

In applications the weight function has to be specified. Bondell and Reich (2009) proposed weights determined by $w_{ij}^{(l)} = (k_l + 1)^{-1}\{(n_i^{(l)} + n_j^{(l)})/n\}^{1/2}$, where $n_i^{(l)}$ denotes the number of observations on level $i$ of predictor $x_l$.

Example 6.8: Munich Rent

All larger German cities compose so-called rent standards to obtain a decision-making instrument available to tenants, landlords, renting advisory boards, and experts. These rent standards are used in particular for the determination of the local comparative rent. For the composition of the rent standards, a representative random sample is drawn from all relevant households. The data analyzed here come from 2053 households interviewed for the Munich rent standard 2003. The response is monthly rent per square meter in Euro. The predictors are ordered as well as unordered and also include binary factors. They include urban district (nominal, labeled by numbers $1, \ldots, 25$), year of construction (ordered classes $[1910, 1919]$, $[1920, 1929], \ldots$), number of rooms (taken as an ordinal factor with levels $1, 2, \ldots, 6$), quality of residential area (ordinal, with levels "fair," "good," "excellent"), floor space (square meters, given in ordered classes $(0, 30)$, $[30, 40)$, $[40, 50)$, $\ldots$, $[140, \infty)$), hot water supply (binary, yes/no), central heating (binary, yes/no), tiled bathroom (binary, yes/no), supplementary equipment in bathroom (binary, yes/no), and well-equipped kitchen (binary, yes/no). The data can be downloaded from the data archive of the Department of Statistics at the University of Munich (http://www.stat.uni-muenchen.de/service/datenarchiv). In order to find relevant variables and identify clusters of categories that have the same effect on the predictor regularized estimates with penalty term (6.13) are computed. With weights chosen by cross-validation, the only predictor that is completely excluded from the model is the binary factor, which indicates if supplementary equipment in the bathroom is available. However, some categories of nominal and ordinal predictors are clustered, for example, houses constructed in the 1930s and 1940s, or urban districts 14, 16, 22, and 24. The original 25 districts of Munich have been reduced to merely 10 categories, which have differing rent levels (see Table 6.2). The regularization paths given in Figure 1.4 show how categories are combined. For the districts, which are treated as nominal, any combination of categories is allowed. For the year of construction ordering over decades is assumed. The regularization paths show how adjacent categories are fused to build clusters. A map of Munich with clusters (Figure 6.18) illustrates the 7 found clusters.

□

### 6.5.3 Selection of Variables

The selection of whole variables refers to the selection of groups of corresponding dummy variables. Within the framework of boosting, the selection of groups is easily obtained by the use of blockwise boosting, where blocks refer to one categorical predictor. An alternative strategy is the group lasso considered in Section 6.2.2. It tends to select the whole group of

TABLE 6.2: Estimated regression coefficients for Munich rent standard data using adaptive weights with refitting, and (cross-validation score minimizing) $s/s_{\max} = 0.61$.

| predictor | label | coefficient |
|---|---|---|
| intercept | | 12.597 |
| urban district | 14, 16, 22, 24 | −1.931 |
| | 11, 23 | −1.719 |
| | 7 | −1.622 |
| | 8, 10, 15, 17, 19, 20, 21, 25 | −1.361 |
| | 6 | −1.061 |
| | 9 | −0.960 |
| | 13 | −0.886 |
| | 2, 4, 5, 12, 18 | −0.671 |
| | 3 | −0.403 |
| year of construction | 1920s | −1.244 |
| | 1930s, 1940s | −0.953 |
| | 1950s | −0.322 |
| | 1960s | 0.073 |
| | 1970s | 0.325 |
| | 1980s | 1.121 |
| | 1990s, 2000s | 1.624 |
| number of rooms | 4, 5, 6 | −0.502 |
| | 3 | −0.180 |
| | 2 | 0.000 |
| quality of residential area | good | 0.373 |
| | excellent | 1.444 |
| floor space ($m^2$) | $[140, \infty)$ | −4.710 |
| | $[90, 100)$, $[100, 110)$, $[110, 120)$, $[120, 130)$, $[130, 140)$ | −3.688 |
| | $[60, 70)$, $[70, 80)$, $[80, 90)$ | −3.443 |
| | $[50, 60)$ | −3.177 |
| | $[40, 50)$ | −2.838 |
| | $[30, 40)$ | −1.733 |
| hot water supply | no | −2.001 |
| central heating | no | −1.319 |
| tiled bathroom | no | −0.562 |
| suppl. equipment in bathroom | yes | 0.506 |
| well equipped kitchen | yes | 1.207 |

coefficients linked to one categorical predictor. The basic group lasso uses the penalty

$$J(\boldsymbol{\beta}) = \sum_{j=1}^{G} \sqrt{df_j} \|\boldsymbol{\beta}_j\|_2,$$

where $\|\boldsymbol{\beta}_j\|_2 = (\beta_{j1}^2 + \cdots + \beta_{j,df_j}^2)^{1/2}$ and $\boldsymbol{\beta}_j$ denotes the parameter of the $j$th group from the partitioned predictor $\boldsymbol{x}_i^T = (\boldsymbol{x}_{i1}^T, \ldots, \boldsymbol{x}_{i,G}^T)$. Thus the group of coefficients collected in $\boldsymbol{\beta}_j$ is shrunken (by use of a ridge-type penalty), but there is no selection effect within the group.

For ordered categories one can incorporate smoothing across categories by using the transformation from the preceding section. With (0-1)-coding and the first category as a reference category of a predictor with $k_j$ categories, one replaces the term $\|\boldsymbol{\beta}_j\|_2$ in the penalty by

$$\|\boldsymbol{\beta}_j^T \boldsymbol{K}_j \boldsymbol{\beta}_j\|_2,$$

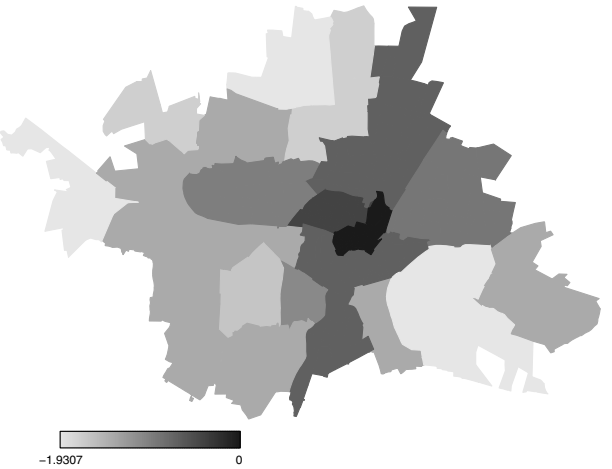


FIGURE 6.18: Map of Munich indicating clusters of urban districts; colors correspond to estimated dummy coefficients from Table 6.2.

where $\boldsymbol{K}_j = \boldsymbol{D}_j^T \boldsymbol{D}_j$ and $\boldsymbol{D}_j$ is the $((k_j - 1) \times (k_j - 1))$-matrix given in equation (4.9) without the first column and $\boldsymbol{\beta}_j^T = (\beta_{j2}, \ldots, \beta_{jk})$. As shown in Section 4.4.3, the penalty can be transformed into a penalty that uses split-coding of covariates. The transformation is helpful because then software designed for the group lasso can be used to fit the model (for more details and examples see Gertheiss et al., 2011).

## 6.6 Bayesian Approach

In Bayesian approaches to regularization, a prior distribution $p(\theta)$ is specified together with a sampling model $p(\boldsymbol{y}|\theta)$ for observations $\boldsymbol{y}$. Then the updated knowledge after the data have been seen is given by the posterior distribution

$$p(\theta|\boldsymbol{y}) = \frac{p(\boldsymbol{y}|\theta)p(\theta)}{\int p(\boldsymbol{y}|\theta)p(\theta)d\theta} \propto p(\boldsymbol{y}|\theta)p(\theta).$$

In regression modeling, typically the unknown parameter is the vector of coefficients $\boldsymbol{\beta}$ and one obtains

$$p(\boldsymbol{\beta}|\boldsymbol{y}) \propto p(\boldsymbol{y}|\boldsymbol{\beta})p(\boldsymbol{\beta}).$$

The posterior mode estimator, which maximizes the posterior distribution, may be obtained by maximizing the logarithm of the posterior, yielding

$$\text{argmax}_{\boldsymbol{\beta}} = \text{argmax}_{\boldsymbol{\beta}}(l(\boldsymbol{\beta}) + \log(p(\boldsymbol{\beta}))),$$

where $l(\boldsymbol{\beta}) = \log(p(\boldsymbol{y}|\boldsymbol{\beta}))$ is the log-likelihood. Therefore, the posterior mode estimator is equivalent to the penalized likelihood estimator for an appropriately chosen prior distribution.

For the Gaussian prior $\boldsymbol{\beta} \sim N(\boldsymbol{0}, \tau^2 \boldsymbol{I})$, the log-prior has the form of the ridge penalty, which, apart from additive constants, has the form

$$\log(p(\boldsymbol{\beta})) = -\frac{1}{2\tau^2}\boldsymbol{\beta}^T\boldsymbol{\beta}.$$

For the normal distribution linear model with fixed variance $\sigma^2$, the posterior is given by

$$\boldsymbol{\beta}|\boldsymbol{y} \sim \mathrm{N}((\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}\boldsymbol{X}^T\boldsymbol{y}, \sigma^2(\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}),$$

where $\lambda = \sigma^2/\tau^2$. The assumption of an iid Laplace prior (also called double-exponential prior), which has density

$$p(\boldsymbol{\beta}) = \prod_{j=1}^{p} \frac{\lambda}{2} e^{-\lambda|\beta_j|}$$

yields, apart from constants, the lasso penalty

$$\log(p(\boldsymbol{\beta})) = -\lambda \sum_{j=1}^{p} |\beta_j|.$$

Therefore, frequentist regularization corresponds to posterior mode estimation if all the other parameters, for example dispersion parameters, are fixed. However, from a Bayesian point of view, maximizing the posterior is not the best way to obtain estimates. A fully Bayesian approach will use the mean or median of the posterior to estimate the coefficient vector. The Bayesian lasso, proposed by Park and Casella (2008), uses posterior median estimates. It does not automatically perform variable selection but provides standard errors and Bayesian credible intervals that can be used to select variables. Park and Casella (2008) used the representation of the double-exponential distribution as a mixture of normals to generate a simple Gibbs sampler. A direct representation of the posterior distribution was used by Hans (2009). An overview of regularization techniques, including bridge regression penalties and elastic net penalties, is given by Fahrmeir and Kneib (2009).

## 6.7 Further Reading

*Alternative Methods and Surveys*. One of the first methods to obtain variable selection by shrinkage was the non-negative garotte, proposed by Breiman (1995). Although it is consistent (Zou, 2006), its performance is often poor in highly correlated settings. An overview on regularization in linear models is found in Hastie, Tibshirani, and Friedman (2009). More recently, Bühlmann and Van De Geer (2011) gave a thorough mathematical treatment of regularization methods.

*R Packages*. Lasso and elastic-net regularized generalized linear models can be fitted with the R package *glmnet*, which allows one to fit Gaussian, binomial, Poisson, and *multinomial* responses (Friedman et al., 2008). An alternative is *glmpath* (Park and Hastie, 2007), which fits models with Gaussian, binomial, and Poisson responses. The package *penalized* (Goeman, 2010) is designed for the Cox model but also fits logit and Poisson distribution models. The group lasso for metric response, the logit, and the log-linear Poisson model can be fitted by use of the R package *grplasso* (Meier et al., 2008). The package *ordPens* is able to handle ordinal predictors. Boosting is available in the packages *mboost* (Hothorn et al., 2009) and *GAMBoost*.

## 6.8 Exercises

6.1 Consider the simple linear regression model.

(a) Give the lasso and ridge estimators as functions of the ML estimate for a one-dimensional predictor.

(b) Give the lasso and ridge estimators as functions of the ML estimate for an orthogonal design.

6.2 Show by using a second-order Taylor approximation that the log-likelihood of a GLM at the maximum likelihood estimate $\hat{\boldsymbol{\beta}}_{ML}$ can be approximated by $l(\hat{\boldsymbol{\beta}}_{ML}) + \frac{1}{2}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_{ML})^T \boldsymbol{F}(\hat{\boldsymbol{\beta}}_{ML})(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_{ML})$, where $\boldsymbol{F}(\hat{\boldsymbol{\beta}}_{ML})$ denotes the Fisher matrix.

6.3 The birth weight data, which have also been considered by Hosmer and Lemeshow (1989) and Venables and Ripley (2002), are available from the R package MASS (dataset birthwt). The data contain 189 observations that were collected at Baystate Medical Center, Springfield, Massachusetts, during 1986. The binary response is an indicator of birth weight less than 2.5 kg. The predictor variables are mother's age in years (age), mother's weight in pounds at last menstrual period (lwt), mother's race (1 = white, 2 = black, 3 = other), smoking status during pregnancy (smoke), number of previous premature labours (ptl), history of hypertension (ht), presence of uterine irritability (ui), and number of physician visits during the first trimester (ftv).

(a) Use regularized estimates to fit a binary regression model; use in particular ML, ridge, lasso, SCAD, elastic net, and boosting. R packages that might be useful are mentioned in the previous section.

(b) Compare the performance of the fitting methods in terms of prediction error by splitting the dataset several times into training data (for fitting) and test data (for evaluation of prediction).

# Chapter 7

# Regression Analysis of Count Data

In many applications the response variable is given in the form of event counts, where an event count refers to the number of times an event occurs. Simple examples are

- number of insolvent firms within a fixed time internal,
- number of insurance claims within a given period of time,
- number of epileptic seizures per day,
- number of cases with a specific disease in epidemiology.

In all of theses examples the response $y$ may be viewed as a non-negative integer-valued random variable with $y \in \{0, 1, 2, \dots\}$. Although in many applications there is an upper bound for the response, because the number of firms or potential insurance claims is finite, the upper bound is often very large and considered as irrelevant in modeling. In other cases, for example, for the number of epileptic seizures, no upper bound is given. In the following some further examples are given.

### Example 7.1: Number of Children

The German General Social Survey Allbus provides micro data, which allow one to model the dependence of the number of children on explanatory variables. We will consider women only and the predictors age in years (age), duration of school education (dur), nationality (nation, 0: German, 1: otherwise), religion (answer categories to "God is the most important in man", 1: strongly agree,. . . , 5: strongly disagree, 6: never thought about it), university degree (univ, 0: no, 1: yes). □

### Example 7.2: Encephalitis

In a study on the occurrence of encephalitis in Central Europe (Karimi et al., 1998), the number of cases of herpes encephalitis in children was observed between 1980 and 1993 in Bavaria and Lower Saxony. Table 7.1 shows the resulting counts. □

The count data in Example 7.2 form a contingency table. Therefore, one might think of using tools for the analysis of contingency tables, a topic that is treated extensively in Chapter 12. Classical analysis of contingency tables treats the rows and columns as factors and thus does not not use the full information available in the potential predictors. It seems more appropriate to model time as a metric variable rather than a qualitative variable. Then the regression problem is determined by the qualitative explanatory variable country and the metric covariate time.

TABLE 7.1: Encephalitis infection in children.

| | Bavaria | Lower Saxony |
|---|---|---|
| 1980 | 1 | 2 |
| 1981 | 0 | 1 |
| 1982 | 1 | 2 |
| 1983 | 2 | 5 |
| 1984 | 2 | 4 |
| 1985 | 3 | – |
| 1986 | 8 | – |
| 1987 | 5 | 6 |
| 1988 | 13 | 7 |
| 1989 | 12 | 7 |
| 1990 | 6 | 7 |
| 1991 | 13 | 3 |
| 1992 | 10 | 4 |
| 1993 | 12 | 2 |

The benchmark model for count data is the Poisson distribution. Therefore, we will first consider the Poisson regression model, which can be treated within the framework of generalized linear models. One of the disadvantages of the Poisson distribution is that it is a one parameter distribution. Consequently, the Poisson regression is frequently not flexible enough to adapt to the given data. One step to more flexible models is the inclusion of an overdispersion parameter (Section 7.5). An alternative, more flexible model is the negative binomial model, which can be motivated as a mixture model (Section 7.6). Models for data that show overdispersion through excess zeros are considered in Section 7.7 and Section 7.8.

## 7.1 The Poisson Distribution

The Poisson distribution is a standard model for count data and was derived as a limiting case of the binomial by Poisson (1837). The discrete random variable $Y$ is Poisson-distributed with intensity or rate parameter $\lambda, \lambda > 0$, if the density is given by

$$P(Y = y) = \begin{cases} \frac{\lambda^y}{y!} e^{-\lambda} & \text{for } y \in \{0, 1, 2, \dots\} \\ 0 & \text{otherwise.} \end{cases} \tag{7.1}$$

An abbreviation is $Y \sim \mathrm{P}(\lambda)$. Figure 7.1 shows several examples of densities of Poisson-distributed variables. A closer look at the form of the densities is obtained by considering the proportion of probabilities for counts $y$ and $y - 1$ $(y \geq 1)$ given by

$$\frac{P(Y = y)}{P(Y = y - 1)} = \frac{\lambda^y e^{-\lambda}/y!}{\lambda^{y-1} e^{-\lambda}/(y-1)!} = \frac{\lambda}{y}.$$

If $\lambda < 1$ one has $\lambda/y < 1$ and therefore the density is decreasing across integers, the largest probability occurs at $y = 0$. If $\lambda > 1$, the probabilities are increasing up to the integer value of $\lambda, [\lambda]$; for $y > [\lambda]$ the density is decreasing. Thus, for non-integer-valued $\lambda$ the density is unimodal with the mode given by $[\lambda]$. If $\lambda$ is integer-valued, the probabilities $P(Y = \lambda)$ and $P(Y = \lambda - 1)$ are equal.

The first two central moments of the Poisson distribution are given by $\mathrm{E}(Y) = \mathrm{var}(Y) = \lambda$. Equality of the mean and variances is often referred to as the *equidispersion property* of the Poisson distribution. Thus, in contrast to the normal distribution, for which the mean and variance are unlinked, the Poisson distribution implicitly models stronger variability for larger

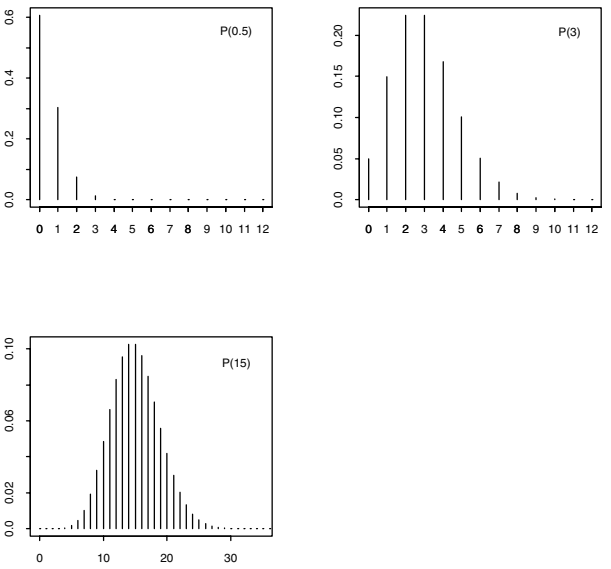


FIGURE 7.1: Probability mass functions of Poisson distributions with $\lambda = 0.5, 3, 15$.

means, a property that is often found in real-life data. On the other hand, in real-life data one frequently finds that the variance exceeds the mean, and the effect is overdispersion, which has to be modeled separately (see Section 7.5).

In the following some connections between distributions and some additional properties of the Poisson distribution are given. The connections between distributions help us to get a clearer picture of the range of applications of the Poisson distribution.

### Poisson Distribution as the Law of Rare Events

The Poisson distribution may be obtained as a limiting case of the binomial distribution. Let $Y$ denote the total number of successes in a large number $n$ of independent Bernoulli trials with the successes probability $\pi$ being small and linked to the number of trials by $\pi = \lambda/n$. As an example, one may consider the number of incoming telephone calls in a fixed time internal of unit length. Let $\lambda$ denote the fixed mean number of calls. Now consider the division of the time interval into $n$ subintervals with equal width. For small intervals, each interval may be considered as one trial with the success being defined as an incoming call. Then the number of total incoming calls can be modeled by a binomial distribution with the probability specified by $\lambda/n$. Then it may be shown that for increasing $n$ the binomial distribution becomes the Poisson distribution.

More formally, let $Y$ have a binomial distribution with parameter $n$ and $\pi = \lambda/n, Y \sim \mathrm{B}(n, \pi = \lambda/n)$. Then one has

$$\lim_{\substack{n\pi=\lambda \\ n\to\infty}} \binom{n}{y} \pi^y (1-\pi)^{n-y} = \frac{\lambda^y}{y!} e^{-\lambda}.$$

The law of rare events refers to this derivation from the binomial distribution, where the number of trials increases while the probability of success decreases correspondingly. However, the term is somewhat misleading because the mean $\lambda$ may be arbitrarily large. The Poisson distribution is not restricted to small values of the mean.

The phone calls example is an idealization with some missing details. For example, in one time subinterval more than one phone call could occur. A more concise derivation for this example is obtained by considering Poisson processes.

### Poisson Process

The Poisson distribution is closely linked to the Poisson process. Let $\{N(t), t \geq 0\}$ be a counting process with $N(t)$ denoting the event counts up to time $t$. $N(t)$ is a non-negative and integer-valued random variable, and the process is a collection of these random variables satisfying the property that $N(s) \leq N(t)$ if $s < t$. The Poisson process is a specific counting process that has to fulfill several properties. With $N(t, t + \Delta t)$ denoting the number of counts in interval $(t, t + \Delta t)$, one postulates

(a) *Independence of intervals*
For disjunct time intervals $(s, s + \Delta s)$ and $(t, t + \Delta t)$, the increments $N(s, s + \Delta s)$ and $N(t, t + \Delta t)$ are independent.

(b) *Stationarity*
The distribution of the counts in the interval $(t, t + \Delta t)$ depends only on the length of the interval $\Delta t$ (it does not depend on $t$).

(c) *Intensity rate*
The probability of no or one event in interval $(t, t + \Delta t)$ is given by

$$P(N(t, t + \Delta t) = 1) = \lambda \Delta t + o(\Delta t),$$
$$P(N(t, t + \Delta t) = 0) = 1 - \lambda \Delta t + o(\Delta t),$$

where $o(h)$ denotes a remainder term with the property $o(h)/h \to 0$ as $h \to 0$.

For this process, the number of events occurring in the interval $(t, t + \Delta t)$ is Poisson-distributed with mean $\lambda \Delta t$:

$$N(t, t + \Delta t) \sim \mathrm{P}(\lambda \Delta t).$$

In particular, one has $N(t) = N(0, \Delta t) \sim \mathrm{P}(\lambda \Delta t)$. To obtain a Poisson distribution in the phone call example one has to postulate not only that the probability of occurrence in one time interval depends only on the length of the interval but also the independence of counts in intervals, stationarity, and that the probability that no or one event has a specific form in the limit.

The Poisson process is also strongly connected to the exponential distribution. If the Poisson process is a valid model, the waiting time between events follows an exponential distribution. It is immediately seen that for the waiting time for the first event, $W_1$, the outcome $W_1 > t$ occurs if no events occur in the interval $(0, t)$:

$$P(W_1 > t) = P(N(0, t) = 0) = \mathrm{e}^{-\lambda t}.$$

Therefore, $W_1$ follows the exponential distribution with parameter $\lambda$. The same distribution may be derived for the waiting time between any two events. Alternative characterizations of the Poisson distributions are considered, for example, in Cameron and Trivedi (1998).

### Further Properties

(1) Sums of independent Poisson-distributed random variables are Poisson-distributed. More concrete, if $Y_i \sim \mathrm{P}(\lambda_i), i = 1, 2, \ldots$, are independent and $\sum_i \lambda_i < \infty$, then $\sum_i Y_i \sim \mathrm{P}(\sum_i \lambda_i)$.

(2) There is a strong connection between the Poisson and the multinomial distribution. If $Y_1, \ldots, Y_N$ are independent Poisson variables, $Y_i \sim \mathrm{P}(\lambda_i)$, and one conditions on the total sum $n_0 = Y_1, + \cdots + Y_N$, one obtains for $(Y_1, \ldots, Y_N)$ *given* $n_0$ the multinomial distribution $\mathrm{M}(n_0, (\pi_1, \ldots, \pi_N))$ with $\pi_i = \lambda_i/(\lambda_1 + \cdots + \lambda_N)$.

(3) For large values of $\lambda$ the Poisson distribution $\mathrm{P}(\lambda)$ may be approximated by the normal distribution $\mathrm{N}(\lambda, \lambda)$. It is seen from Figure 7.1 that for small $\lambda$ the distribution is strongly skewed. For larger values of $\lambda$ the distribution becomes more symmetric. For details of approximations see, for example, McCullagh and Nelder (1989), Chapter 6.

## 7.2 Poisson Regression Model

The Poisson regression model is the standard model for count data. Let $(y_i, \boldsymbol{x}_i)$ denote $n$ independent observations and $\mu_i = \mathrm{E}(y_i|\boldsymbol{x}_i)$. One assumes that $y_i|\boldsymbol{x}_i$ is Poisson-distributed with mean $\mu_i$, and that the mean is determined by

$$\mu_i = h(\boldsymbol{x}_i^T\boldsymbol{\beta}) \quad or \quad g(\mu_i) = \boldsymbol{x}_i^T\boldsymbol{\beta}, \tag{7.2}$$

where $g$ is a known link function and $h = g^{-1}$ denotes the response function. Since the Poisson distribution is from the simple exponential family, the model is a generalized linear model. The most widely used model uses the canonical link function by specifying

$$\mu_i = \exp(\boldsymbol{x}_i^T\boldsymbol{\beta}) \quad \text{or} \quad \log(\mu_i) = \boldsymbol{x}_i^T\boldsymbol{\beta}.$$

Since the logarithm of the conditional mean is linear in the parameters, the model is called a *log-linear* model.

The log-linear version of the model is particulary attractive because interpreting the parameters is very easy. The model implies that the conditional mean given $\boldsymbol{x}^T = (x_1, \ldots, x_p)$ has a multiplicative form given by

$$\mu(\boldsymbol{x}) = \exp(\boldsymbol{x}^T\boldsymbol{\beta}) = \mathrm{e}^{x_1\beta_1} \ldots \mathrm{e}^{x_p\beta_p}.$$

Thus $\mathrm{e}^{\beta_j}$ represents the multiplicative effect on $\mu(\boldsymbol{x})$ if the variable $x_j$ changes by one unit to $x_j + 1$ (given that the rest of the variables are fixed). One obtains

$$\frac{\mu(x_1, \ldots, x_j + 1, \ldots, x_p)}{\mu(x_1, \ldots, x_j, \ldots, x_p)} = \mathrm{e}^{\beta_j}$$

or, equivalently,

$$\log\, \mu(x_1, \ldots, x_j + 1, \ldots, x_p) - \log\, \mu(x_1, \ldots, x_j, \ldots, x_p) = \beta_j.$$

While $\beta_j$ is the change in log-means if $x_j$ increases by one unit, $\mathrm{e}^{\beta_j}$ is the multiplicative effect, which is easier to interpret because it directly effects upon the mean.

For illustration, let us consider Example 1.5, where the dependent variable is the number of insolvent firms and there is only one covariate, namely, time. The log-linear Poisson model

$$\log(\mu) = \beta_0 + \text{time}\beta$$

specifies the number of insolvent firms in dependence on time (1 to 36 for January 1994 to December 1996). One obtains the estimates $\hat{\beta}_0 = 4.25$ and $\hat{\beta} = 0.0097$, yielding $e^{\hat{\beta}} = 1.01$. Therefore, the log-mean increases additively by 0.0097 every month or, more intuitively, the mean increases by the factor 1.01 every month.

The canonical link has the additional advantage that the mean $\mu(\boldsymbol{x})$ is always positive whatever values the (estimated) parameters take. For alternative models like the linear model, $\mu_i = \boldsymbol{x}_i^T\boldsymbol{\beta}$ problems may occur when the mean is predicted for new observations $\boldsymbol{x}_i$, although $\mu_i$ may take admissible values in the original sample. Therefore, in most applications one uses the log-link. Nevertheless, the results can be misleading if the link is grossly misspecified. When nothing is known about the link one can also use non-parametric approaches to link specification (see Section 5.2 and Section 6.4).

## 7.3 Inference for the Poisson Regression Model

### Maximum Likelihood Estimation

For inference, the whole machinery of generalized linear models may be used. For model (7.2) one obtains the log-likelihood

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{n} y_i \log(\mu_i) - \mu_i - \log(y_i!) = \sum_{i=1}^{n} y_i \log(h(\boldsymbol{x}_i^T\boldsymbol{\beta})) - h(\boldsymbol{x}_i^T\boldsymbol{\beta}) - \log(y_i!).$$

The score function $\boldsymbol{s}(\boldsymbol{\beta}) = \partial l(\boldsymbol{\beta})/\partial\boldsymbol{\beta}$ is given by

$$\boldsymbol{s}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \boldsymbol{x}_i \frac{h'(\boldsymbol{x}_i^T\boldsymbol{\beta})}{h(\boldsymbol{x}_i\boldsymbol{\beta})}(y_i - h(\boldsymbol{x}_i^T\boldsymbol{\beta})),$$

and the Fisher matrix is given by

$$\boldsymbol{F}(\boldsymbol{\beta}) = \mathrm{E}(-\partial^2 l(\boldsymbol{\beta})/\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T) = \sum_{i=1}^{n} \boldsymbol{x}_i\boldsymbol{x}_i^T \frac{h'(\boldsymbol{x}_i^T\boldsymbol{\beta})^2}{h(\boldsymbol{x}_i^T\boldsymbol{\beta})}. \tag{7.3}$$

For the canonical link $h(\eta) = \exp(\eta)$, these terms simplify to

$$\begin{aligned}
l(\boldsymbol{\beta}) &= \sum_{i=1}^{n} y_i\boldsymbol{x}_i^T\boldsymbol{\beta} - \exp(\boldsymbol{x}_i^T\boldsymbol{\beta}) - \log(y_i!),\\
\boldsymbol{s}(\boldsymbol{\beta}) &= \sum_{i=1}^{n} \boldsymbol{x}_i(y_i - \exp(\boldsymbol{x}_i^T\boldsymbol{\beta})),\\
\boldsymbol{F}(\boldsymbol{\beta}) &= \sum_{i=1}^{n} \boldsymbol{x}_i\boldsymbol{x}_i^T \exp(\boldsymbol{x}_i^T\boldsymbol{\beta}).
\end{aligned}$$

Under regularity conditions, $\hat{\boldsymbol{\beta}}$ defined by $\boldsymbol{s}(\hat{\boldsymbol{\beta}}) = \mathbf{0}$ is consistent and asymptotically normal distributed:

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

where $\boldsymbol{F}(\boldsymbol{\beta})$ may be replaced by $\boldsymbol{F}(\hat{\boldsymbol{\beta}})$ to obtain standard errors.

### Deviance and Goodness-of-Fit

The deviance as a measure of discrepancy between the fit and the data compares the log-likelihood of the fitted value for observation $y_i$, denoted by $l_i(\hat{\mu}_i) = y_i \log(\hat{\mu}_i) - \hat{\mu}_i - \log(y_i!)$, to the log-likelihood of the perfect fit $l_i(y_i) = y_i \log(y_i) - y_i - \log(y_i!)$, yielding

$$D = -2\sum_i l_i(\hat{\mu}_i) - l_i(y_i) = 2\sum_i \{y_i \log\left(\frac{y_i}{\hat{\mu}_i}\right) + [(\hat{\mu}_i - y_i)]\}.$$

If an intercept is included, the term in brackets, $\hat{\mu}_i - y_i$, may be omitted. Within the framework of GLMs the deviance is used as a goodness-of-fit statistic with a known asymptotic distribution when the observations are grouped. Let $y_{i1}, \ldots, y_{in_i}$, $i = 1, \ldots, N$, denote independent observations at a fixed measurement point $\boldsymbol{x}_i$ with $y_{it} \sim \mathrm{P}(\tilde{\mu}_i), \tilde{\mu}_i = h(\boldsymbol{x}_i^T \boldsymbol{\beta})$. Then $y_i = n_i \bar{y}_i = \sum_{t=1}^{n_i} y_{it} \sim \mathrm{P}(n_i \tilde{\mu}_i)$, which may be written as $y_i \sim \mathrm{P}(\mu_i)$, where $\mu_i = n_i \tilde{\mu}_i$. When using $\mu_i$, the deviance for grouped observations has the same form as for single observations:

$$D = 2 \sum_{i=1}^{N} \{ y_i \log \left( \frac{y_i}{\hat{\mu}_i} \right) + [(\hat{\mu}_i - y_i)] \}. \tag{7.4}$$

$D$ is asymptotically $\chi^2$-distributed with $N - p$ degrees of freedom, where $p$ is the dimension of the parameter vector. The underlying asymptotic concept is fixed cells asymptotic, where $N$ is fixed and $n_i \to \infty$ for all $i$. In the form (7.4), where $n_i$ is only implicitly contained in $\mu_i = n_i \tilde{\mu}_i$, one assumes $\mu_i \to \infty$. The assumption $\mu_i \to \infty$ is slightly more general because one does not have to assume that $y_i$ is composed from repeated measurements. As an alternative goodness-of fit statistic one may also use the Pearson statistic considered in Chapter 3. The specific form is given in the box.

**Goodness-of-Fit for Poisson Regression Model**

$$D = 2 \sum_{i=1}^{N} y_i \log \left( \frac{y_i}{\hat{\mu}_i} \right)$$

$$\chi^2_P = \sum_{i=1}^{N} \frac{(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i}$$

For $\mu_i \to \infty$ one obtains the approximation

$$D, \chi^2_P \sim \chi^2(N - p)$$

The use of the deviance and the Pearson statistic depends on whether asymptotic results apply. Usually one expects all of the means to be larger than three. Fienberg (1980) showed that the approximation might work even if for a small percentage the mean is only one; see also Read and Cressie (1988). If $D$ and $\chi^2_P$ are quite different, one might always suspect that the approximation is inadequate.

## Testing of Hierarchical Models

The deviance may also be used to test hierarchical models $\tilde{M} \subset M$, where $\tilde{M}$ is determined by a linear hypothesis $\boldsymbol{C}\boldsymbol{\beta} = \xi$. The difference between deviances for the fit of model $\tilde{M}$ (yielding $\tilde{\boldsymbol{\beta}}$ and $\tilde{\mu}_i$) and model $M$ (yielding $\hat{\boldsymbol{\beta}}$ and $\hat{\mu}_i$) is

$$D(\tilde{M}|M) = 2 \sum_i y_i \log(\frac{\tilde{\mu}_i}{\hat{\mu}_i}) + (\hat{\mu}_i - \tilde{\mu}_i),$$

which has asymptotically a $\chi^2$-distribution with the degrees of freedom given by the rank of $\boldsymbol{C}$.

TABLE 7.2: Parameter estimates and log-linear Poisson model for number of children.

| | Estimate | Std. Error | $z$-Value | Pr($>\|z\|$) |
|---|---|---|---|---|
| Intercept | –12.280 | 1.484 | –8.27 | 0.0000 |
| age | 0.935 | 0.124 | 7.55 | 0.0000 |
| age$^2$ | –0.025 | 0.004 | –6.57 | 0.0000 |
| age$^3$ | 0.000 | 0.000 | 5.78 | 0.0000 |
| age$^4$ | –0.000 | 0.000 | –5.14 | 0.0000 |
| dur | 0.112 | 0.067 | 1.68 | 0.0929 |
| dur$^2$ | –0.008 | 0.003 | –2.77 | 0.0054 |
| nation | 0.056 | 0.138 | 0.41 | 0.6816 |
| god2 | –0.010 | 0.059 | –1.73 | 0.0826 |
| god3 | –0.144 | 0.068 | –2.14 | 0.0327 |
| god4 | –0.128 | 0.071 | –1.80 | 0.0711 |
| god5 | –0.036 | 0.067 | –0.54 | 0.5886 |
| god6 | –0.092 | 0.075 | –1.23 | 0.2182 |
| univ | 0.637 | 0.173 | 3.68 | 0.0002 |

TABLE 7.3: Deviances for Poisson model, number of children as response.

| | DF | Difference | DF | Deviance |
|---|---|---|---|---|
| All effects | | | 1747 | 1718.6 |
| age | 4 | 215.5 | 1751 | 1940.7 |
| dur | 2 | 40.3 | 1749 | 1758.9 |
| nationality | 1 | 0.2 | 1748 | 1718.8 |
| religion | 5 | 6.7 | 1752 | 1725.3 |
| univ | 1 | 13.5 | 1748 | 1732.1 |

### Example 7.3: Number of Children

For the data described in Example 7.1 a log-linear Poisson model with the number of children as the dependent variable has been fitted. Table 7.2 shows the estimates. Since it is hardly to be expected that the metric predictors age and duration of school education have a linear effect, the polynomial terms are included in the predictor, which turn out to be highly significant. The only variable that seems to be not influential is nationality. However, the analysis of deviance, given in Table 7.3, shows that the effect of the variable religion is also not significant. The table gives the deviance of the model that contains all the predictors and the differences between deviances when one predictor is omitted. Since the effects of polynomial terms are hard to see from the estimates, the effects are plotted in Figure 7.2, with all other covariates considered as fixed. It is seen that especially for women between 20 and 40, age makes a difference as far as the expected number of children is concerned. The duration of school years shows that unfortunately the time spent in school decreases the number of children. For a more flexible modelling see Example 10.4, Chapter 10. □

### Example 7.4: Encephalitis

For the encephalitis dataset, the number of infections is modeled in dependence on country (BAV=1: Bavaria, BAV=0: Lower Saxony) and TIME (1–14, corresponding to 1980–1993). The compared models are the log-linear Poisson model, the normal distribution model with log-link and the identity link. Table 7.4 shows the fits with an interaction effect between country and time. When using a normal distribution model, the log-linear model is to be preferred because its log-likelihood is larger. Comparison across distributions cannot be recommended because log-likelihoods are not comparable. The points in

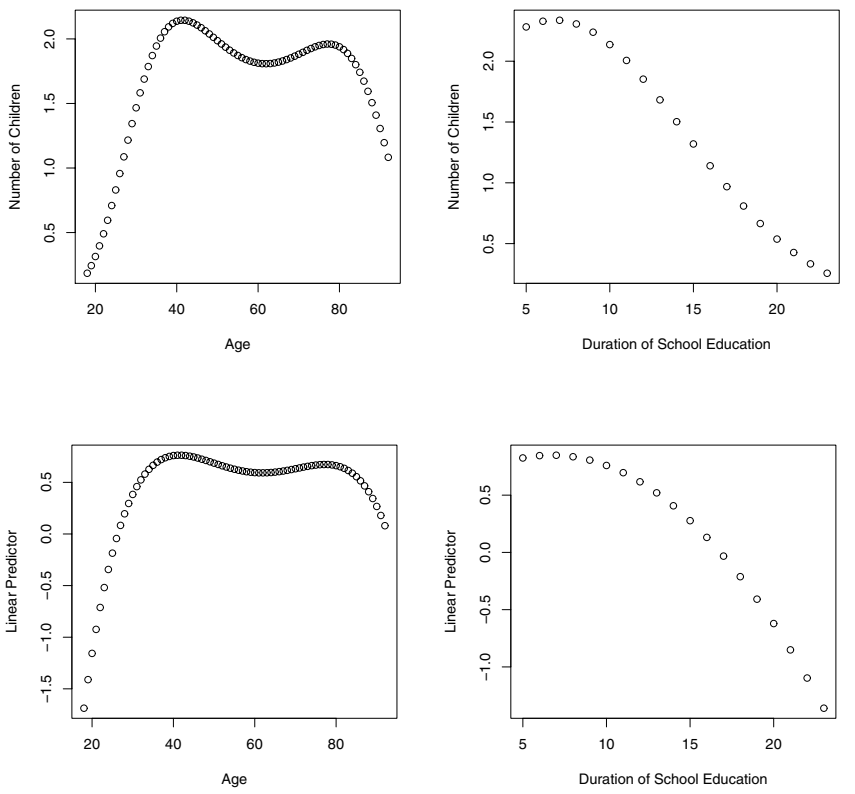


FIGURE 7.2: Number of children versus age and duration of education

TABLE 7.4: Models for encephalitis data.

| | Log-Linear Poisson Model | | Log-Linear Normal Model | | Linear Normal Model | |
|---|---|---|---|---|---|---|
| | Estimate | $p$-Value | Estimate | $p$-value | Estimate | $p$-Value |
| Intercept | -0.255 | 0.622 | -0.223 | 0.705 | 0.397 | 0.815 |
| TIME | 0.513 | 0.000 | 0.499 | 0.0002 | 1.154 | 0.014 |
| TIME$^2$ | -0.030 | 0.0001 | -0.029 | 0.0002 | -0.065 | 0.030 |
| BAV | -1.587 | 0.006 | -1.478 | 0.017 | -4.414 | 0.014 |
| BAV.TIME | 0.211 | 0.003 | 0.198 | 0.001 | 0.853 | 0.000 |
| Log-likelihood | -47.868 | | -51.398 | | -54.905 | |

favor of the Poisson model are that data are definitely discrete and the equidispersion property of the Poisson model. It is seen from Figure 7.3 that large means tend to have larger variability. □

### Example 7.5: Insolvent Firms

For the number of insolvent firms between 1994 and 1996 (see Example 1.5) a log-linear Poisson model is fitted with time as the predictor. Time is considered as a number from 1 to 36, denoting months, starting with January 1994 and ending with December 1996. Since the counts are not too small, one might also fit a model that assumes normally distributed responses. The models that are compared are the log-linear model $\log(\mu) = \beta_0 + x\beta_1$ and the model $\log(\mu) = \beta_0 + x\beta_1 + x^2\beta_2$ with $x \in \{1, \ldots, 36\}$.

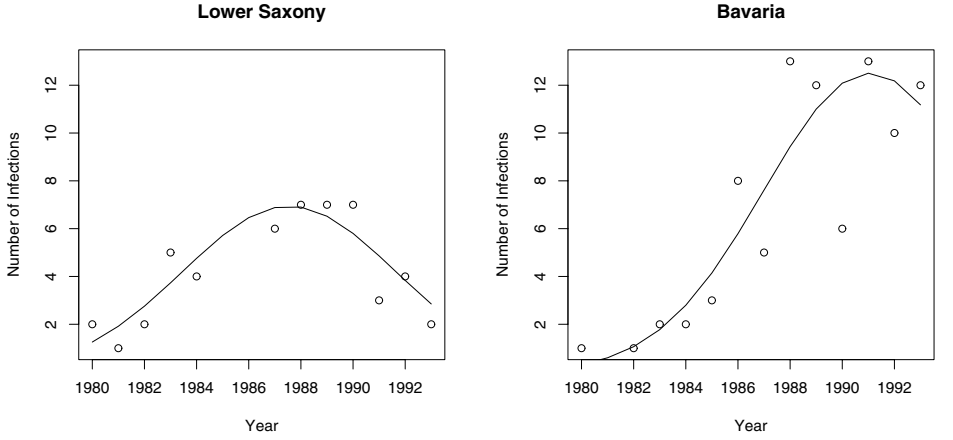


FIGURE 7.3: Estimated means against time for log-linear Poisson model for encephalitis data, Lower Saxony (upper panel) and Bavaria (lower panel).

The results given in Table 7.5 show that the quadratic terms seem to be unnecessary. Nevertheless, in Figure 7.4, which shows the mean response against months, the quadratic term is included. □

TABLE 7.5: Log-linear Poisson models for insolvency data.

| | Estimate | Standard Error | $z$-Value |
|---|---|---|---|
| $\beta_0$ | 4.192 | 0.062 | 67.833 |
| $\beta_1$ | 0.020 | 0.007 | 2.677 |
| $\beta_2$ | –0.00027 | 0.00019 | –1.408 |

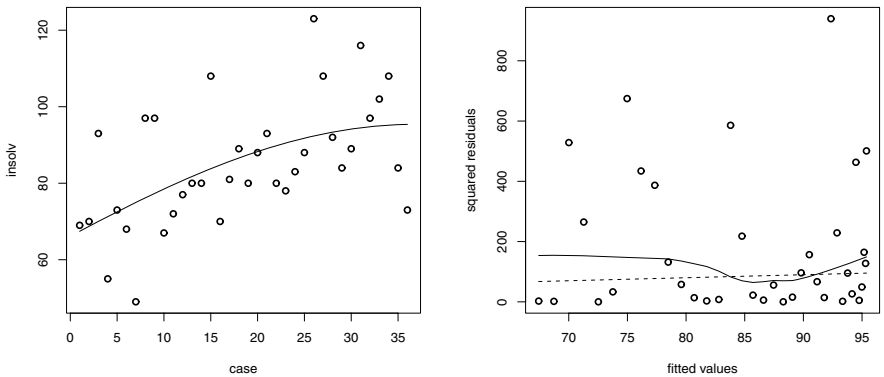


FIGURE 7.4: Log-linear model for insolvency data plotted against months (left) and squared residuals (right) against fitted values.

## 7.4 Poisson Regression with an Offset

In the standard log-linear Poisson model it is assumed that the log-mean of the response depends directly on the covariates in linear form, $\log(\mu_i) = \boldsymbol{x}_i^T\boldsymbol{\beta}$. In many applications, the response results from different levels of aggregation, and it is more appropriate to model the underlying

driving force. For example, in epidemiology, if the incidence of infectious diseases is studied, count data may refer to geographical districts with varying population sizes. Thus the number of people at risk has to be taken into account. The same effect is found if the counts are observed in different time intervals. Let us consider the latter case. Then, one can use the strong connection between the Poisson distribution and the Poisson process as described in Section 7.1. By assuming a Poisson process with intensity rate $\lambda$, one obtains for the counts in intervals of length $\Delta$ the Poisson distribution $y \sim \mathrm{P}(\lambda\Delta)$. Consequently, the mean $\mu = \lambda\Delta$ depends on the length of the interval. Let data be given by $(y_i, \boldsymbol{x}_i)$, where $y_i$ denotes the counts in intervals of length $\Delta_i$. If counts arise from a Poisson process with intensity rate $\lambda_i$ (depending on $\boldsymbol{x}_i$), one obtains $y_i \sim \mathrm{P}(\Delta_i\lambda_i)$. If the dependence of the intensity rate is modeled in log-linear form, $\log(\lambda_i) = \boldsymbol{x}_i^T\boldsymbol{\beta}$, one obtains for mean counts $\mu_i = \mathrm{E}(y_i|\boldsymbol{x}_i)$

$$\log(\mu_i) = \log(\Delta_i) + \boldsymbol{x}_i^T\boldsymbol{\beta} \tag{7.5}$$

or, equivalently,

$$\mu_i = \exp(\log(\Delta_i) + \boldsymbol{x}_i^T\boldsymbol{\beta}).$$

Therefore, the specification of the driving force, represented by $\lambda_i$, yields a model with a fixed term $\log(\Delta_i)$ in it. For an application, see Santner and Duffy (1989).

A similar form of the model results if one has different levels of aggregation. Let independent responses $y_{i1}, \ldots, y_{in_i}$ be observed for fixed values of explanatory variables $\boldsymbol{x}_i$. If $y_{it} \sim \mathrm{P}(\lambda_i)$, one obtains for the sum of responses

$$y_i = \sum_{t=1}^{n_i} y_{it} = n_i\bar{y}_i \sim \mathrm{P}(n_i\lambda_i),$$

where $\bar{y}_i = y_i/n$. Since the local sample sizes $n_i$ vary across measurements, one obtains with $\log(\lambda_i) = \boldsymbol{x}_i^T\boldsymbol{\beta}$ for the mean responses $\mu_i = \mathrm{E}(y_i|\boldsymbol{x}_i)$

$$\log(\mu_i) = \log(n_i) + \boldsymbol{x}_i^T\boldsymbol{\beta}. \tag{7.6}$$

The number $n_i$ can be given implicitly, for example, as the population size in a given geographical district when one looks at the number of cases of a specific disease. From the form $\log(\mu_i/n_i) = \boldsymbol{x}_i^T\boldsymbol{\beta}$ it is seen that the appropriately standardized rate of occurrence is modeled rather than the number of cases itself.

Models (7.5) and (7.6) have the general form

$$\log(\mu_i) = \gamma_i + \boldsymbol{x}_i^T\boldsymbol{\beta},$$

where $\gamma_i$ is a known parameter $\gamma_i = \log(\Delta_i)$ for the varying length of time intervals, and $\gamma_i = \log(n_i)$ for varying sample sizes. The parameter $\gamma_i$ may be treated as an offset that remains fixed across interactions. The log-likelihood and the score function are given by $l(\boldsymbol{\beta}) = \sum_i y_i \log(\gamma_i + \boldsymbol{x}_i^T\boldsymbol{\beta}) - \exp(\gamma_i + \boldsymbol{x}_i^T\boldsymbol{\beta}) - \log(y_i!)$ and $\boldsymbol{s}(\boldsymbol{\beta}) = \sum_i \boldsymbol{x}_i(y_i - \exp(\gamma_i + \boldsymbol{x}_i^T\boldsymbol{\beta}))$. The Fisher matrix has the form

$$\boldsymbol{F}(\boldsymbol{\beta}) = \sum_{i=1}^{N} \boldsymbol{x}_i\boldsymbol{x}_i^T \exp(\gamma_i + \boldsymbol{x}_i^T\boldsymbol{\beta})) = \sum_{i=1}^{N} \boldsymbol{x}_i\boldsymbol{x}_i^T \exp(\boldsymbol{x}_i^T\boldsymbol{\beta}) \exp(\gamma_i).$$

It is directly seen how $\gamma_i$ determines the accuracy of the estimates. Since the ML estimate $\hat{\boldsymbol{\beta}}$ has approximate covariance $cov(\hat{\boldsymbol{\beta}}) \approx \boldsymbol{F}(\hat{\boldsymbol{\beta}})^{-1}$, standard errors decrease with increasing parameters $\gamma_i$. As is to be expected, larger time intervals or larger sample sizes yield better estimates.

## 7.5 Poisson Regression with Overdispersion

In many applications count data are overdispersed, with the conditional variance exceeding the conditional mean. One cause for this can be unmodeled heterogeneity among subjects. In the following several modeling approaches that account for overdispersion are considered. The first one is based on quasi-likelihood, and the second one models the heterogeneity among subjects explicitly. A specific model that models heterogeneity explicitly is the Gamma-Poisson or negative binomial model considered in Section 7.6. Also, models for excess zeros like the zero-inflated model (Section 7.7) and the hurdle model (Section 7.8) imply overdispersion.

### 7.5.1 Quasi-Likelihood Methods

Maximum likelihood estimates are based on the assumption $\mu_i = h(\boldsymbol{x}_i^T\boldsymbol{\beta})$ and $y_i|\boldsymbol{x}_i \sim \mathrm{P}(\mu_i)$. The estimates are obtained by setting the score function equal to zero. These assumptions may be weakened within the quasi-likelihood framework (see Section 3.11, Chapter 3). The link between the mean and the linear predictor has the usual GLM form $\mu_i = h(\boldsymbol{x}_i^T\boldsymbol{\beta})$, but instead of assuming a fixed distribution for $y_i$ only a mean–variance relationship is assumed. The ML estimation equation $\boldsymbol{s}(\hat{\boldsymbol{\beta}}) = \mathbf{0}$ has the general form

$$\sum_{i=1}^{n} \boldsymbol{x}_i \frac{\partial \mu_i}{\partial \eta} \frac{y_i - \mu_i}{\sigma_i^2} = \mathbf{0}, \tag{7.7}$$

where $\mu_i = h(\eta_i)$ and $\sigma_i^2$ is the variance that has the form $\sigma_i^2 = \phi v(\mu_i)$ with variance function $v(\mu_i)$. Since $\phi$ cancels out, the estimation depends only on the specification of the mean and the variance function. For the Poisson distribution, the latter has the form $v(\mu_i) = \mu_i$. For alternative variance functions, which do not necessarily correspond to a Poisson distribution, (7.7) is considered as the estimation equation yielding quasi-likelihood estimates.

#### Model with Overdispersion Parameter

A simple quasi-likelihood approach uses the variance function $v(\mu_i) = \mu_i$, which yields the variance $\sigma_i^2 = \phi\mu_i$ for some unknown constant $\phi$. The case $\phi > 1$ represents the *overdispersion* of the Poisson model, and the case $\phi < 1$, which is rarely found in applications, is called the *underdispersion*. If $\sigma_i^2 = \phi\mu_i$ is used in (7.7), $\phi$ drops out. Thus the estimation equation is identical to the likelihood equation for Poisson models. Consequently, parameter estimates are identical. However, the variance is inflated by overdispersion, since one obtains the asymptotic covariance

$$\mathrm{cov}(\hat{\boldsymbol{\beta}}) = \phi \boldsymbol{F}(\boldsymbol{\beta})^{-1},$$

with $\boldsymbol{F}(\boldsymbol{\beta})$ denoting the Fisher matrix from equation (7.3). Wedderburn (1974) proposed estimating the dispersion parameter by

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

where $p$ is the number of model parameters and $n - p$ is a degrees-of-freedom correction. The motivation for this estimator is that the variance function $v(\mu_i) = \mu_i$ implies $\mathrm{E}(y_i - \mu_i)^2 = \phi\mu_i$ and hence $\phi = \mathrm{E}((y_i - \mu_i)^2/\mu_i)$. $\hat{\phi}$ is motivated as a moment estimator with a degrees-of-freedom correction. There is a strong connection to the Pearson statistic, because $\hat{\phi} = \chi^2_P/(n-p)$. The approximation $\mathrm{E}(\chi^2_P) \approx n - p$ holds if $\sigma^2 = \mu$ is the underlying variance. For $\sigma^2 = \phi\mu$, one has $\mathrm{E}(\chi^2_P/\phi) \approx n - p$ and therefore $\mathrm{E}(\chi^2_P/(n-p)) \approx \phi$.

In summary, the variance $\sigma^2 = \phi\mu$ is easy to handle. One fits the usual Poisson model and uses the ML estimate. To obtain the correct covariance matrix of $\hat{\boldsymbol{\beta}}$ one multiplies the maximum likelihood covariance by $\hat{\phi}$. Maximum likelihood standard errors are multiplied by $\sqrt{\hat{\phi}}$ and $t$-statistics are divided by $\sqrt{\hat{\phi}}$.

### Alternative Variance Functions

Alternative variance functions usually continue to model the variance as a function of the mean. A general variance function that is in common use has the form

$$v(\mu_i) = \mu_i + \gamma\mu_i^m$$

for a fixed value of $m$. The choice $m = 2$ corresponds to the assumption of the negative binomial distribution (see Section 7.6) while the choice $m = 1$ yields $v(\mu_i) = (1 + \gamma)\mu_i$. Hence, the case $m = 1$ is equivalent to assuming $v(\mu_i) = \phi\mu_i$. Breslow (1984) used the negative binomial type variance within a quasi-likelihood approach.

#### Example 7.6: Demand for Medical Care

Deb and Trivedi (1997) analyzed the demand for medical care for individuals, aged 66 and over, based on a dataset from the U.S. National Medical Expenditure survey in 1987/88. The data are available from the archive of the *Journal of Applied Econometrics* and the *Journal of Statistical Software*; see also Kleiber and Zeileis (2008), and Zeileis et al. (2008). Like Zeileis et al. (2008) we consider the number of physician/non-physician office and hospital outpatient visits (ofp) as dependent variable. The regressors used in the present analysis are the number of hospital stays (hosp), self-perceived health status (poor, medium, excellent), number of chronic conditions (numchron), age, marital status, and number of years of education (shool). Since the effects vary across gender, only male patients are used in the analysis. Table 7.6 shows the fits of a log-linear Poisson model without and with overdispersion (residual deviance is $9665.7$ on $1770$ degrees of freedom). With an estimated overdispersion parameter $\hat{\phi} = 7.393$ the data are highly overdispersed. The negative binomial model (Table 7.8) shows similar effects but slightly smaller standard errors ($\hat{\nu} = 1.079$, with standard error 0.048, and the residual deviance is $-9607.73$). The Poisson model yields the log-likelihood value $-7296.398$ ($df = 8$), and the negative binomial model, which uses just one more parameter, reduces the likelihood to $-4803.867(df = 9)$. □

TABLE 7.6: Log-linear Poisson and quasi-Poisson models for health care data (males).

| | Poisson | | | Quasi-Poisson | |
|---|---|---|---|---|---|
| | Estimate | Std. Error | $p$-Value | Std. Error | $p$-Value |
| Intercept | 0.746 | 0.136 | 0.000 | 0.370 | 0.044 |
| hosp | 0.188 | 0.009 | 0.000 | 0.025 | 0.000 |
| healthpoor | 0.221 | 0.030 | 0.000 | 0.081 | 0.006 |
| healthexcellent | –0.229 | 0.045 | 0.000 | 0.122 | 0.060 |
| numchron | 0.153 | 0.007 | 0.000 | 0.020 | 0.000 |
| age | 0.004 | 0.017 | 0.833 | 0.047 | 0.938 |
| married[yes] | 0.132 | 0.027 | 0.000 | 0.073 | 0.072 |
| school | 0.043 | 0.003 | 0.000 | 0.008 | 0.000 |

In the preceding example overdispersion was found, which occurs quite frequently in applications. An exception is the log-linear model for the number of children in Example 7.3. When fitting a log-linear model with variance $\phi\mu_i$, one obtains the estimate $\hat{\phi} = 0.847$, which means weak underdispersion. Therefore, $p$-values for the corresponding parameter estimates are slightly smaller than the values given in Table 7.2.

### 7.5.2 Random Effects Model

One possible cause for overdispersion in the Poisson model is unobserved heterogeneity among subjects. A way of handling heterogeneity is to model it explicitly. It is assumed that the mean of observation $y_i$ is given by

$$\lambda_i = b_i \mu_i = b_i \exp(\boldsymbol{x}_i^T \boldsymbol{\beta}), \tag{7.8}$$

where $b_i$ is a subject-specific parameter, which is itself drawn from a mixing distribution. It represents the heterogeneity of the population that is not captured by the observed variables $\boldsymbol{x}_i$. The model assumption $y_i \sim \mathrm{P}(\lambda_i)$ is understood conditionally *for given* $b_i$ (and $\boldsymbol{x}_i$).

Model (7.8) may also be written in the form

$$\lambda_i = \exp(a_i + \boldsymbol{x}_i^T \boldsymbol{\beta}),$$

where $a_i = \log(b_i)$ is a random intercept within the linear predictor. With $f(b_i)$ denoting the density of $b_i$, the marginal probability of the response value $y_i$ is obtained in the usual way as

$$P(y_i) = \int f(y_i|b_i)\, f(b_i) db_i.$$

There are various ways of specifying the distribution of $b_i$ and $a_i$, respectively. Hinde (1982) assumes a normal distribution for $a_i = \log(b_i)$. Then $b_i$ follows the log-normal distribution. For the specific normal distribution $a_i \sim \mathrm{N}(-\sigma^2/2, \sigma^2)$ one obtains for $b_i$ the mean $\mathrm{E}(b_i) = 1$ and the variance $\mathrm{var}(b_i) = \exp(\sigma^2) - 1$. In particular $\mathrm{E}(b_i) = 1$ is a sensible choice for the model (7.8), where $\lambda_i$ is given by $\lambda_i = b_i \exp(\boldsymbol{x}_i^T \boldsymbol{\beta})$, since then the log-linear Poisson model is the limiting case when the variance of $b_i$ tends to zero. Dean et al. (1989) consider a random effects model by using the inverse normal distribution. An alternative choice that yields an explicit marginal distribution is based an the Gamma mixing distribution. The corresponding Poisson-Gamma model is considered in the next section.

In general, maximization of the marginal likelihood is computationally intensive because the integrals have to be approximated, for example, by Gauss-Hermite integration (see Chapter 14).

## 7.6 Negative Binomial Model and Alternatives

Quasi-likelihood methods seem to provide a sufficiently flexible tool for the estimation of overdispersed count data. The assumptions are weak; by using only the first two moments one does not have to specify a distribution function. Nevertheless, parametric models have advantages. In particular, they are useful as building blocks of mixture models as considered in Sections 7.7 and 7.8. The type of overdispersion found in these mixture models cannot be modeled within the framework of quasi-likelihood methods. Therefore, in the following we will consider alternative parametric models.

There are several distribution models that are more flexible than the Poisson model but include it as a limiting case. A frequently used model is the negative binomial distribution. In contrast to the Poisson distribution, it is a two-parameter distribution and therefore more flexible than the Poisson model; in particular, it can model overdispersed counts. In the following we first consider the negative binomial model, which can be derived as a mixture of Poisson distributions. The second extension that will be considered is the generalized Poisson distribution.

## Negative Binomial Model as Gamma-Poisson-Model

A specific choice for the mixing distribution in model (7.8), which allows a closed form of the marginal distribution, is the Gamma-distribution. The Gamma-distribution $b_i \sim \Gamma(\nu, \alpha)$ is given by the density

$$f(b_i) = \begin{cases} 0 & b_i \leq 0 \\ \frac{\alpha^\nu}{\Gamma(\nu)} b_i^{\nu-1} \, \mathrm{e}^{-\alpha b_i} & b_i > 0. \end{cases}$$

The mean and variance are $\mathrm{E}(b_i) = \nu/\alpha, \mathrm{var}(b_i) = \nu/\alpha^2$. If one assumes for the random parameter $b_i$ the Gamma-distribution $\Gamma(\nu, \nu)$, the mean fulfills $\mathrm{E}(b_i) = 1$ and one obtains for the marginal probability

$$\begin{aligned} P(y_i) &= \int f(y_i|b_i) f(b_i) d\, b_i \\ &= \int \left( \mathrm{e}^{-b_i \mu_i} \, \frac{(b_i \mu_i)^{y_i}}{y_i!} \right) \left( \frac{\nu^\nu}{\Gamma(\nu)} b_i^{\nu-1} \, \mathrm{e}^{-\nu b_i} \right) d\, b_i \\ &= \frac{\Gamma(y_i + \nu)}{\Gamma(\nu)\Gamma(y_i + 1)} \left( \frac{\mu_i}{\mu_i + \nu} \right)^{y_i} \left( \frac{\nu}{\mu_i + \nu} \right)^{\nu}. \end{aligned} \tag{7.9}$$

The density (7.9) represents the *negative binomial distribution* $\mathrm{NB}(\nu, \mu_i)$, with mean and variance given by

$$\mathrm{E}(y_i) = \mu_i = \exp(\boldsymbol{x}_i^T \boldsymbol{\beta}), \quad \mathrm{var}(y_i) = \mu_i + \mu_i^2/\nu.$$

While the mean is the same as for the simple Poisson model, the variance exceeds the Poisson variances by $\mu_i^2/\nu$. The Poisson model may be seen as a limiting case ($\nu \to \infty$). The scaling of $\nu$ is such that small values signal strong overdispersion when compared to the Poisson model, while for large values of $\nu$ the model is similar to the Poisson model. Therefore, $1/\nu$ is considered the dispersion parameter. For illustration, Figure 7.5 shows three densities of the negative binomial distribution. It is seen that NB(100,3), which is is close to the Poisson distribution, is much more concentrated around the mean than NB(5,3). For *known* $\nu$ the *negative binomial model* can be estimated within the GLM framework.

In summary, the negative binomial model was motivated by the assumptions $y_i|\lambda_i \sim \mathrm{P}(\lambda_i)$, $b_i \sim \Gamma(\nu, \nu), \lambda_i = b_i \mu_i$ and is given by

$$y_i|\boldsymbol{x}_i \sim \mathrm{NB}(\nu, \mu_i), \quad \mu_i = \exp(\boldsymbol{x}_i^T \boldsymbol{\beta}). \tag{7.10}$$

The additional dispersion parameter makes the model more flexible than the simple Poisson model. If the parameter $\nu$ is fixed, for example, by $\nu = 1$, which corresponds to the geometric distribution, the additional flexibility is lost, but nevertheless variance functions that do not postulate equidispersion are used.

Alternatively, the Gamma-Poisson model may be derived from assuming that $y_i$ is conditionally Poisson-distributed $P(\lambda_i)$ *for given* $\lambda_i$ and specifying $\lambda_i$ as a random variable that is Gamma-distributed $\Gamma(\nu_i, \frac{\nu_i}{\mu_i})$ with density function

$$f(\lambda_i) = \frac{1}{\Gamma(\nu_i)} (\frac{\nu_i}{\mu_i})^{\nu_i} \lambda_i^{\nu_i - 1} \exp(-\frac{\nu_i}{\mu_i} \lambda_i)$$

for $\lambda_i > 0$. Then one has mean $\mathrm{E}(\lambda_i) = \mu_i$ and variance $\mathrm{var}(\lambda_i) = \mu_i^2/\nu_i$. If the link between the mean and the linear predictor is specified by $\mu_i = \exp(\boldsymbol{x}_i^T \boldsymbol{\beta})$, one obtains for the conditional distribution $y_i|\lambda_i \sim \mathrm{P}(\exp(\boldsymbol{x}_i^T \boldsymbol{\beta}))$ and for the marginal distribution the discrete

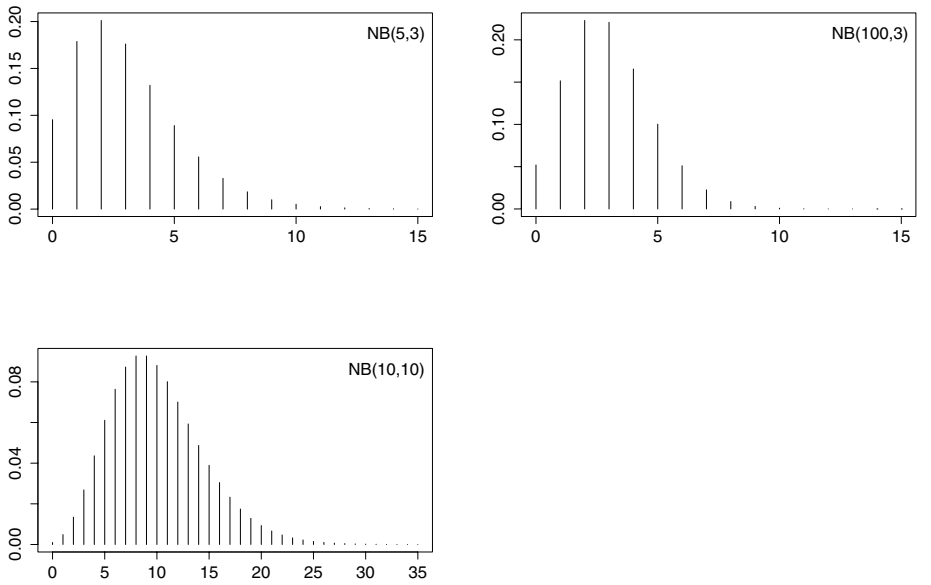


FIGURE 7.5: Probability mass functions of negative binomial distributions.

density

$$P(y_i;\nu_i,\mu_i) = \int f(y_i|\lambda_i)f(\lambda_i)d\,\lambda_i$$
$$= \frac{\Gamma(y_i+\nu_i)}{\Gamma(y_i+1)\Gamma(\nu_i)}\left(\frac{\mu_i}{\mu_i+\nu_i}\right)^{y_i}\left(\frac{\nu_i}{\mu_i+\nu_i}\right)^{\nu_i}. \qquad (7.11)$$

Density (7.11) is the negative binomial distribution function with mean $\mathrm{E}(y_i)=\mu_i$ and variance $\mathrm{var}(y_i)=\mu_i+\frac{1}{\nu_i}\mu_i^2$. Therefore, the negative binomial model (7.10) may also be motivated by the assumptions $y_i|\lambda_i \sim \mathrm{P}(\lambda_i), \lambda_i \sim \Gamma(\nu,\frac{\nu}{\mu_i}), \mu_i=\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})$.

The essential advantage of the negative binomial model over the Poisson model is that, by introducing a second parameter, more flexible variance structures are possible. However, the variance $\mathrm{var}(y_i)=\mu_i+\mu_i^2/\nu$ is also restrictive in a certain sense. Since $\nu>0$, the variance can only be larger than assumed in the Poisson model. Therefore, underdispersion cannot be modeled adequately by the negative binomial model. When underdispersion occurs, as in Example 7.3, where $\hat{\phi}=0.847$, the fitting of the negative binomial value typically yields the Poisson model with a very large value $\hat{\nu}$.

### Example 7.7: Insolvent Firms

As in Example 7.5, a log-linear link is assumed, $\log(\mu)=\beta_0+x\beta_1+x^2\beta_2$, where $x$ denotes the months ranging from 1 to 36. The fitted models are the log-linear Poisson model, the log-linear Poisson model with dispersion parameter $\phi$, the log-linear negative binomial model, and a mixture model that assumes a normal distribution of individual effects $a_i$. Since the counts are rather large, in addition a normal distribution model with log-link has been fitted. It is seen from Table 7.7 that the quadratic effect seems negligible for all models. The Poisson model is certainly not the best choice since the data are

overdispersed. The estimate $\hat{\phi} = 2.313$ signals that the variance is about twice what one would expect for the Poisson model. A check for overdispersion is shown in Figure 7.4 (right panel), where the quadratic residuals are plotted against the fitted values. The straight line shows what is to be expected when the Poisson model holds, namely, $\mathrm{E}(y_i - \mu_i)^2 = \mu_i$. The smooth fit shows that squared residual tends to be much larger than expected. Overdispersion is also seen from the fit of the negative binomial model. When compared to the Poisson model, the variance increases by $\mu_i^2/\nu$. For values $\mu_i \in [70, 100]$ and $\hat{\nu} = 77.93$, that means a substantial increase in variance between 63 and 128. The estimated effects for the models compare well with the exception of the Poisson model, for which the standard errors are definitely too small. AIC for the Poisson model was 306.82; for the normal distribution model, which is more flexible, one gets the smaller value 296.54; the smallest value, 296.27, is obtained by the negative binomial model. □

TABLE 7.7: Log-linear model for insolvencies (standard errors in brackets).

| | Log-Linear Poisson Model $\phi = 1$ | Log-Linear Dispersion Poisson Model | Negative Binomial Model | Log-Linear Normal Distribution Model | Mixture Gauss-Hermite |
|---|---|---|---|---|---|
| Variance | $\mathrm{var}(y_i) = \mu_i$ | $\mathrm{var}(y_i) = \phi\mu_i$ | $\mathrm{var}(y_i) = \mu_i + \frac{\mu_i^2}{\nu}$ | $\sigma^2$ | |
| Intercept | 4.192 (0.062) | 4.192 (0.094) | 4.195 (0.086) | 4.184 (0.101) | 4.186 (0.086) |
| Month | 0.020 (0.007) | 0.020 (0.0112) | 0.019 (0.011) | 0.021 (0.01?) | 0.0196 (0.0105) |
| Month$^2$ | –0.00026 (0.00019) | –0.00026 (0.00028) | –0.00025 (0.00027) | –0.00029 (0.00029) | –0.00026 (0.00027) |
| Dispersion | — | $\hat{\phi} = 2.313$ | $\hat{\nu} = 77.93$ (35.49) | $\hat{\sigma} = 13.90$ | $\hat{\sigma}$ =0.113 (0.025) |

### Example 7.8: Demand for Medical Care

Table 7.8 shows the estimates of the negative binomial model for the medical care data (Example 7.6). It is seen that the effects are similar to the effects of a quasi-Poisson model, but the standard errors are slightly smaller ($\hat{\nu} = 1.079$, with standard error 0.048, and the residual deviance is $-9607.73$). The Poisson model yields the log-likelihood value $-7296.398$ ($df = 8$), and the negative binomial model, which uses just one more parameter, reduces the likelihood to $-4803.867$ ($df = 9$). □

TABLE 7.8: Negative binomial model for health care data (males).

| | Estimate | Std. Error | $p$-Value |
|---|---|---|---|
| Intercept | 0.556 | 0.333 | 0.094 |
| hosp | 0.245 | 0.033 | 0.000 |
| healthpoor | 0.255 | 0.083 | 0.002 |
| healthexcellent | –0.206 | 0.096 | 0.032 |
| numchron | 0.182 | 0.020 | 0.000 |
| age | 0.021 | 0.042 | 0.622 |
| married[yes] | 0.148 | 0.063 | 0.020 |
| school | 0.040 | 0.007 | 0.000 |

### Generalized Poisson Distribution

An alternative distribution that allows for overdispersion is the generalized Poisson distribution, which was investigated in detail in Consul (1998). A random variable $Y$ follows a generalized Poisson distribution with parameters $\mu > 0$ and $\gamma$, $Y \sim GP(\mu, \gamma)$, if the density is given by

$$P(Y = y) = \begin{cases} \dfrac{\mu(\mu + y(\gamma - 1))^{y-1}\gamma^{-y}e^{-(\mu+y(\gamma-1))/\gamma}}{y!} & \text{for } y \in \{0, 1, 2, \dots\} \\ 0 & \text{for } y > m, \quad \text{if } \gamma < 1. \end{cases}$$

Additional constraints on the parameters are $\gamma \geq \max\{1/2, 1 - \mu/m\}$, where $m \geq 4$ is the largest natural number such that $\mu + m(\gamma - 1) > 0$ if $\gamma < 1$.

It is seen that the distribution becomes the Poisson distribution for $\gamma = 1$. For small values of $\gamma$ the generalized Poisson distribution is very similar to the negative binomial distribution; for large values the negative binomial distribution puts more mass on small values of $y$. For the generalized Poisson distribution one obtains

$$\mathrm{E}(Y) = \mu \quad \mathrm{var}(Y) = \gamma^2\mu.$$

The parameter $\gamma^2$ can be seen as a dispersion parameter; for $\gamma^2 > 1$ one obtains greater dispersion than for the Poisson model, and for $\gamma^2 < 1$ one obtains underdispersion. An advantage of the generalized Poisson distribution is that the dispersion parameter also allows for underdispersion in contrast to the negative binomial model, for which the variance is $\mathrm{var}(Y) = \mu + \mu^2/\nu$. Like the negative binomial model, the generalized Poisson distribution can be derived as a mixture of Poisson distributions (Joe and Zhu, 2005). Gschoessl and Czado (2006) fitted a regression model based on the generalized Poisson distribution and compared several models for overdispersion from a Bayesian perspective.

## 7.7 Zero-Inflated Counts

In many applications one observes more zero counts than is consistent with the Poisson (or an alternative count data) model; the data display overdispersion through excess zeros. Often one may think of data as resulting from a mixture of distributions. If a person is asked, "How many times did you eat mussels in the past 3 months?" one records zero responses from people who never eat mussels and from those who do but happen not to have done so during the time interval in question.

In general, a zero-inflated count model may be motivated from a mixture of two subpopulations, the non-responders who are "never at risk" and the responders who are at risk. With $C$ denoting the class indicator of subpopulations ($C_i = 1$ for responders and $C_i = 0$ for non-responders) one obtains the mixture distribution

$$P(y_i = y) = P(y_i = y|C_i = 1)\pi_i + P(y_i = y|C_i = 0)(1 - \pi_i),$$

where $\pi_i = P(C_i = 1)$ are the mixing probabilities. When one assumes that counts within the responder subpopulation are Poisson-distributed, one obtains with $P(y_i = 0|C_i = 0) = 1$

$$P(y_i = 0) = P(y_i = 0|C_i = 1)\pi_i + (1 - \pi_i) = \pi_i\,\mathrm{e}^{-\mu_i} + 1 - \pi_i,$$

and for $y > 0$

$$P(y_i = y) = P(y_i = y|C_i = 1)\pi_i = \pi_i\,\mathrm{e}^{-\mu_i}\,\mu_i^y/y!,$$

where $\mu_i$ is the mean of the Poisson distribution of population $C_i = 1$. One obtains

$$\mathrm{E}(y_i) = \pi_i\mu_i, \quad \mathrm{var}(y_i) = \pi_i\mu_i + \pi_i(1 - \pi_i)\mu_i^2 = \pi_i\mu_i(1 + \mu_i(1 - \pi_i))$$

(Exercise 7.4). Since $\text{var}(y_i) > \text{E}(y_i)$, excess zeros imply overdispersion if $\pi_i < 1$. Of course, the Poisson model is included as the special case where all observations refer to responders and $\pi_i = 1$.

When covariates are present one may specify a Poisson distribution model for $y|C_i = 1$ and a binary response model for $C_i \in \{0, 1\}$, for example,

$$\log(\mu_i) = \boldsymbol{x}_i^T\boldsymbol{\beta}, \quad \text{logit}(\pi_i) = \boldsymbol{z}_i^T\boldsymbol{\gamma},$$

where $\boldsymbol{x}_i, \boldsymbol{z}_i$ may be different sets of covariates. The simplest mixture model assumes only an intercept in the binary model, $\text{logit}(\pi_i) = \gamma_0$. For increasing $\gamma_0$ one obtains in the limit the Poisson model without zero inflation.

The joint log-likelihood function after omitting constants is given by

$$\begin{aligned} l &= \sum_{i=1}^{n} l_i(y_i) \\ &= \sum_{i=1}^{n} I(y_i = 0) \log\{1 + \frac{\exp(\boldsymbol{z}_i^T\boldsymbol{\gamma})}{1 + \exp(\boldsymbol{z}_i^T\boldsymbol{\gamma})}(\exp(-\exp(\boldsymbol{x}_i^T\boldsymbol{\beta}) - 1))\} \\ &+ (1 - I(y_i = 0))\{\boldsymbol{z}_i^T\boldsymbol{\gamma} - \log(1 + \exp(\boldsymbol{z}_i^T\boldsymbol{\gamma})) - \exp(\boldsymbol{x}_i^T\boldsymbol{\beta}) + y_i(\boldsymbol{x}_i^T\boldsymbol{\beta})\}, \end{aligned}$$

where $I(y_i = 0)$ denotes an indicator variable that takes value 1 if $y_i = 0$, and 0 otherwise. Lambert (1992) suggested the use of the EM algorithm to maximize the log-likelihood. Zeileis et al. (2008) obtain ML estimates by using optimization functions from R and allow to specify starting values estimated by the EM algorithm. They compute the covariance matrix as the numerically determined Hessian matrix.

The zero-inflated Poisson model has been extended to the zero-inflated generalized Poisson model in which the Poisson distribution is replaced by the generalized Poisson distribution (Famoye and Singh, 2003; Famoye and Singh, 2006; Gupta et al., 2004; Czado et al., 2007). The resulting family of models is rather large, comprising a zero-inflated Poisson regression and a generalized Poisson regression. Min and Czado (2010) discussed the use of the Wald and likelihood ratio tests for the investigation of zero inflation (or zero deflation).

### Example 7.9: Demand for Medical Care

Table 7.9 shows the fit of a zero-inflated model for the health care data. The counts are modeled as a log-linear Poisson model, with all variables included. The binary response uses the logit link with intercept only. Therefore, it is assumed that all probabilities $\pi_i$ are equal. The estimate $-1.522$ corresponds to a probability of $0.295$, which means that the portion of responders is not very large and overdispersion has to be expected. That is in agreement with the fitted quasi-Poisson model (Table 7.6). Table 7.10 shows the fit of the zero-inflated model when all the variables can have an effect on the mixture component. With the exception of the health status, all variables seem to contribute to the inflation component. For the log-likelihood of the fitted models one obtains $-6544$ on 9 *df* (zero-inflated Poisson model with an intercept for the logit model) and $-6455$ on 16 *df* (zero-inflated Poisson model with all variables in both components). Compared to the Poisson model with log-likelihood $-7296.398$ ($df = 8$), already the zero-inflated model with just an intercept in the binary component is a distinct improvement. It is noteworthy that the results differ with respect to significance. The predictor married is not significant when a quasi-Poisson or a negative binomial model is fitted but seems not neglectable when a zero-inflated model is fitted. □

TABLE 7.9: Zero-inflated models, Poisson and logit, for health care data (males).

| Count Model Coefficients (Poisson with Log Link) | | | | |
|---|---|---|---|---|
| | Estimate | Std. Error | $z$-Value | $p$-Value |
| Intercept | 1.639 | 0.140 | 11.678 | 0.0 |
| hosp | 0.165 | 0.009 | 17.477 | 0.0 |
| healthpoor | 0.241 | 0.030 | 8.078 | 0.0 |
| healthexcellent | −0.176 | 0.047 | −3.771 | 0.0 |
| numchron | 0.103 | 0.008 | 13.527 | 0.0 |
| age | −0.048 | 0.018 | −2.721 | 0.007 |
| married[yes] | 0.009 | 0.027 | 0.342 | 0.732 |
| school | 0.029 | 0.003 | 10.134 | 0.0 |
| **Zero-Inflation Model Coefficients (Binomial with Logit Link)** | | | | |
| | Estimate | Std. Error | $z$-Value | $p$-Value |
| Intercept | -1.522 | 0.063 | -24.10 | 0.0 |

TABLE 7.10: Zero-inflated models, Poisson and logit, for health care data (males).

| Count Model Coefficients (Poisson with Log Link) | | | | |
|---|---|---|---|---|
| | Estimate | Std. Error | $z$-Value | $p$-Value |
| (Intercept) | 1.672 | 0.140 | 11.980 | 0.0 |
| hosp | 0.165 | 0.009 | 17.451 | 0.0 |
| healthpoor | 0.240 | 0.030 | 8.057 | 0.0 |
| healthexcellent | −0.163 | 0.045 | −3.587 | 0.0003 |
| numchron | 0.101 | 0.008 | 13.351 | 0.0 |
| age | −0.050 | 0.017 | −2.839 | 0.0045 |
| married[yes] | 0.005 | 0.027 | 0.168 | 0.8663 |
| school | 0.028 | 0.003 | 9.919 | 0.0 |
| **Zero-Inflation Model Coefficients (Binomial with Logit Link)** | | | | |
| | Estimate | Std. Error | $z$-Value | $p$-Value |
| (Intercept) | 3.152 | 0.890 | 3.541 | 0.0004 |
| hosp | −0.604 | 0.156 | −3.869 | 0.0001 |
| healthpoor | 0.214 | 0.245 | 0.874 | 0.3822 |
| healthexcellent | 0.260 | 0.213 | 1.221 | 0.2221 |
| numchron | −0.477 | 0.065 | −7.305 | 0.0 |
| age | −0.348 | 0.115 | −3.042 | 0.0024 |
| married[yes] | −0.700 | 0.148 | −4.745 | 0.0 |
| school | −0.092 | 0.017 | −5.505 | 0.0 |

## 7.8 Hurdle Models

An alternative model that is able to account for excess zeros is the hurdle models (Mullahy, 1986; Creel and Loomis, 1990). It allows one to model overdispersion through excess zeros for baseline models such as the Poisson model and the negative binomial model. The model specifies two processes that generate the zeros and the positives. The combination of both models, a binary model that determines whether the outcome is zero or positive and a truncated-at-zero count model, gives the model.

In general, one assumes that $f_1, f_2$ are the probability mass functions with support $\{0, 1, 2, \dots\}$. The hurdle model is given by

$$\begin{aligned} P(y = 0) &= f_1(0), \\ P(y = r) &= f_2(r)\frac{1 - f_1(0)}{1 - f_2(0)}, \quad r = 1, 2, \dots . \end{aligned}$$

The model may be seen as a stage-wise decision model. At the first stage a binary variable $C$ determines whether a count variable has a zero or a positive outcome. $C = 1$ means that the "hurdle is crossed" and the outcome is positive, while $C = 0$ means that zero will be observed. The binary decision between zero and a positive outcome is determined by the $f_1$-distribution in the form

$$P(C = 1) = 1 - f_1(0), \quad P(C = 0) = f_1(0).$$

At the second stage the condition distribution given $C$ is specified. If the hurdle is crossed, the response is determined by the truncated count model with probability mass function

$$P(y = r|C = 1) = f_2(r)/(1 - f_2(0)) \qquad r = 1, 2, \dots$$

If the hurdle is not crossed, the probability for zero outcome is 1, $P(y = 0|C = 0) = 1$. One obtains the hurdle model from $P(y = r) = P(y = r|C = 0)P(C = 0) + P(y = r|C = 1)P(C = 1)$ , which yields

$$\begin{aligned} P(y = 0) &= P(C = 0) = f_1(0) \\ P(y = r) &= P(y = r|C = 1)P(C = 1) \\ &= \{f_2(r)/(1 - f_2(0))\}(1 - f_1(0)), \ r = 1, 2, \dots . \end{aligned}$$

The derivation shows that the hurdle model is a finite mixture of the truncated count model $P(y = r|C = 1)$ and the degenerate distribution $P(y = r|C = 0)$. In contrast to the zero inflated-counts models from Section 7.7, $C$ is an observed variable and not an unobservable mixture. The truncated count model is determined by the probability mass function $f_2$, which has been called the *parent process* by Mullahy (1986). If $f_1 = f_2$, the model collapses to the parent model $f_2$.

The model is quite flexible and allows for both under- and overdispersion. This is seen by considering the mean and variance. With $\gamma = (1-f_1(0))/(1-f_2(0)) = P(y > 0)/(1-f_2(0))$, the mean is given by

$$\mathrm{E}(y) = \sum_{r=1}^{\infty} r \ f_2(r)\gamma = P(y > 0)\,\mathrm{E}(y|y > 0)$$

and the variance has the form

$$\mathrm{var}(y) = P(y > 0)\,\mathrm{var}(y|y > 0) + P(y > 0)(1 - P(y > 0))\,\mathrm{E}(y|y > 0)^2.$$

Let us consider as a specific model, the *hurdle Poisson model*, which assumes that $f_2$ is the probability mass function of a Poisson distribution with mean $\mu_2$. Let $y_2$ denote the corresponding random variable (Poisson distribution with mean $\mu_2$). Then one has $\mathrm{E}(y_2) = \mu_2$ and $\mu_2 = \mathrm{var}(y_2) = \mathrm{E}(y_2^2) - \mathrm{E}(y_2)^2$, yielding $\mathrm{E}(y_2^2) = \mu_2 + \mu_2^2 = \mu_2(1 + \mu_2)$. One obtains for the mean and variance of $y$

$$\begin{aligned} \mathrm{E}(y) &= \gamma\mu_2, \\ \mathrm{var}(y) &= \sum_{r=1}^{\infty} r^2 f_2(r)\gamma - (\sum_{r=1}^{\infty} r\, f_2(r)\gamma)^2 = \mu_2(1 + \mu_2)\gamma - \mu_2^2\gamma^2, \end{aligned}$$

and therefore

$$\frac{\text{var}(y)}{\text{E}(y)} = 1 + \mu_2(1-\gamma).$$

This means that for the non-trivial case $\mu_2 > 0$ one obtains *overdispersion* if $0 < \gamma < 1$ and *underdispersion* if $1 < \gamma < (1+\mu_2)/\mu_2$, where the upper threshold is determined by the restriction $\text{var}(y) > 0$. For $\gamma = 1$, the hurdle Poisson becomes the Poisson model.

The hurdle model is determined by the choices of $f_1$ and $f_2$. There is much flexibility because $f_1$ and $f_2$ may be Poisson, geometric, or negative binomial distributions. Moreover, the distributions do not have to be the same. One can also combine a binary logit model for the truncated (right-censored at $y = 1$) distribution of $f_1$ and a Poisson or negative binomial model for $f_2$.

Concrete parameterizations are obtained by linking the two distributions to explanatory variables. For illustration we consider the hurdle Poisson model where both $f_1$ and $f_2$ correspond to Poisson distributions with means $\mu_1$ and $\mu_2$, respectively. For observations $(y_i, \boldsymbol{x}_i)$ one may specify for

$$\mu_{i1} = \exp(\boldsymbol{x}_i^T \boldsymbol{\beta}_1), \quad \mu_{i2} = \exp(\boldsymbol{x}_i^T \boldsymbol{\beta}_2),$$

yielding the model

$$\begin{aligned} P(y_i = 0) &= \exp(-\mu_{i1}), \\ P(y_i = r) &= \frac{\mu_{i2}^r}{r!}\, \text{e}^{-\mu_i 2}\, \frac{1-\exp(-\mu_{i1})}{1-\exp(-\mu_{i2})}. \end{aligned}$$

The log-likelihood is given by

$$l(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2) = -\sum_{y_i=0} \mu_{i1} + \sum_{y_i>0} \log\left(\frac{1-\text{e}^{-\mu_{i1}}}{1-\text{e}^{-\mu_{i2}}}\, \frac{\mu_{i2}^{y_i}}{y_i!}\, \text{e}^{-\mu_{i2}}\right),$$

which decomposes into $l(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2) = l_1(\boldsymbol{\beta}_1) + l_2(\boldsymbol{\beta}_2)$ with

$$\begin{aligned} l_1(\boldsymbol{\beta}_1) &= -\sum_{y_i=0} \mu_{i1} + \sum_{y_i>0} \log(1-\text{e}^{-\mu_{i1}}), \\ l_2(\boldsymbol{\beta}_2) &= \sum_{y_i>0} y_i \log(\mu_{i2}) - \mu_{i2} - \log(1-\text{e}^{-\mu_{i2}}) - \log(y_i!). \end{aligned}$$

Since the components depend on a one-parameter vector, only the two components can be maximized separately. In general, the regressors for the two model components do not have to be the same. But, if the same regressors as well as the same count models are used, as in the preceding Poisson example, a test of the hypothesis $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_2$ tests whether the hurdle is needed. Although most hurdle models use the hurdle zero, the specification of more general models where the hurdle is some positive number is straightforward.

### Example 7.10: Demand for Medical Care

Table 7.11 and Table 7.12 show the fits of hurdle models for the health care data. The counts are modeled as a log-linear Poisson model, with all variables included, and the binary response uses the logit link with intercept only (Table 7.11) or with all the covariates included (Table 7.12). For the log-likelihood of the fitted models one obtains $-6549$ on 9 *df* (hurdle model with an intercept for the logit model) and $-6456$ on 16 *df* (hurdle model with all variables in both components). Both models have a much better fit than the simple Poisson model ($-7296.398$ on $df = 8$). Moreover, the model with all covariates in the binary component is to be preferred over the pure intercept model. A comparison of the hurdle model and the

TABLE 7.11: Hurdle model, Poisson and logit, for health care data (males).

| Count Model Coefficients (Truncated Poisson with Log Link) | | | | |
|---|---|---|---|---|
| | Estimate | Std. Error | $z$-Value | $p$-Value |
| Intercept | 1.673 | 0.139 | 12.001 | 0.0 |
| hosp | 0.165 | 0.009 | 17.450 | 0.0 |
| healthpoor | 0.240 | 0.030 | 8.062 | 0.0 |
| healthexcellent | −0.164 | 0.046 | −3.592 | 0.0 |
| numchron | 0.101 | 0.008 | 13.346 | 0.0 |
| age | −0.050 | 0.017 | −2.848 | 0.0 |
| marriedyes | 0.005 | 0.027 | 0.170 | 0.86 |
| school | 0.028 | 0.003 | 9.920 | 0.0 |
| **Zero Hurdle Model Coefficients (Binomial with Logit Link)** | | | | |
| | Estimate | Std. Error | $z$-Value | $p$-Value |
| Intercept | 1.483 | 0.061 | 24.28 | 0.0 |

TABLE 7.12: Hurdle model, Poisson and logit, for health care data (males).

| Count Model Coefficients (Truncated Poisson with Log Link) | | | | |
|---|---|---|---|---|
| | Estimate | Std. Error | $z$-Value | $p$-Value |
| (Intercept) | 1.673 | 0.139 | 12.001 | 0.0 |
| hosp | 0.165 | 0.009 | 17.450 | 0.0 |
| healthpoor | 0.240 | 0.030 | 8.062 | 0.0 |
| healthexcellent | −0.164 | 0.046 | −3.592 | 0.0003 |
| numchron | 0.101 | 0.008 | 13.346 | 0.0 |
| age | −0.050 | 0.017 | −2.848 | 0.0043 |
| marriedyes | 0.005 | 0.027 | 0.170 | 0.8652 |
| school | 0.028 | 0.003 | 9.920 | 0.0 |
| **Zero Hurdle Model Coefficients (Binomial with Logit Link):** | | | | |
| | Estimate | Std. Error | $z$-Value | $p$-Value |
| (Intercept) | −3.104 | 0.871 | −3.564 | 0.0004 |
| hosp | 0.611 | 0.155 | 3.944 | 0.0 |
| healthpoor | −0.199 | 0.244 | −0.815 | 0.4151 |
| healthexcellent | −0.274 | 0.208 | −1.318 | 0.1876 |
| numchron | 0.482 | 0.064 | 7.476 | 0.0 |
| age | 0.336 | 0.118 | 3.010 | 0.0026 |
| married[yes] | 0.690 | 0.146 | 4.743 | 0.0 |
| school | 0.094 | 0.016 | 5.711 | 0.0 |

zero-inflated model shows that the parameter estimates and $p$-values are comparable; only the signs for the parameters of the mixture have changed since the response $y = 0$ is modeled in the hurdle models. □

## 7.9 Further Reading

*Surveys and Books*. A source book for the modeling of count data that includes many applications is Cameron and Trivedi (1998). An econometric view on count data is outlined in Winkelmann (1997) and Kleiber and Zeileis (2008). The negative binomial model is treated extensively in Hilbe (2011)

*Tests on Zero Inflation*. Tests that investigate the need for zero inflation have been suggested for the case of constant overdispersion. The most widely used test is the score test, because it

requires only the fit under the null model; see van den Broek (1995), Deng and Paul (2005), and Gupta et al. (2004).

*Hurdle Models*. The Poisson hurdle and the geometric and hurdle have been examined by Mullahy (1986), and hurdle negative binomial models have been considered by Pohlmeier and Ulrich (1995). Zeileis et al. (2008) describe how regression models for count data, including zero-inflated and hurdle models, can be fitted in R.

*R Packages*. GLMs as the Poisson model can be fitted by using of the model fitting functions *glm* from the *MASS* package. Many tools for diagnostic and inference are available. *MASS* also allows one to fit negative binomial models with fixed dispersion parameters (function *negative.binomial*) and for estimating regression parameters and dispersion parameters (function *glm.nb*). Estimation procedures for zero-inflated and hurdle models are available in the *pscl* package (for details see Zeileis et al., 2008).

## 7.10 Exercises

7.1 In Example 1.5, the dependent variable is the number of insolvent firms depending on year and month (see Table 1.3).

(a) Consider time as the only covariate ranging from 1 to 36. Fit a log-linear Poisson model, an overdispersed model, a negative binomial model, and a log-linear normal distribution model with linear time (compare to Example 7.7).

(b) Fit the models from part (a) with the linear predictor determined by the factors year and month and an interaction effect if needed.

(c) Discuss the difference between the models fitted in parts (a) and (b).

TABLE 7.13: Cellular differentiation data from Piegorsch et al. (1988).

| Number of Cells Differentiating | Dose of TNF(U/ml) | Dose of IFN(U/ml) |
|---|---|---|
| 11 | 0 | 0 |
| 18 | 0 | 4 |
| 20 | 0 | 20 |
| 39 | 0 | 100 |
| 22 | 1 | 0 |
| 38 | 1 | 4 |
| 52 | 1 | 20 |
| 69 | 1 | 100 |
| 31 | 10 | 0 |
| 68 | 10 | 4 |
| 69 | 10 | 20 |
| 128 | 10 | 100 |
| 102 | 100 | 0 |
| 171 | 100 | 4 |
| 180 | 100 | 20 |
| 193 | 100 | 100 |

7.2 The R package *pscl* provides the dataset *bioChemists*.

(a) Use descriptive tools to learn about the data.

(b) Fit a zero-inflated Poisson model and a hurdle model by using the R package *pscl*.

7.3 Investigate the effect of explanatory variables on the number of children for men in analogy to Example 7.3 by using the dataset *children* from the the package *catdat*.

7.4 For the zero-inflated count model a mixture of two subnopulations is assumed, with $C$ denoting the class indicator ($C_i = 1$, for responders and $C_i = 0$ for non-responders). When one assumes a Poisson model if $C_i = 1$, one has $P(y_i = 0) = \pi_i \, \mathrm{e}^{-\mu_i} + 1 - \pi_i$, $P(y_i = y) = \pi_i \, \mathrm{e}^{-\mu_i} \, \mu_i^y / y!$. Show that mean and variance have the form $\mathrm{E}(y_i) = \pi_i \mu_i$, $\mathrm{var}(y_i) = \pi_i \mu_i (1 + \mu_i (1 - \pi_i))$.

7.5 Table 7.13, which is reproduced from Piegorsch et al. (1988), shows data from a biomedical study of the immuno-activating ability of two agents, TNF (tumor necrosis factor) and IFN (interferon). Both agents induce cell differentiation. The number of cells that exhibited markers of differentiation after exposure to TNF and/or IFN was recorded. At each of the 16 dose combinations of TNF/INF, 200 cells were examined. It is of particular interest to investigate if the two agents stimulate cell differentiation synergistically or independently.

(a) Fit a log-linear Poisson model that includes an interaction term and investigate the effects.

(b) Use diagnostic tools to investigate the model fit.

(b) Fit alternative log-linear models that allow for overdispersion and compare the results to the Poisson model.

# Chapter 8

# Multinomial Response Models

In many regression problems the response is restricted to a fixed set of possible values, the so-called response categories. Response variables of this type are called *polytomous* or *multicategory* responses. In economical applications, the response categories may refer to the choice of different brands or to the choice of the transport mode (Example 1.3). In medical applications, the response categories may represent different side effects of medical treatment or several types of infection that may follow an operation. Most rating scales have fixed response categories that measure, for example, the medical condition after some treatment in categories like good, fair, and poor or the severeness of symptoms in categories like none, mild, moderate, marked. These examples show that there are at least two cases to be distinguished, namely, the case where response categories are mere labels that have no inherent ordering and the case where categories are ordered. In the first case, the response $Y$ is measured on a *nominal scale*. Instead of using the numbers $1, \ldots, k$ for the response categories, any set of $k$ numbers would do. In the latter case, the response is measured on an *ordinal scale*, where the ordering of the categories and the corresponding numbers may be interpreted but not the distance or spacing between categories. Figures 8.1 and 8.2 illustrate different scalings of response categories. In the nominal case the response categories are given in an unsystematic way, while in the ordinal case the response categories are given on a straight line, thus illustrating the ordering of the categories.

Another type of response category that contains more structure than the nominal case but is not captured by simple ordering occurs in the form of nested or hierarchical response categories. Figure 8.3 shows an example where the basic response is in the categories "no infection," "infection type I", and "infection type II." However, for infection type I two cases have to be distinguished, namely, infection with and without additional complications. Thus, one has splits on two levels, first the split into basic categories and then the conditional split within outcome "infection type I."

In this chapter we will consider the modeling of responses with unordered categories. Modeling of ordered response categories is treated in Chapter 9. In the following some examples are given.

### Example 8.1: Preference for Political Parties

Table 8.1 shows counts from a survey on the preference for political parties. The four German parties were the Christian Democratic Union (CDU), the Social Democratic Party (SPD), the Green Party, and the Liberal Party (FDP). The covariates are gender and age in categories. □

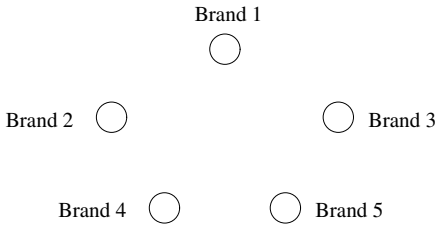


FIGURE 8.1: Choice of brand as nominal response categories.

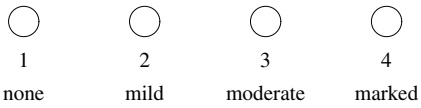


FIGURE 8.2: Severness of symptoms as ordered categories.

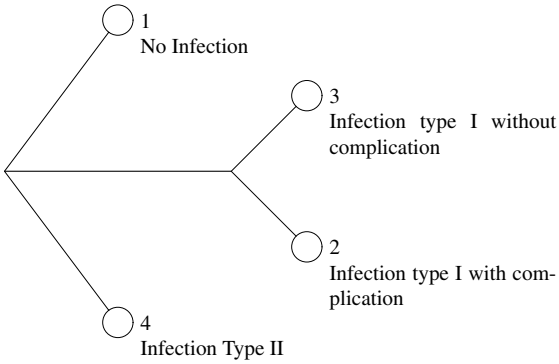


FIGURE 8.3: Type of infection as nested structure.

TABLE 8.1: Cross-classification of preference for political parties and gender.

| | | Preferred Party | | | |
|---|---|---|---|---|---|
| Gender | Age | CDU/CSU | SPD | Green Party | FDP |
| male | 1 | 114 | 224 | 53 | 10 |
| | 2 | 134 | 226 | 42 | 9 |
| | 3 | 114 | 174 | 23 | 8 |
| | 4 | 339 | 414 | 13 | 30 |
| female | 1 | 42 | 161 | 44 | 5 |
| | 2 | 88 | 171 | 60 | 10 |
| | 3 | 90 | 168 | 31 | 8 |
| | 4 | 413 | 375 | 14 | 23 |

Example 8.2: Addiction

In a survey people were asked, "Is addiction a disease or are addicts weak-willed ?" The response was in three categories, "addicts are weak-willed," "addiction is a disease," or both alternatives hold. One wants to investigate how the response depends on predictors like gender, age, and education level. The dataset is available at http://www.stat.uni-muenchen.de/service/datenarchiv/sucht/sucht.html. □

## 8.1 The Multinomial Distribution

The multinomial distribution is a natural generalization of the binomial distribution. It allows for more than two possible outcomes. For example, in a sample survey respondents might be asked for their preference for political parties. Then the number of outcomes will depend on the number of competing parties.

Let the possible outcomes be denoted by $1, \ldots, k$, which occur with probabilities $\pi_1, \ldots, \pi_k$. For the random variable $Y$, which takes values $1, \ldots, k$, one has the simple relationship $P(Y = r) = \pi_r$. However, the categories of the random variable $Y$ hide that the response is genuinely multivariate, since each response category refers to a dimension of its own. A more appropriate representation is by a vector-valued random variable. In the general form of the multinomial distribution one usually considers a sample of, say, $m$ responses. Then the components of the vector $\boldsymbol{y}^T = (y_1, \ldots, y_k)$ give the cell counts in categories $1, \ldots, k$. The vector $\boldsymbol{y}^T = (y_1, \ldots, y_k)$ has probability mass function

$$f(y_1, \ldots, y_k) = \begin{cases} \frac{m!}{y_1! \cdots y_k!} \pi_1^{y_1} \ldots \pi_k^{y_k} & y_i \in \{0, \ldots, m\},\ \Sigma_i y_i = m \\ 0 & \text{otherwise.} \end{cases}$$

A response (vector) with this probability mass function follows *a multinomial distribution* with parameters $m$ and $\boldsymbol{\pi}^T = (\pi_1, \ldots, \pi_k)$. Of course the probabilities are restricted by $\pi_i \in [0, 1], \Sigma_i \pi_i = 1$.

Since $\Sigma_i y_i = m$ there is some redundancy in the representation. Therefore, one often uses for the representation the shorter vector $\boldsymbol{y}^T = (y_1, \ldots, y_q)$, $q = k-1$, obtaining for the relevant part of the mass function

$$f(y_1, \ldots, y_q) = \frac{m!}{y_1! \ldots y_q! (m - y_1 \cdots - y_q)!} \pi_1^{y_1} \ldots \pi_q^{y_q} \cdot (1 - \pi_1 - \cdots - \pi_q)^{m - y_1 - \ldots - y_q}.$$

In the following the abbreviation $\boldsymbol{y} \sim \mathrm{M}(m, \boldsymbol{\pi})$ for the multinomial distribution will always refer to the latter version with $q = k - 1$ components. In this representation it also becomes obvious that the binomial distribution is a special case of the multinomial distribution where $k = 2$ $(q = 1)$, since

$$f(y_1) = \frac{m!}{y_1!(m - y_1)!} \pi_1^{y_1} (1 - \pi_1)^{m - y_1} = \binom{m}{y_1} \pi_1^{y_1} (1 - \pi_1)^{m - y_1}.$$

For the components of the multinomial distribution $\boldsymbol{y}^T = (y_1, \ldots y_q)$ one derives

$$\mathrm{E}(y_i) = m\pi_i, \ \ \mathrm{var}(y_i) = m\pi_i(1 - \pi_i), \ \ \mathrm{cov}(y_i, y_j) = -m\pi_i\pi_j$$

(Exercise 8.1). In vector form, the covariance is given as $\mathrm{cov}(\boldsymbol{y}) = m(\mathrm{diag}(\boldsymbol{\pi}) - \boldsymbol{\pi}\boldsymbol{\pi}^T)$, where $\boldsymbol{\pi}^T = (\pi_1, \ldots, \pi_q)$ and $\mathrm{diag}(\boldsymbol{\pi})$ is a diagonal matrix with entries $\pi_1, \ldots, \pi_q$.

For *single* observations, where only one respondent $(m = 1)$ is considered, one obtains $Y = r \Leftrightarrow y_r = 1$, with probabilities given by $\pi_r = P(Y = r) = P(y_r = 1)$ and possible outcome vectors of length $k - 1$ given by $(1, 0, \dots), (0, 1, 0, \dots) \dots (0, 0, \dots, 1)$.

The *scaled multinomial* distribution uses the vector of relative frequencies $\bar{\boldsymbol{y}} = (y_1/m, \dots, y_q/m) = \boldsymbol{y}^T/m$. It has mean $\mathrm{E}(\bar{\boldsymbol{y}}) = \boldsymbol{\pi}$ and covariance matrix $\mathrm{cov}(\bar{\boldsymbol{y}}) = (\mathrm{diag}(\boldsymbol{\pi}) - \boldsymbol{\pi}\boldsymbol{\pi}^T)/m$.

## 8.2 The Multinomial Logit Model

The multinomial logit model is the most widely used regression model that links a categorical response variable with unordered categories to explanatory variables. Again let $Y \in \{1, \dots, k\}$ denote the response in categories $1, \dots, k$ and $\boldsymbol{y}^T = (y_1, \dots, y_k)$ the corresponding multinomial distribution (for $m = 1$). Let $\boldsymbol{x}$ be a vector of explanatory variables. The binary logit model (Chapter 2) has the form

$$P(Y = 1|\boldsymbol{x}) = \frac{\exp(\boldsymbol{x}^T\boldsymbol{\beta})}{1 + \exp(\boldsymbol{x}^T\boldsymbol{\beta})}$$

or, equivalently,

$$\log\left(\frac{P(Y = 1|\boldsymbol{x})}{P(Y = 2|\boldsymbol{x})}\right) = \boldsymbol{x}^T\boldsymbol{\beta}.$$

The multinomial logit model uses the same linear form of logits. But instead of only one logit, one has to consider $k - 1$ logits. One may specify

$$\log\left(\frac{P(Y = r|\boldsymbol{x})}{P(Y = k|\boldsymbol{x})}\right) = \boldsymbol{x}^T\boldsymbol{\beta}_r, \quad r = 1, \dots, q, \tag{8.1}$$

where the log-odds compare $P(Y = r|\boldsymbol{x})$ to the probability $P(Y = k|\boldsymbol{x})$. In this presentation $k$ serves as the reference category since all probabilities are compared to the last category. It should be noted that the vector $\boldsymbol{\beta}_r$ depends on $r$ because comparison of $Y = r$ to $Y = k$ should be specific for $r$. The $q$ logits $\log(P(Y = 1|\boldsymbol{x})/P(Y = k|\boldsymbol{x})), \dots, \log(P(Y = q|\boldsymbol{x})/P(Y = k|\boldsymbol{x}))$ specified in (8.1) determine the response probabilities $P(Y = 1|\boldsymbol{x}), \dots, P(Y = k|\boldsymbol{x})$ uniquely. From $P(Y = r|\boldsymbol{x}) = P(Y = k|\boldsymbol{x}) \exp(\boldsymbol{x}^T\boldsymbol{\beta}_r)$ one obtains $\sum_{r=1}^{k-1} P(Y = r|\boldsymbol{x}) = P(Y = k|\boldsymbol{x}) \sum_{r=1}^{k-1} \exp(\boldsymbol{x}^T\boldsymbol{\beta}_r)$. By adding $P(Y = k|\boldsymbol{x})$ on the left- and right-hand sides one obtains

$$P(Y = k|\boldsymbol{x}) = \frac{1}{1 + \sum_{r=1}^{k-1} \exp(\boldsymbol{x}^T\boldsymbol{\beta}_r)}.$$

When $P(Y = k|\boldsymbol{x})$ is inserted into (8.1) one obtains the probabilities of the multinomial model given in the following box.

**Multinomial Logit Model with Reference Category $k$**

$$\log\left(\frac{P(Y=r|\boldsymbol{x})}{P(Y=k|\boldsymbol{x})}\right) = \boldsymbol{x}^T\boldsymbol{\beta}_r, \qquad r=1,\ldots,k-1, \tag{8.2}$$

or, equivalently,

$$P(Y=r|\boldsymbol{x}) = \frac{\exp(\boldsymbol{x}^T\boldsymbol{\beta}_r)}{1+\sum_{s=1}^{k-1}\exp(\boldsymbol{x}^T\boldsymbol{\beta}_s)}, \qquad r=1,\ldots,k-1, \tag{8.3}$$

$$P(Y=k|\boldsymbol{x}) = \frac{1}{1+\sum_{s=1}^{k-1}\exp(\boldsymbol{x}^T\boldsymbol{\beta}_s)}.$$

The representation of the nomnial logit model depends on the choice of the reference category. Instead of $k$, any category from $1,\ldots,k$ could have been chosen as the reference category. The necessity to specify a reference category is due to the constraint $\Sigma_r P(Y=r|\boldsymbol{x})=1$. The consequence of this constraint is that only $q=k-1$ response categories may be specified; the remaining probability is implicitly determined. A generic form of the logit model is given by

$$P(Y=r|\boldsymbol{x}) = \frac{\exp(\boldsymbol{x}^T\boldsymbol{\beta}_r)}{\sum_{s=1}^{k}\exp(\boldsymbol{x}^T\boldsymbol{\beta}_s)}, \tag{8.4}$$

where additional side constraints have to be specified to fulfill $\Sigma_r P(Y=r|\boldsymbol{x})=1$. It is obvious that without side constraints the parameters $\boldsymbol{\beta}_1,\ldots\boldsymbol{\beta}_k$ are not identifiable. If $\boldsymbol{\beta}_r$ is replaced by $\boldsymbol{\beta}_r+\boldsymbol{c}$ with $\boldsymbol{c}$ denoting some fixed vector, the form (8.4) also holds with parameters $\tilde{\boldsymbol{\beta}}_r=\boldsymbol{\beta}_r+\boldsymbol{c}$.

**Generic Multinomial Logit Model**

$$P(Y=r|\boldsymbol{x}) = \frac{\exp(\boldsymbol{x}^T\boldsymbol{\beta}_r)}{\Sigma_{s=1}^{k}\exp(\boldsymbol{x}^T\boldsymbol{\beta}_s)} \tag{8.5}$$

with optional side constraints

$$\begin{aligned}
\boldsymbol{\beta}_k^T &= (0,\ldots,0) && \text{reference category } k\\
\boldsymbol{\beta}_{r_0}^T &= (0,\ldots,0) && \text{reference category } r_0\\
\sum_{s=1}^{k}\boldsymbol{\beta}_s &= (0,\ldots,0) && \text{symmetric side constraint}
\end{aligned}$$

### Side Constraints

The side constraint $\boldsymbol{\beta}_k = \mathbf{0}$ immediately yields the logit model with reference category $k$. If one chooses the side constraints $\boldsymbol{\beta}_{r_0} = \mathbf{0}$, one obtains

$$\log\left(\frac{P(Y=r|\boldsymbol{x})}{P(Y=r_0|\boldsymbol{x})}\right) = \boldsymbol{x}^T\boldsymbol{\beta}_r,$$

which is equivalent to choosing $r_0$ as the reference category. It should be noted that the choice of the reference category is essential for interpreting the parameters. If $r_0$ is the reference category, $\boldsymbol{\beta}_r$ determines the logits $\log(P(Y=r|\boldsymbol{x})/P(Y=r_0|\boldsymbol{x}))$.

A symmetric form of the side constraint is given by

$$\sum_{s=1}^{k}\boldsymbol{\beta}_s = \mathbf{0}.$$

Then parameter interpretation is quite different; it refers to the "median" response. Let the median response be defined by the geometric mean

$$GM(\boldsymbol{x}) = \sqrt[k]{\prod_{s=1}^{k}P(Y=s|\boldsymbol{x})} = (\prod_{s=1}^{k}P(Y=s|\boldsymbol{x}))^{1/k}.$$

Then one can derive from (8.5)

$$\log\left(\frac{P(Y=r|\boldsymbol{x})}{GM(\boldsymbol{x})}\right) = \boldsymbol{x}^T\boldsymbol{\beta}_r$$

(Exercise 8.2). Therefore, $\boldsymbol{\beta}_r$ reflects the effects of $\boldsymbol{x}$ on the logits when $P(Y=r|\boldsymbol{x})$ is compared to the geometric mean response $GM(\boldsymbol{x})$.

It should be noted that whatever side constraint is used, the log-odds between two response probabilities and the corresponding weight are given by

$$\log\left(\frac{P(Y=r|\boldsymbol{x})}{P(Y=s|\boldsymbol{x})}\right) = \boldsymbol{x}^T(\boldsymbol{\beta}_r - \boldsymbol{\beta}_s),$$

which follows from (8.5) for any choice of response categories $r, s \in \{1,\ldots,k\}$. The transformations between different side constraints are rather simple. Let $\boldsymbol{\beta}_1,\ldots,\boldsymbol{\beta}_q$ denote the vectors with side constraint $\boldsymbol{\beta}_k = \mathbf{0}$ and $\boldsymbol{\beta}_1^*,\ldots,\boldsymbol{\beta}_q^*$ denote the vectors with symmetric side constraints. Then one obtains $\boldsymbol{\beta}_r = 2\boldsymbol{\beta}_r^* + \sum_{s\neq r, s<k}\boldsymbol{\beta}_s^*$ (Exercise 8.3).

The following example illustrates the interpretation of effects in the simple case with just one categorical covariate.

#### Example 8.3: Preference for Political Parties

Let us model the data from Table 8.1 with gender as the single explanatory variable (1: female, 0: male). The effect of gender on the preference is investigated by use of the logit model

$$\log\left(\frac{P(Y=r|x)}{P(Y=1|x)}\right) = \beta_{r0} + x_G\beta_r,$$

where $x_G = 1$ for female respondents and $x_G = 0$ for male respondents. Implicitly category 1 (CDU) has been chosen as the reference category ($\beta_{10} = 0$). The interpretation of parameters follows from

$$\beta_{r0} = \log\left(\frac{P(Y=r|x_G=0)}{P(Y=1|x_G=0)}\right), \qquad \mathrm{e}^{\beta_{10}} = \frac{P(Y=r|x_G=0)}{P(Y=1|x_G=0)},$$

$$\beta_r = \log\left(\frac{P(Y=r|x_G=1)/P(Y=1|x_G=1)}{P(Y=r|x_G=0)/P(Y=1|x_G=0)}\right), \quad \mathrm{e}^{\beta_r} = \frac{P(Y=r|x_G=1)/P(Y=1|x_G=1)}{P(Y=r|x_G=0)/P(Y=1|x_G=0)}.$$

Thus $e^{\beta_{r0}}$ represents the odds of preference for party $r$ instead of reference party 1 for male respondents, and $e^{\beta_r}$ represents the odds ratio that compares the odds for female respondents to the odds for male respondents. The parameters are given in Table 8.2. For example, for male respondents, the odds of preference for party 3 instead of reference party 1 are $0.187$. A comparison of the odds from female and male respondents yields $1.259$, signaling that female respondents prefer the Green Party stronger than male respondents. □

TABLE 8.2: Parameter estimates for party preference data with covariate gender and reference category CDU.

| | $\beta_{0r}$ | $e^{\beta_{0r}}$ | $\beta_r$ | $e^{\beta_r}$ |
|---|---|---|---|---|
| CDU (1) | 0 | 1 | 0 | 1 |
| SPD (2) | 0.392 | 1.480 | −0.068 | 0.934 |
| Greens (3) | −1.677 | 0.187 | 0.230 | 1.259 |
| Liberals (4) | −2.509 | 0.081 | −0.112 | 0.894 |

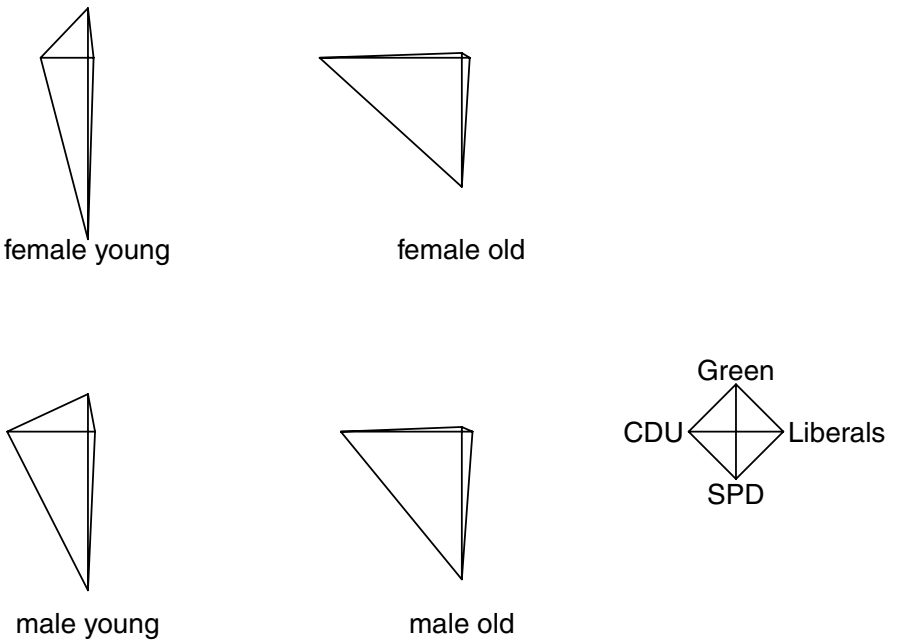


FIGURE 8.4: Star-plots for subpopulations of party preference data.

A simple way to visualize the response probabilities is by use of star-plots, which are in common use in multivariate statistics. The star-plot applied to response probabilities codes the probabilities or relative frequencies into the length of the rays emanating from the center of the plot. Figure 8.4 shows the resulting star-plots of relative frequencies for four subpopulations of the party preference data including one symmetric plot that serves to label the rays. It illustrates the strong effects of gender and age. In all the plots a strong preference for SPD is seen. But in the younger population there is a much stronger tendency toward the Green Party than in the older population; older voters prefer the CDU. The shifting of preference toward the Green Party is much stronger for females.

In more complex models, and when continuous predictors are included, it can be advantageous to represent the exponentials of parameters rather than subpopulations as star-plots. For illustration we will consider the main effect model for the party preference data (see also Exercise 8.5). Table 8.3 shows the fitted parameters and the exponentials. The latter represent the odds ratios and therefore the modification of the probabilities in comparison to the reference category. Figure 8.5 shows the corresponding star-plots. The first star-plot, which gives the exponentials of the intercept, represents the fitted odds in the reference population (male, age category 1). In all the plots the reference category among responses is CDU and the corresponding ray length is 1. The other plots of the exponentials of parameters show the modifications resulting from the covariates. It is in particular seen that females have a stronger tendency toward the Green Party when compared to the reference category of gender (male). For age, with reference category 1, it is seen that especially in age category 4 the tendency toward the Green Party is strongly reduced.

TABLE 8.3: Parameter estimates and exponentials for party preference data with covariates gender and age and reference category CDU.

| | CDU | SPD | Greens | Liberals |
|---|---|---|---|---|
| intercept | 0 | 0.905 | −0.656 | −2.3090 |
| gender | 0 | −0.006 | 0.429 | −0.0916 |
| age2 | 0 | −0.321 | −0.328 | −0.1100 |
| age3 | 0 | −0.386 | −0.898 | −0.1930 |
| age4 | 0 | −0.854 | −2.910 | −0.2970 |
| exp(intercept) | 1 | 2.471 | 0.518 | 0.099 |
| exp(gender) | 1 | 0.994 | 1.535 | 0.912 |
| exp(age2) | 1 | 0.725 | 0.720 | 0.895 |
| exp(age3) | 1 | 0.679 | 0.407 | 0.824 |
| exp(age4) | 1 | 0.425 | 0.054 | 0.743 |

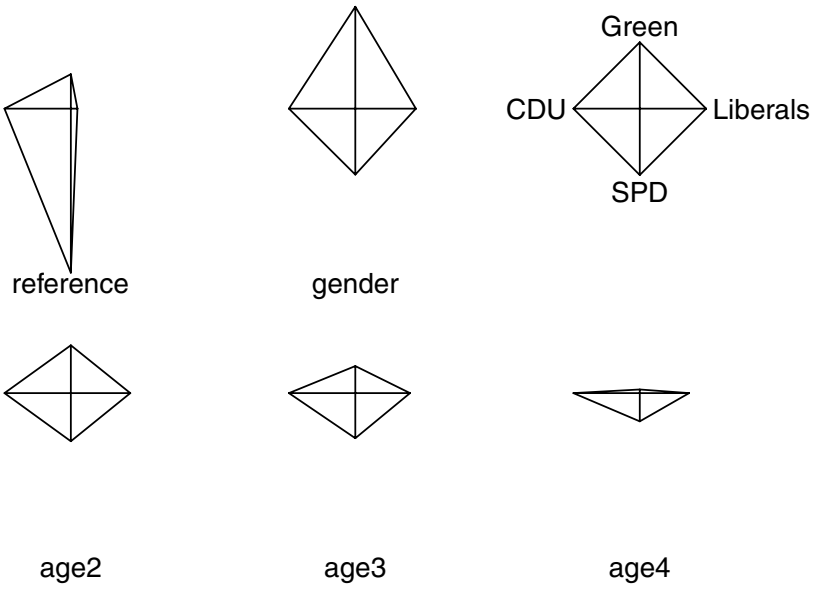


FIGURE 8.5: Star-plots of exponentials of fitted parameters for main effect model of party preference data.

## 8.3 Multinomial Model as Random Utility Model

In Section 8.2 the multinomial logit model was considered as a generalization of the binary logit model. There is an alternative motivation of the model that is not based on binary models but on random utilities. Although random utilities are considered more extensively within the framework of discrete choice models (Section 8.8), they are briefly sketched here because it helps to structure the linear predictor.

Let $U_r$ be an unobservable random utility associated with the $r$th response category. For example, $U_r$ is the (subjective) utility of a brand when the choice is among brands $1, \ldots, k$ or the "attractiveness" of the $r$th political party. Let $U_r$ be given by

$$U_r = u_r + \varepsilon_r,$$

where $u_r$ is a fixed value, representing the utility associated with the $r$th response category, and $\varepsilon_1, \ldots, \varepsilon_k$ are iid random variables with distribution function $F$. Now let the response $Y$ be determined by the *principle of maximum random utility*, which specifies the link between the observable $Y$ and the unobservable random utility by

$$Y = r \quad \Leftrightarrow \quad U_r = \max_{j=1,\ldots,k} U_j.$$

Therefore, the alternative $r$ is chosen that maximizes the random utility. If one assumes that $\varepsilon_r, \ldots, \varepsilon_k$ are iid variables with distribution function $F(x) = \exp(-\exp(-x))$, which is the Gumbel or maximum extreme value distribution, one obtains

$$P(Y = r) = \frac{\exp(u_r)}{\sum_{j=r}^{k} \exp(u_j)}$$

(e.g., Yellott, 1977; McFadden, 1973). The resulting model corresponds to the generic form of the logit model (8.5). The fixed utilities are unique only up to additive constants. Therefore, one needs some additional side constraints; for example, one may consider the differences $u_r - u_k, r = 1, \ldots, k-1$, which is equivalent to considering $k$ as the reference category. As shown in the next section, if one lets the fixed utilities $u_1, \ldots, u_k$ depend linearly on covariates one obtains the parametric multinomial logit model.

## 8.4 Structuring the Predictor

The covariates come into the multinomial logit model in the form of linear predictors:

$$\eta_r = \boldsymbol{x}^T \boldsymbol{\beta}_r.$$

In the same way as in univariate response models, the linear predictor may contain dummy variables for categorical covariates, polynomial terms for continuous variables, and interaction terms between both types of variables.

Apart from these transformations of the original observations for the structuring of the linear predictor, it is often useful to distinguish between two types of covariates, namely, *global* and *category-specific* variables. For example, when an individual chooses among alternatives $1, \ldots, k$, one may model the effect of characteristics of the individual like age and gender, which are global variables, but also account for measured attributes of the alternatives $1, \ldots, k$,

which are category-specific variables. When the choice refers to transportation mode, the potential attributes are price and duration, which vary across the alternatives and therefore are category-specific.

Let $\boldsymbol{x}$ denote the individual characteristics and $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_k$ denote the set of attributes of alternatives, where $\boldsymbol{w}_r$ are the attributes of category $r$. The first type of variable is called *global*, and the latter type *category-specific*. Then the set of linear predictors may be generalized to

$$\eta_r = \boldsymbol{x}^T\boldsymbol{\beta}_r + (\boldsymbol{w}_r - \boldsymbol{w}_k)^T\boldsymbol{\alpha}, \qquad r = 1, \ldots, k-1, \tag{8.6}$$

where $k$ is chosen as the reference category. The first term specifies the effect of the global variables, and the second term specifies the effect of the difference $\boldsymbol{w}_r - \boldsymbol{w}_k$ on the choice between category $r$ and the reference category. When $\boldsymbol{w}_r$ stands for the price of alternative $r$, it is quite natural to assume that the choice between alternatives $r$ and $k$ is determined by the difference.

The predictor (8.6) may be derived by maximizing the latent utilities. Let the latent utility of category $r$ be specified by $u_r = \boldsymbol{x}^T\boldsymbol{\gamma}_r + \boldsymbol{w}_r^T\boldsymbol{\alpha}$. Then the difference is

$$u_r - u_k = \boldsymbol{x}^T(\boldsymbol{\gamma}_r - \boldsymbol{\gamma}_k) + (\boldsymbol{w}_r - \boldsymbol{w}_k)^T\boldsymbol{\alpha},$$

which has the form given in (8.6) with $\boldsymbol{\beta}_r = (\boldsymbol{\gamma}_r - \boldsymbol{\gamma}_k)$. Of course one may also specify interactions between the two types of variables. Let, for example, $x_G$ be a dummy variable for gender and $w_r$ denote the price of alternative $r$. Then the model

$$\eta_r = \beta_{0r} + x_G\beta_G + (w_r - w_k)\alpha_1 + x_G(w_r - w_k)\alpha_2$$

allows that the effect of prices depends on gender.

Formally, it is not necessary to distinguish between global and category-specific variables. One may always define one long vector of variables that contains all the specified variables. For example, the predictor with only global variables $\eta_r = \boldsymbol{x}^T\boldsymbol{\beta}_r$ may also be written as $\eta_r = (\mathbf{0}^T, \ldots, \boldsymbol{x}^T, \ldots, \mathbf{0}^T)\boldsymbol{\beta}$, where $\boldsymbol{\beta}^T = (\boldsymbol{\beta}_1^T, \ldots, \boldsymbol{\beta}_q^T)$. The model with category-specific variables $\eta_r = \boldsymbol{x}^T\boldsymbol{\beta}_r + (\boldsymbol{w}_r^T - \boldsymbol{w}_k^T)\boldsymbol{\alpha}$ has the form $\eta_r = (\mathbf{0}^T, \ldots, \boldsymbol{x}^T, \ldots, \mathbf{0}^T, \boldsymbol{w}_r^T - \boldsymbol{w}_k^T)\boldsymbol{\beta}$, where $\boldsymbol{\beta}$ is now given as $\boldsymbol{\beta}^T = (\boldsymbol{\beta}_1^T, \ldots, \boldsymbol{\beta}_q^T, \boldsymbol{\alpha}^T)$. Thus one always obtains the form $\eta_r = \boldsymbol{x}_r^T\boldsymbol{\beta}$, where $\boldsymbol{x}_r$ may or may not depend on $r$.

In econometrics, sometimes also for category-specific predictors category-specific effects are assumed. The latent utility $u_r = \boldsymbol{x}^T\boldsymbol{\gamma}_r + \boldsymbol{w}_r^T\boldsymbol{\alpha}_r$ with category-specific effect $\boldsymbol{\alpha}_r$ yields the difference of utilities that defines the linear predictor for reference category $k$

$$\eta_r = u_r - u_k = \boldsymbol{x}^T\boldsymbol{\beta}_r + \boldsymbol{w}_r^T\boldsymbol{\alpha}_r - \boldsymbol{w}_k^T\boldsymbol{\alpha}_k,$$

where $\boldsymbol{\beta}_r = (\boldsymbol{\gamma}_r - \boldsymbol{\gamma}_k)$. The total set of parameters that defines the total vector $\boldsymbol{\beta}$ now contains the $k-1$ $\boldsymbol{\beta}$-parameters $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_{k-1}$, and the $k$ $\boldsymbol{\alpha}$-vectors $\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_k$.

### Example 8.4: Travel Mode

The choice of travel mode of $n = 840$ passengers in Australia was investigated by Greene (2003). The data are available from the R package *Ecdat*. The alternatives of travel mode were air, train, bus, and car, which have frequencies $0.276, 0.300, 0.142$, and $0.280$. Air serves as the reference category. As category-specific variables we consider travel time in vehicle (timevc) and cost, and as the global variable we consider household income (income). The estimates given in Table 8.4 show that income seems to be influential for the preference of train and bus over airplane. Moreover, time in vehicle seems to matter for the preference of the travel mode. Cost turns out to be non-influential if income is in the predictor (see also Exercise 8.10). □

TABLE 8.4: Estimated coefficients for travel mode data.

| | Estimate | Std. Error | $z$-Value | Pr(>\|$z$\|) |
|---|---|---|---|---|
| train | 3.525 | 0.654 | 5.381 | 0.0 |
| bus | 2.278 | 0.717 | 3.174 | 0.001 |
| car | 1.533 | 0.706 | 2.170 | 0.029 |
| train:income | −0.056 | 0.012 | −4.588 | 0.0 |
| bus:income | −0.035 | 0.013 | −2.705 | 0.006 |
| car:income | −0.002 | 0.010 | −0.226 | 0.820 |
| timevc | −0.003 | 0.001 | −3.274 | 0.001 |
| cost | −0.001 | 0.005 | −0.293 | 0.769 |

## 8.5 Logit Model as Multivariate Generalized Linear Model

For simplicity, let $\pi_r = P(Y = r|\boldsymbol{x}, \{\boldsymbol{w}_j\})$ denote the response probability for one observation $Y$ with covariates $\boldsymbol{x}, \{\boldsymbol{w}_j\}$, where $\boldsymbol{w}_j$ are category-specific attributes. Then, with $\tilde{\boldsymbol{w}}_r = \boldsymbol{w}_r - \boldsymbol{w}_k$ and the linear predictor $\eta_r = \boldsymbol{x}^T\boldsymbol{\beta}_r + \tilde{\boldsymbol{w}}_r^T\boldsymbol{\alpha}$, one may write the $q$ equations that specify the nominal logit model with reference category $k$ in matrix form by

$$\begin{pmatrix} \log\left(\pi_1/(1-\pi_1-\cdots-\pi_q)\right) \\ \vdots \\ \log\left(\pi_q/(1-\pi_1-\cdots-\pi_q)\right) \end{pmatrix} = \begin{pmatrix} \boldsymbol{x}^T & & 0 & \tilde{\boldsymbol{w}}_1^T \\ & \ddots & & \vdots \\ 0 & & \boldsymbol{x}^T & \tilde{\boldsymbol{w}}_q^T \end{pmatrix} \begin{pmatrix} \boldsymbol{\beta}_1 \\ \vdots \\ \boldsymbol{\beta}_q \\ \boldsymbol{\alpha} \end{pmatrix}. \tag{8.7}$$

The predictor for the $r$th logit $\log(\pi_r/(1-\pi_1-\ldots-\pi_q))$ has the form

$$\eta_r = \boldsymbol{x}^T\boldsymbol{\beta}_r + \tilde{\boldsymbol{w}}_r^T\boldsymbol{\alpha} = (0,\ldots,0,\boldsymbol{x}^T,0,\ldots,\tilde{\boldsymbol{w}}_r^T)\boldsymbol{\beta} = \boldsymbol{x}_r\boldsymbol{\beta},$$

where $\boldsymbol{\beta}^T = (\boldsymbol{\beta}_1^T,\ldots\boldsymbol{\beta}_q^T,\boldsymbol{\alpha}^T)$ and $\boldsymbol{x}_r = (0,\ldots,0,\boldsymbol{x}^T,0,\ldots,0,\tilde{\boldsymbol{w}}_r)$ is the corresponding design vector. Thus the general form of (8.7) is

$$g(\boldsymbol{\pi}) = \boldsymbol{X}\boldsymbol{\beta},$$

where $\boldsymbol{\pi}^T = (\pi_1,\ldots,\pi_q)$ is the vector the of response probabilities, $\boldsymbol{X}$ is a design matrix that corresponds to the total parameter vector $\boldsymbol{\beta}$, and $g$ is the link function. For the logit model (8.7) the vector-valued *link function* $g = (g_1,\ldots,g_q) : \mathbb{R}^q \to \mathbb{R}^q$ is given by

$$g_r(\pi_1,\ldots,\pi_q) = \log\left(\frac{\pi_r}{1-\pi_1-\cdots-\pi_q}\right).$$

As usual in generalized linear models, an equivalent form is

$$\boldsymbol{\pi} = h(\boldsymbol{X}\boldsymbol{\beta}),$$

where $h = (h_1,\ldots,h_q) = g^{-1}$ is the response function, which in the present case has components

$$h_r(\eta_1,\ldots,\eta_q) = \frac{\exp(\eta_r)}{1+\sum_{s=1}^q \exp(\eta_s)}.$$

Thus, for *one* observation, the nominal logit model has the general form

$$g(\boldsymbol{\pi}) = \boldsymbol{X}\boldsymbol{\beta} \quad \text{or} \quad \boldsymbol{\pi} = h(\boldsymbol{X}\boldsymbol{\beta})$$

for an appropriately chosen vector-valued link function, design matrix, and parameter vector. For the given data $(\boldsymbol{y}_i, \boldsymbol{x}_i), i = 1, \ldots, n$, one has

$$g(\boldsymbol{\pi}(\boldsymbol{x}_i)) = \boldsymbol{X}_i\boldsymbol{\beta} \quad \text{or} \quad \boldsymbol{\pi}(\boldsymbol{x}_i) = h(\boldsymbol{X}_i\boldsymbol{\beta}),$$

where $\boldsymbol{\pi}(\boldsymbol{x})^T = (\pi_1(\boldsymbol{x}), \ldots, \pi_q(\boldsymbol{x}))$, $\pi_r(\boldsymbol{x}_i) = P(Y_i = r|\boldsymbol{x}_i)$, and $\boldsymbol{X}_i$ is composed from the covariates $\boldsymbol{x}_i$.

## 8.6 Inference for Multicategorical Response Models

Let the data be given by $(\boldsymbol{y}_i, \boldsymbol{x}_i), i = 1, \ldots, N$, where $\boldsymbol{x}_i$ contains all the covariates that are observed, including category-specific variables. Given $\boldsymbol{x}_i$, one assumes a multinomial distribution, $\boldsymbol{y}_i \sim \mathrm{M}(n_i, \boldsymbol{\pi}_i)$. This means that at measurement point $\boldsymbol{x}_i$ one has $n_i$ observations. In particular, if $\boldsymbol{x}_i$ contains categorical variables, usually more than one observation is collected at a fixed value of covariates. For example, in a sample survey respondents may be characterized by gender and educational level. Then, for fixed values of gender and educational level, one usually observes more than one response. This may be seen as a grouped data case, where the number of counts stored in $\boldsymbol{y}_i$ is the sum of the responses of single respondents, with each one having multinomial distribution $\mathrm{M}(1, \boldsymbol{\pi}_i)$. Instead of the multinomial distribution $\boldsymbol{y}_i \sim \mathrm{M}(n_i, \boldsymbol{\pi}_i)$ one may also consider the scaled multinomials or proportions $\bar{\boldsymbol{y}}_i = \boldsymbol{y}_i/n_i$, which are also denoted by $\boldsymbol{p}_i = \bar{\boldsymbol{y}}_i$. The components of $\boldsymbol{p}_i^T = (p_{i1}, \ldots, p_{iq})$ contain the relative frequencies, and $p_{ir}$ is the proportion of observations in category $r$. The proportions $\bar{\boldsymbol{y}}_i = \boldsymbol{p}_i$ have the advantage that $\mathrm{E}(\boldsymbol{p}_i) = \boldsymbol{\pi}_i$, whereas for $\boldsymbol{y}_i = n_i\boldsymbol{p}_i$ one has $\mathrm{E}(\boldsymbol{y}_i) = n_i\boldsymbol{p}_i$. The model that is assumed to hold has the form

$$g(\boldsymbol{\pi}_i) = \boldsymbol{X}_i\boldsymbol{\beta} \quad \text{or} \quad \boldsymbol{\pi}_i = h(\boldsymbol{X}_i\boldsymbol{\beta}).$$

### 8.6.1 Maximum Likelihood Estimation

The multinomial distribution has the form of a multivariate exponential family. Let $\boldsymbol{y}_i^T = (y_{i1}, \ldots, y_{iq}) \sim \mathrm{M}(n_i, \boldsymbol{\pi}_i)$, $i = 1, \ldots, N$, denote the multinomial distribution with $k = q + 1$ categories. Then the probability mass function is

$$\begin{aligned} f(\boldsymbol{y}_i) &= \frac{n_i!}{y_{i1}! \cdots y_{iq}!(n_i - y_{i1} - \cdots - y_{iq})!} \pi_{i1}^{y_{i1}} \cdot \cdots \cdot \pi_{iq}^{y_{iq}} (1 - \pi_{i1} - \ldots - \pi_{iq})^{(n_i - y_{i1} - \ldots - y_{iq})} \\ &= \exp\left(\boldsymbol{y}_i^T\boldsymbol{\theta}_i + n_i \log(1 - \pi_{i1} - \ldots - \pi_{iq}) + \log(c_i)\right) \\ &= \exp\left([\boldsymbol{p}_i^T\boldsymbol{\theta}_i + \log(1 - \pi_{i1} - \ldots - \pi_{iq})]/(1/n_i) + \log(c_i)\right), \end{aligned}$$

where the canonical parameter vector is $\boldsymbol{\theta}_i^T = (\theta_{i1}, \ldots, \theta_{iq})$, $\theta_{ir} = \log(\pi_{ir}/(1 - \pi_{i1} - \cdots - \pi_{iq}))$, the dispersion parameter is $1/n_i$, and $c_i = n_i!/(y_{i1}! \ldots y_{iq}!(n_i - y_{i1} - \cdots - y_{iq})!)$. One obtains the likelihood

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{N} c_i \, \pi_{i1}^{y_{i1}} \cdot \cdots \cdot \pi_{iq}^{y_{iq}} (1 - \pi_{i1} - \cdots - \pi_{iq})^{n_i - y_{i1} - \cdots - y_{iq}}$$

and the log-likelihood $l(\boldsymbol{\beta}) = \sum_{i=1}^{N} l_i(\boldsymbol{\pi}_i)$ with

$$\begin{aligned} l_i(\boldsymbol{\pi}_i) = n_i \left\{ \sum_{r=1}^{q} p_{ir} \log\left(\frac{\pi_{ir}}{1 - \pi_{i1} - \cdots - \pi_{iq}}\right) + \log(1 - \pi_{i1} - \cdots - \pi_{iq}) \right\} \\ + \log(c_i). \end{aligned} \tag{8.8}$$

The score function $\boldsymbol{s}(\boldsymbol{\beta}) = \partial l(\boldsymbol{\beta})/\partial\boldsymbol{\beta}$ for the model $\boldsymbol{\pi}_i = h(\boldsymbol{X}_i\boldsymbol{\beta})$ has the form

$$\boldsymbol{s}(\boldsymbol{\beta}) = \sum_{i=1}^{N} \boldsymbol{X}_i^T \boldsymbol{D}_i(\boldsymbol{\beta}) \boldsymbol{\Sigma}_i^{-1}(\boldsymbol{\beta})(\boldsymbol{p}_i - \boldsymbol{\pi}_i),$$

where $\boldsymbol{\eta}_i = \boldsymbol{X}_i\boldsymbol{\beta}$ and $\boldsymbol{D}_i(\boldsymbol{\beta}) = \partial h(\boldsymbol{\eta}_i)/\partial\boldsymbol{\eta} = (\partial g(\boldsymbol{\pi}_i)/\partial\boldsymbol{\pi})^{-1}$. It should be noted that the matrix $\boldsymbol{D}_i(\boldsymbol{\beta})$ with entries $\partial h_r(\boldsymbol{\eta}_i)/\partial\eta_s$ is not a symmetric matrix. The covariance $\boldsymbol{\Sigma}_i(\boldsymbol{\beta})$ is determined by the multinomial distribution and has the form

$$\begin{aligned}\boldsymbol{\Sigma}_i(\boldsymbol{\beta}) &= \frac{1}{n_i}\begin{pmatrix} \pi_{i1}(1-\pi_{i1}) & -\pi_{i1}\pi_{i2} & \dots & -\pi_{i1}\pi_{iq} \\ & \pi_{i2}(1-\pi_{i2}) & & \\ & & \ddots & \\ -\pi_{iq}\pi_{i1} & & & \pi_{iq}(1-\pi_{iq}) \end{pmatrix} \\ &= [\,\text{Diag}\,(\boldsymbol{\pi}_i) - \boldsymbol{\pi}_i\boldsymbol{\pi}_i^T]/n_i.\end{aligned}$$

In closed form one obtains

$$\boldsymbol{s}(\boldsymbol{\beta}) = \boldsymbol{X}^T\boldsymbol{D}(\boldsymbol{\beta})\boldsymbol{\Sigma}(\boldsymbol{\beta})^{-1}(\boldsymbol{p}-\boldsymbol{\pi}) = \boldsymbol{X}^T\boldsymbol{W}(\boldsymbol{\beta})\boldsymbol{D}(\boldsymbol{\beta})^{-T}(\boldsymbol{p}-\boldsymbol{\pi}),$$

where $\boldsymbol{X}^T = (\boldsymbol{X}_1^T, \dots, \boldsymbol{X}_N^T)$, $\boldsymbol{D}(\boldsymbol{\beta})$ and $\boldsymbol{\Sigma}(\boldsymbol{\beta})$ are block-diagonal matrices with blocks $\boldsymbol{D}_i(\boldsymbol{\beta})$, $\boldsymbol{\Sigma}_i(\boldsymbol{\beta})$, respectively, and $\boldsymbol{W}(\boldsymbol{\beta}) = \boldsymbol{D}(\boldsymbol{\beta})\boldsymbol{\Sigma}(\boldsymbol{\beta})^{-1}\boldsymbol{D}(\boldsymbol{\beta})^T$ is a block-diagonal matrix with blocks $\boldsymbol{W}_i(\boldsymbol{\beta}) = \boldsymbol{D}_i(\boldsymbol{\beta})\boldsymbol{\Sigma}_i(\boldsymbol{\beta})^{-1}\boldsymbol{D}_i(\boldsymbol{\beta})^T$. For $\boldsymbol{\Sigma}_i(\boldsymbol{\beta})^{-1}$ an explicit form is available (see Exercise 8.6).

The expected information or Fisher matrix $\boldsymbol{F}(\boldsymbol{\beta}) = \text{E}\left(-\partial l/\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T\right) = \text{cov}(\boldsymbol{s}(\boldsymbol{\beta}))$ has the form

$$\boldsymbol{F}(\boldsymbol{\beta}) = \sum_{i=1}^{N} \boldsymbol{X}_i^T\boldsymbol{W}_i(\boldsymbol{\beta})\boldsymbol{X}_i = \boldsymbol{X}^T\boldsymbol{W}(\boldsymbol{\beta})\boldsymbol{X}.$$

The blocks $\boldsymbol{W}_i(\boldsymbol{\beta})$ in the weight matrix can also be given in the form $\boldsymbol{W}_i(\boldsymbol{\beta}) = (\frac{\partial g(\boldsymbol{\pi}_i)}{\partial\boldsymbol{\pi}^T}\boldsymbol{\Sigma}_i(\boldsymbol{\beta})\frac{\partial g(\boldsymbol{\pi}_i)}{\partial\boldsymbol{\pi}})^{-1}$, which is an approximation to the inverse of the covariance of $g(\boldsymbol{p}_i)$ when the model holds. For the logit model, which corresponds to the canonical link, simpler forms of the score function and the Fisher matrix can be found (see Exercise 8.4).

The estimate $\hat{\boldsymbol{\beta}}$ is under regularity conditions asymptotically ($n_1+\dots+n_N \to \infty$) normally distributed with

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

for details see Fahrmeir and Kaufmann (1985). The score function and Fisher matrix have the same forms as in univariate GLMs, namely, $\boldsymbol{s}(\boldsymbol{\beta}) = \boldsymbol{X}^T\boldsymbol{D}(\boldsymbol{\beta})\boldsymbol{\Sigma}(\boldsymbol{\beta})^{-1}(\boldsymbol{p}-\boldsymbol{\pi})$ and $\boldsymbol{F}(\boldsymbol{\beta}) = \boldsymbol{X}^T\boldsymbol{W}(\boldsymbol{\beta})\boldsymbol{X}$. But for multicategorical responses the design matrix is composed from matrices for single observations, and the weight matrix $\boldsymbol{W}(\boldsymbol{\beta})$ as well as the matrix of derivatives $\boldsymbol{D}(\boldsymbol{\beta})$ are block-diagonal matrices in contrast to the univariate models, where $\boldsymbol{W}(\boldsymbol{\beta})$ and $\boldsymbol{D}(\boldsymbol{\beta})$ are diagonal matrices.

### Separate Fitting of Binary Models

When one considers only two categories, say $r$ and $k$, the multinomial model looks like a binary logit model. If $k$ is chosen as reference category, one has

$$\log(\frac{P(Y=r|\boldsymbol{x})}{P(Y=k|\boldsymbol{x})}) = \boldsymbol{x}^T\boldsymbol{\beta}_r.$$

Therefore, the parameters $\boldsymbol{\beta}_r$ can also be estimated by fitting a binary logit model using only observations in categories $r$ and $k$. The resulting estimates are conditional on classification in categories $r$ and $k$. The estimates obtained by fitting $k-1$ separate binary models differ from the estimates obtained from the full likelihood of the multinomial model. In particular, they tend to have larger standard errors, although the effect is usually small if the reference category is chosen as the category with most of the observations (Begg and Gray, 1984). Another advantage of the multinomial likelihood is that the testing of hypotheses that refer to parameters that are linked to different categories is straightforward, for example, by using likelihood ratio tests.

## 8.6.2 Goodness-of-Fit

As in univariate GLMs, the goodness-of-fit may be checked by the Pearson statistic and the deviance. Again asymptotic results are obtained only for grouped data, where the number of repetitions $n_i$ taken at observation vector $\boldsymbol{x}_i$ is not too small.

### Pearson Statistic

When considering the discrepancy between observations and fit, one should have in mind that responses are vector-valued. One wants to compare the observation vectors $\boldsymbol{p}_i = \boldsymbol{y}_i/n$ and the fitted vector $\boldsymbol{\pi}_i$, where both vectors have dimension $q$, since one category (in our case the last one) is omitted from the vector. By using for the last category the observation $p_{ik} = 1 - p_{i1} - \cdots - p_{iq}$ and the fit $\hat{\pi}_{ik} = 1 - \hat{\pi}_{i1} - \cdots - \hat{\pi}_{iq}$, one defines the quadratic *Pearson residual* of the $i$th observation by

$$\chi^2_P(\boldsymbol{p}_i, \hat{\boldsymbol{\pi}}_i) = \sum_{r=1}^{k} n_i \frac{(p_{ir} - \hat{\pi}_{ir})^2}{\hat{\pi}_{ir}}.$$

The corresponding Pearson statistic is given by

$$\chi^2_P = \sum_{i=1}^{N} \chi^2_P(\boldsymbol{p}_i, \hat{\boldsymbol{\pi}}_i) = \sum_{i=1}^{N} (\boldsymbol{p}_i - \hat{\boldsymbol{\pi}}_i)^T \boldsymbol{\Sigma}_i^{-1}(\hat{\boldsymbol{\beta}})(\boldsymbol{p}_i - \hat{\boldsymbol{\pi}}_i),$$

where the last form is derived by explicitly deriving the inverse of the $q \times q$-matrix $\boldsymbol{\Sigma}_i(\hat{\boldsymbol{\beta}}) = [\,\text{Diag}\,(\hat{\boldsymbol{\pi}}_i) - \boldsymbol{\pi}_i \boldsymbol{\pi}_i^T]/n_i$.

### Deviance

The deviance for multinomial (grouped) observations is given by $D = 2\sum_{i=1}^{N} l_i(\boldsymbol{p}_i) - l_i(\hat{\boldsymbol{\pi}}_i)$, yielding

$$D = 2\sum_{i=1}^{N} n_i \sum_{r=1}^{k} p_{ir} \log\left(\frac{p_{ir}}{\hat{\pi}_{ir}}\right).$$

The corresponding quadratic *deviance residuals* are given by

$$\chi^2_D(\boldsymbol{p}_i, \hat{\boldsymbol{\pi}}_i) = 2n_i \sum_{r=1}^{k} p_{ir} \log\left(\frac{p_{ir}}{\hat{\pi}_{ir}}\right).$$

Under the assumptions of the fixed cells asymptotic ($n_i/N \to \lambda_i \in (0,1)$) and regularity conditions, $\chi^2_P$ and $D$ are asymptotically $\chi^2$-distributed with $N(k-1)-p$ degrees of freedom, where $N$ is the number of (grouped) observations, $k$ is the number of response categories and $p$ is the number of estimated parameters.

**Goodness-of-Fit Tests**

Pearson statistic

$$\chi^2_P = \sum_{i=1}^{N} \chi^2_P(\boldsymbol{p}_i, \hat{\boldsymbol{\pi}}_i)$$

$$\text{with} \qquad \chi^2_P(\boldsymbol{p}_i, \hat{\boldsymbol{\pi}}_i) = n_i \sum_{r=1}^{k} (p_{ir} - \hat{\pi}_{ir})^2 / \hat{\pi}_{ir}$$

Deviance

$$D = \sum_{i=1}^{N} \chi^2_D(\boldsymbol{p}_i, \hat{\boldsymbol{\pi}}_i)$$

$$\text{with} \qquad \chi^2_D(\boldsymbol{p}_i, \hat{\boldsymbol{\pi}}_i) = 2n_i \sum_{r=1}^{k} p_{ir} \log\left(\frac{p_{ir}}{\hat{\pi}_{ir}}\right)$$

$D, \chi^2_P$ are compared to the asymptotic $\chi^2$-distribution with $\mathrm{N}(k-1) - p$ degrees of freedom ($n_i/N \to \lambda_i \in [0,1]$).

As in the case of the binomial distribution, one should be cautious when using these goodness-of-fit statistics if $n_i$ is small. For the ungrouped case, for example, if covariates are continuous, one has $n_i = 1$ for all observations and the deviance becomes

$$D = -2 \sum_{r=1}^{k} I(Y_i = r) \log(\hat{\pi}_{ir}) = -2 \sum_{i=1}^{N} \log(\hat{\pi}_{iY_i}),$$

where $Y_i \in \{1, \ldots, k\}$ denotes the $i$th observation, and $I$ is the indicator function with $I(a) = 1$ if a holds and $I(a) = 0$ otherwise. The value of $D$ should certainly not be compared to quantiles of the $\chi^2$-distribution since one has $N$ observations and $N(k-1)$ degrees of freedom. For small values of $n_i$ alternative asymptotic concepts should be used (see also next section).

### Power-Divergence Family

A general family of goodness-of-fit statistics that comprises the deviance and the Pearson statistic is the power-divergence family, which for $\lambda \in (-\infty, \infty)$ has the form

$$S_\lambda = \sum_{i=1}^{N} SD_\lambda(\boldsymbol{p}_i, \hat{\boldsymbol{\pi}}_i),$$

where

$$SD_\lambda(\boldsymbol{p}_i, \hat{\boldsymbol{\pi}}_i) \;=\; \frac{2n_i}{\lambda(\lambda+1)} \sum_{r=1}^{k} p_{ir} \left[ \left( \frac{p_{ir}}{\hat{\pi}_{ir}} \right)^{\lambda} - 1 \right]. \tag{8.9}$$

As special cases, one obtains for $\lambda = 1$ the Pearson statistic and for the limit $\lambda \to 0$ the deviance. However, the family includes further statistics that have been proposed in the literature; in particular one obtains for the limit $\lambda \to -1$ Kullback's minimum discrimination information statistic with $SD_{-1}(\boldsymbol{p}_i, \hat{\boldsymbol{\pi}}_i) = n_i \sum_r \hat{\pi}_{ir} \log(\hat{\pi}_{ir}/p_{ir})$ and for $\lambda = -2$ Neyman's minimum modified $\chi^2$-statistic with $SD_{-2}(\boldsymbol{p}_i, \hat{\boldsymbol{\pi}}_i) = n_i \sum_r (p_{ir} - \hat{\pi}_{ir})^2/p_{ir}$.

Under the assumptions of fixed cells asymptotics one obtains the same asymptotic $\chi^2$-distribution for any $\lambda$, if estimates $\hat{\pi}_{ir}$ are best asymptotically normal (BAN-) distributed estimates, as for example the ML estimates or estimates obtained from minimizing $S_\lambda$. Fixed cells asymptotics postulates in particular fixed numbers of groups and large values of $n_i$. If the number of observations $n_i$ at a fixed value $\boldsymbol{x}_i$ is small, the usual asymptotic fails. An alternative is *increasing cells asymptotics*, which allows that with increasing sample size $n = n_1 + \cdots + n_N \to \infty$ the number of groups also increases $N \to \infty$. However, under increasing cells asymptotics the asymptotic distribution is normal and depends on $\lambda$. Therefore, if goodness-of-fit statistics like the deviance and Pearson statistic differ strongly, one might suspect that fixed asymptotics does not apply. For more details on increasing cells asymptotics see Read and Cressie (1988) and Osius and Rojek (1992).

Further test statistics are variations of the tests for binary responses considered in Section 4.2.3. One is the Hosmer-Lemeshow statistic, which has been extended to the multinomial model by Pigeon and Heyse (1999), and the other is based on smoothing of residuals. It has been adapted to the multinomial model by Goeman and le Cessie (2006). Of course problems found in the binary case, namely, low power of the Hosmer-Lemeshow statistic and the restriction to low dimensions for smoothed residuals, carry over to the multinomial case.

### 8.6.3 Diagnostic Tools

#### Hat Matrix

For multicategory response models, the iteratively reweighted least-squares fitting procedure (see Chapter 3) has the form

$$\hat{\boldsymbol{\beta}}^{(l+1)} = (\boldsymbol{X}^T \boldsymbol{W}(\boldsymbol{\beta}^{(l)}) \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{W}(\boldsymbol{\beta}^{(l)}) \tilde{\boldsymbol{\eta}}(\hat{\boldsymbol{\beta}}^{(l)}),$$

with $\tilde{\boldsymbol{\eta}}(\hat{\boldsymbol{\beta}}^{(l)}) = \boldsymbol{X}\hat{\beta} + \boldsymbol{D}^{-1}(\hat{\boldsymbol{\beta}})^T(\boldsymbol{p}_i - \boldsymbol{\pi}_i(\boldsymbol{\beta}))$, and one obtains at convergence

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T \boldsymbol{W}(\boldsymbol{\beta}) \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{W}(\boldsymbol{\beta}) \tilde{\boldsymbol{\eta}}(\hat{\boldsymbol{\beta}}).$$

The corresponding hat matrix is

$$\boldsymbol{H} = \boldsymbol{W}^{T/2}(\hat{\boldsymbol{\beta}}) \boldsymbol{X} \boldsymbol{F}^{-1}(\hat{\boldsymbol{\beta}}) \boldsymbol{X}^T \boldsymbol{W}^{1/2}(\hat{\boldsymbol{\beta}}).$$

$\boldsymbol{H}$ is an $(Nq \times Nq)$-matrix with blocks $\boldsymbol{H}_{ij}$. The $(q \times q)$-matrix $\boldsymbol{H}_{ii}$ corresponds to the $i$th observation. As indicators for leverage one can use $\det(\boldsymbol{H}_{ii})$ or $\mathrm{tr}(\boldsymbol{H}_{ii})$.

#### Residuals

The vector of raw residuals is given by $\boldsymbol{p}_i - \hat{\boldsymbol{\pi}}_i$. Correcting for the variances of $\boldsymbol{p}_i$ yields the *Pearson residual*:

$$\boldsymbol{r}_P(\boldsymbol{p}_i, \hat{\boldsymbol{\pi}}_i) = \boldsymbol{\Sigma}_i^{-1/2}(\hat{\boldsymbol{\beta}})(\boldsymbol{p}_i - \hat{\boldsymbol{\pi}}_i),$$

which forms the Pearson statistic $\chi^2_P = \sum_i \boldsymbol{r}_P(\boldsymbol{p}_i, \hat{\boldsymbol{\pi}}_i)^T \boldsymbol{r}_P(\boldsymbol{p}_i, \hat{\boldsymbol{\pi}}_i)$. A plot of the squared Pearson residuals shows which observations have a strong impact on the goodness-of-fit. The Pearson residuals themselves are vector-valued and show how well categories are fitted.

Standardized Pearson residuals, which correct for the variance of the Pearson residual, have the form

$$\boldsymbol{r}_P(\boldsymbol{p}_i, \hat{\boldsymbol{\pi}}_i) = I - \boldsymbol{H}_{ii}^{-1/2} \boldsymbol{\Sigma}_i^{-1/2}(\hat{\boldsymbol{\beta}})(\boldsymbol{p}_i - \hat{\boldsymbol{\pi}}_i),$$

which uses the approximation $\text{cov}(\boldsymbol{p}_i - \hat{\boldsymbol{\pi}}_i) \simeq \boldsymbol{\Sigma}_i^{1/2}(\hat{\boldsymbol{\beta}})(\boldsymbol{I} - \boldsymbol{H}_{ii})\boldsymbol{\Sigma}_i^{T/2}(\hat{\boldsymbol{\beta}})$. The approximation may be derived in the same way as in the univariate case (see Section 3.10). More details on regression diagnostics are found in Lesaffre and Albert (1989).

### Testing Components of the Linear Predictor

Linear hypotheses concerning the linear predictor have the form

$$H_0 : \boldsymbol{C}\boldsymbol{\beta} = \boldsymbol{\xi} \text{ against } H_1 : \boldsymbol{C}\boldsymbol{\beta} \neq \boldsymbol{\xi},$$

where $\boldsymbol{C}$ is a fixed matrix of full rank $s \leq p$ and $\boldsymbol{\xi}$ is a fixed vector. Let, for example, the linear predictor contain only two variables, such that $\eta_r = \beta_{ro} + x_1\beta_{r1} + x_2\beta_{r2}$. Then the hypothesis that variable $x_1$ has no influence has the form

$$H_0 : \beta_{11} = \cdots = \beta_{q1} = 0 \text{ against } H_1 : \beta_{r1} \neq 0 \text{ for one } r.$$

It is easy to find a matrix $\boldsymbol{C}$ and a vector $\boldsymbol{\xi}$ that form the corresponding linear hypothesis. Hypotheses like that make it necessary to treat the multicategorical model as a multivariate model. Since the hypothesis involves parameters that correspond to more than one response category, the fitting of $q$ separate binary models could not be used directly to test if $H_0$ holds. The test statistics in common use are the same as in univariate GLMs, namely, the likelihood ratio statistic, the Wald test, and the score test. The form is the same as given in Section 4.4; one just has to replace the score function and the Fisher matrix by their multivariate analogs.

Test procedures serve to determine if the variables have significant weights. Another aspect is the explanatory value of the predictors, which can be evaluated for example by $R$-squared measures. For measures of this type see Amemiya (1981).

#### Example 8.5: Addiction

We refer to Example 8.2. In the survey people were asked, "Is addiction a disease or are addicts weak-willed?" The response was in three categories, 0: addicts are weak-willed, 1: addiction is a disease, 2: both. Table 8.5 shows the coefficients of the multinomial logit model with the covariates gender (0: male; 1: female), age in years, and university degree (0: no; 1: yes). Category 0 was chosen as the reference category. It is seen that women show a stronger tendency to accept addiction as a disease than men. The same effect is found for respondents with a university degree. Age also shows a significant effect, and at least a quadratic effect should be included since the inclusion of a quadratic effect reduces the deviance by 32.66. Figure 8.6 shows the estimated probabilities against age for males and females with a university degree (compare also the smooth modeling in Example 10.5). □

## 8.7 Multinomial Models with Hierarchically Structured Response

In some applications the response categories have some inherent grouping. For example, when the response categories in a clinical study are given by {no infection, infection type I, infection type II}, it is natural to consider the two latter response categories as one group {infection}. When investigating the effect of the predictors on these responses, one might want to take

TABLE 8.5: Estimated coefficients for addiction survey with quadratic effect of age.

| | Intercept | Gender | Univ | Age | Age$^2$ |
|---|---|---|---|---|---|
| | | | Estimates | | |
| 1 | −3.720 | 0.526 | 1.454 | 0.184 | −0.002 |
| 2 | −3.502 | 0.356 | 0.936 | 0.135 | −0.001 |
| | | | Standard Errors | | |
| 1 | 0.546 | 0.201 | 0.257 | 0.029 | 0.0003 |
| 2 | 0.596 | 0.224 | 0.290 | 0.030 | 0.0003 |

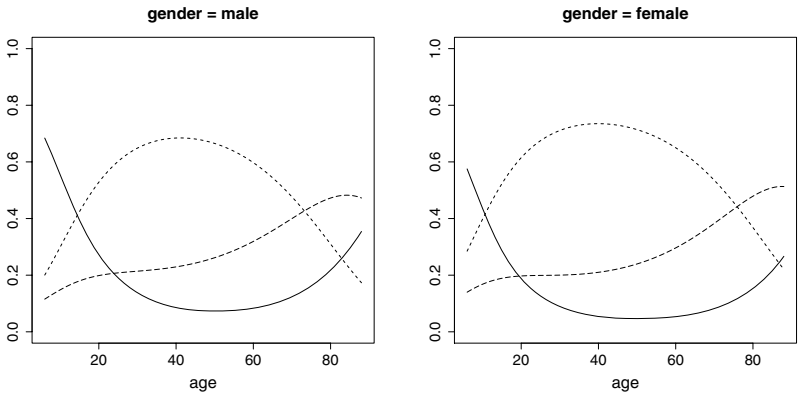


FIGURE 8.6: Estimated probabilities for addiction data with quadratic effect of age (category 0: solid line; category 1: dotted line; category 2: dashed line).

the similarities between the response categories into account to obtain effects with a simple interpretation.

A hierarchical model is obtained by first modeling the response in groups of homogeneous response categories, and in a second step the response within the groups is modeled. In general, let the response categories $K = \{1, \ldots, k\}$ be subdivided into basic sets $S_1, \ldots, S_m$, where $K = S_1 \cup \cdots \cup S_m$. In the first step, let the logit model be

$$P(Y \in S_t | \boldsymbol{x}) = \frac{\exp(\boldsymbol{x}^T \boldsymbol{\beta}_t)}{\sum_{s \in K} \exp(\boldsymbol{x}^T \boldsymbol{\beta}_s)}.$$

In the second step, the conditional response given $S_t$ is modeled as

$$P(Y = r | Y \in S_t, \boldsymbol{x}) = \frac{\exp(\boldsymbol{x}^T \boldsymbol{\beta}_r^{(S_t)})}{\sum_{s \in S_t} \exp(\boldsymbol{x}^T \boldsymbol{\beta}_s^{(S_t)})}.$$

The parameters have to be restricted by side constraints on two levels, for the responses in $S_1, \ldots, S_m$ and for the conditional response. For example, one might use $S_m$ as a reference on the first level by setting $\boldsymbol{\beta}_m$ to zero and choose one category from each set $S_t$ as a reference on the second level by setting $\boldsymbol{\beta}_{r_t}$ to zero for one category $r_t \in S_t$.

For the derivation of the maximum likelihood estimates one needs the marginal probability of response on category $r$, which for $r \in S_t$ has the simple form

$$P(Y = r | \boldsymbol{x}) = P(Y = r | Y \in S_t) P(Y \in S_t).$$

Therefore, assuming multinomially distributed responses $\boldsymbol{y}_i \sim \mathrm{M}(n_i, \boldsymbol{\pi}_i)$, $i = 1, \ldots, N$, $\boldsymbol{\pi}_i = \boldsymbol{\pi}(\boldsymbol{x}_i)$, the log-likelihood is

$$l = \sum_{i=1}^{N} \sum_{r=1}^{k} y_{ir} \log(\pi_{ir}) = \sum_{i=1}^{N} \sum_{t=1}^{m} \sum_{r \in S_t} y_{ir} \log(\pi_{ir}).$$

It decomposes into $l = l_g(\{\boldsymbol{\beta}_t\}) + \sum_{t=1}^{m} l_t(\{\boldsymbol{\beta}_r^{(S_t)}\})$, where

$$l_g(\{\boldsymbol{\beta}_t\}) = \sum_{i=1}^{N} \sum_{t=1}^{m} y_{iS_t} \log(P(Y_i \in S_t)),$$

with $y_{iS_t} = \sum_{r \in S_t} y_{ir}$, is the log-likelihood for the grouped observations on the first level, depending only on $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_m$, and

$$l_t(\{\boldsymbol{\beta}_r^{(S_t)}\}) = \sum_{i=1}^{N} \sum_{r \in S_t} y_{ir} \log(P(Y_i \in r | Y_i \in S_t))$$

is the log-likelihood for responses within $S_t$ depending on $\{\boldsymbol{\beta}_r^{(S_t)}, r \in S_t\}$. Since the log-likelihood decomposes into additive terms, each part of the log-likelihood may be fitted separately by fitting the corresponding logit model. Likelihood ratio tests for individual parameters apply on the level of each model. However, Wald tests, which are typically used to examine single parameters, and the corresponding $p$-values are not trustworthy because they are based only on the components of the total model (Exercise 8.8).

It should be noted that the assumption of a logit model on both levels yields a model that is not equivalent to a one-step logit model. If a logit model holds for all categories, one easily derives that the conditional model $P(Y = r | Y \in S_t, \boldsymbol{x})$ is again a logit model, but that does not hold for the probabilities $P(Y \in S_t | \boldsymbol{x})$.

### Example 8.6: Addiction

We refer again to the addiction data (Example 8.5). The response was in three categories, 0: addicts are weak-willed; 1: addiction is a disease; 2: both, which are grouped into $S_1 = \{0, 1\}, S_2 = \{2\}$. Therefore, the binary logit model that distinguishes between $S_1$ and $S_2$ models if the respondents think that a single cause is responsible for addiction. In the second step, the logit model compares category 1 to category 0, given that the response is in categories $\{0, 1\}$. From the estimates in Tables 8.6 and 8.7 it is seen that the covariates have no effect on the distinction between categories $\{0, 1\}$ and $\{2\}$ but have strong effects on the distinction between causes given that they think that one cause is behind addiction. □

TABLE 8.6: Estimated probabilities for addiction data with response in category 2 compared to categories $\{0, 1\}$.

| | Estimate | Std. Error | $z$-Value | Pr(>\|$z$\|) |
|---|---|---|---|---|
| Intercept | 2.1789 | 0.5145 | 4.23 | 0.000 |
| gender | −0.0172 | 0.1828 | −0.09 | 0.925 |
| university | 0.0895 | 0.2067 | 0.43 | 0.665 |
| age | −0.0342 | 0.0255 | −1.34 | 0.179 |
| age2 | 0.0001 | 0.0003 | 0.45 | 0.650 |

TABLE 8.7: Estimated probabilities for addiction data comparing category 1 to category 0, given the response is in categories $\{0, 1\}$.

| | Estimate | Std. Error | $z$-Value | Pr(>\|$z$\|) |
|---|---|---|---|---|
| Intercept | −3.5468 | 0.5443 | −6.52 | 0.000 |
| gender | 0.5433 | 0.2055 | 2.64 | 0.008 |
| university | 1.4656 | 0.2601 | 5.64 | 0.000 |
| age | 0.1720 | 0.0284 | 6.06 | 0.000 |
| age2 | −0.0017 | 0.0003 | −5.19 | 0.000 |

## 8.8 Discrete Choice Models

Since many economic decisions involve choice among discrete alternatives, probabilistic choice models have become an important research area in contemporary econometrics. In contrast to earlier approaches to model demand on an aggregate level, probabilistic choice models focus on the modeling of individual behavior. In the following some basic concepts are given. They provide a motivation for the multinomial logit model but also give rise to alternative models. More extensive treatments of probabilistic choice have been given, for example, by McFadden (1981).

Let $K = \{1, \ldots, k\}$ denote the total set of alternatives. For a subset $B \subset K$ of available alternatives let $P_B(r)$ denote the probability of choosing $r \in B$, given that a selection must be made from set $B$. A *probabilistic choice system* may be described as a tupel:

$$(K, \mathcal{B}, \{P_B, B \in \mathcal{B}\}),$$

where $\mathcal{B}$ is a family of subsets from $K$. A simple example is the one-member family $\mathcal{B} = \{K\}$. Then one considers only selections from the full set of alternatives $K$. In a pair comparison system with $\mathcal{B} = \{\{i, j\}, i, j \in K\}\}$, the selection is among two alternatives, where all combinations of alternatives are presented. In a complete choice experiment $\mathcal{B}$ contains all subsets $K \subset \mathcal{B}$ with $|K| \geq 2$.

A probabilistic choice system is called a *random utility model* if there exist random variables $U_r, r \in K$, such that

$$P_B(r) = P(U_r = \max_{s \in B}\{U_s\}). \tag{8.10}$$

The random utility $U_r$ represents the utility attached to alternative $r$. Usually it is a latent variable that cannot be measured directly. Its interpretation depends on the application; it may be the subjective utility of a travel mode or of a brand in commodity purchases. If the choice probabilities have a representation (8.10), one says that the *random utility maximization hypothesis* holds.

Let the random utility $U_r$ have the form $U_r = u_r + \varepsilon_r$, where $u_r$ is the structural part or non-random fixed utility and $\varepsilon_r$ is a noise variable. The fixed utility $u_r$ is determined by the characteristics of the decision maker and the attributes of the alternatives, while the random variable $\varepsilon_r$ represents the residual variation.

### Random Utility Models

Equation (8.10) may also be seen as a way of constructing choice probabilities. Let us consider a set $B = \{i_1, \ldots, i_m\}, i_1 < \cdots < i_m$. Then the choice probabilities $P(r) = P_B(r)$ are

obtained as

$$\begin{aligned} P_B(i_r) &= P(U_{i_r} \geq U_j \quad \text{for} \quad j \in B) \\ &= P(U_{i_r} \geq U_{i_1}, \ldots, U_{i_r} \geq U_{i_m}) \\ &= P(u_{i_r} - u_{i_1} \geq \varepsilon_{i_1} - \varepsilon_{i_r}, \ldots, u_{i_r} - u_{i_m} \geq \varepsilon_{i_m} - \varepsilon_{i_r}) \\ &= \int_{-\infty}^{u_{i_r}-u_{i_1}} \cdots \int_{-\infty}^{u_{ir}-u_{i_m}} f_{B,r}(\varepsilon_{i_1 i_r}, \ldots, \varepsilon_{i_m i_r}) d\,\varepsilon_{i_1 i_r} \ldots d\,\varepsilon_{i_m i_r} \\ &= F_{B,r}(u_{i_r} - u_{i_1}, \ldots, u_{i_r} - u_{i_m}), \end{aligned} \tag{8.11}$$

where $\boldsymbol{\varepsilon}_{B,r}^T = (\varepsilon_{i_1 i_r}, \ldots, \varepsilon_{i_m i_r})$, with $\varepsilon_{i_s i_r} = \varepsilon_{i_s} - \varepsilon_{i_r}$, is the $(m-1)$-dimensional vector of differences; $f_{B,r}$ is the density of $\boldsymbol{\varepsilon}_{B,r}$; and $F_{B,r}$ is the cumulative distribution function of $\boldsymbol{\varepsilon}_{B,r}$.

Let us consider the simple case where the full set of alternatives $K$ forms the choice set. Then any distribution of $\boldsymbol{\varepsilon}^T = (\varepsilon_1, \ldots, \varepsilon_k)$ will generate a discrete choice model. A familiar model results when one assumes that $\varepsilon_1, \ldots, \varepsilon_k$ are iid variables with marginal distribution function $F(x) = \exp(-\exp(-x))$, which is the Gumbel or maximum extreme value distribution. Then one obtains for the cumulative distribution function of the differences $(\varepsilon_1 - \varepsilon_r, \ldots, , \varepsilon_k - \varepsilon_r)$

$$F_{m-1}(x_1, \ldots, x_{m-1}) = \frac{1}{1 + \sum_{i=1}^{m-1} \exp(-x_i)}$$

and therefore the logit model

$$P_K(r) = \frac{1}{1 + \sum_{j \neq r} \exp(-(u_r - u_j))} = \frac{\exp(u_r)}{\sum_{j=1}^{k} \exp(u_j)}$$

(e.g. Yellott, 1977). One obtains the familiar multinomial logit model with predictors by assuming that one has a vector $\boldsymbol{x}$ that characterizes the decision maker and the attributes $\boldsymbol{w}_r$ connected to alternative $r$ that form the latent fixed utility:

$$u_r = \boldsymbol{x}^T \boldsymbol{\gamma}_r + \boldsymbol{w}_r^T \boldsymbol{\alpha}.$$

Then the model has the familiar form

$$\begin{aligned} \log\left(\frac{P_K(r)}{P_K(k)}\right) = u_r - u_k &= \boldsymbol{x}^T(\boldsymbol{\gamma}_r - \boldsymbol{\gamma}_k) + (\boldsymbol{w}_r - \boldsymbol{w}_k)^T \boldsymbol{\alpha} \\ &= \boldsymbol{x}^T \boldsymbol{\beta}_r + (\boldsymbol{w}_r - \boldsymbol{w}_k)^T \boldsymbol{\alpha}, \end{aligned}$$

which is equivalent to the logit model considered in Section 8.4.

Random utility models that assume iid distributions for the noise variables have a long history. In psychophysics, Thurstone (1927) proposed the law of comparative judgement, which is based on normally distributed variables. Generally, in the iid case, all differences have the same distribution and one obtains with $\sigma_0^2 = \text{var}(\varepsilon_i)$ for the covariance of differences:

$$\begin{aligned} \text{cov}(\varepsilon_i - \varepsilon_r, \varepsilon_j - \varepsilon_r) &= \text{E}(\varepsilon_i - \varepsilon_r)(\varepsilon_j - \varepsilon_r) \\ &= \text{E}(\varepsilon_i \varepsilon_j - \varepsilon_i \varepsilon_r - \varepsilon_r \varepsilon_j + \varepsilon_r^2) = \text{E}(\varepsilon_r^2) = \sigma_0^2. \end{aligned}$$

Then the covariance matrix of the differences is given by $\boldsymbol{\Sigma}_0 = \sigma_0^2(\mathbf{I} + \mathbf{1}\mathbf{1}^T)$. If one assumes that $\varepsilon_i$ is normally distributed with $\varepsilon_i \sim \text{N}(0, \sigma_0^2)$, one obtains from maximizing the random utility the *multinomial probit model*

$$P_K(r) = \phi_{\mathbf{0}, \boldsymbol{\Sigma}_0}(u_r - u_1, \ldots, u_r - u_k),$$

where $\phi_{\mathbf{0},\boldsymbol{\Sigma}_0}$ is the $(k-1)$-dimensional cumulative distribution function of the normal distribution $\mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma}_0)$. In psychology, the model is also known as *Thurstone's case V*. The multinomial probit model is harder to use for higher numbers, of alternatives because the integral has to be computed numerically.

## Independence from Irrelevant Alternatives

Simple models like the multinomial logit model imply a property that has caused some discussion in economics. The problem occurs if decisions for more than just one fixed set of alternatives are investigated. For example, in a complete choice experiment, the choice probabilities are determined for every subset of alternatives. Then the logit model for subset $B$, derived from the maximization of random utilities $U_r$, has the form

$$P_B(r) = \frac{\exp(u_r)}{\sum_{j\in B}\exp(u_j)},$$

and one obtains for any subset $B, C, r, s \in B \cap C$

$$\frac{P_B(r)}{P_B(s)} = \frac{P_C(r)}{P_C(s)} = \frac{\exp(u_r)}{\exp(u_s)}. \tag{8.12}$$

This means that the proportion of probabilities is identical when different sets of alternatives are considered. The system of choice probabilities satisfies Luce's Choice Axiom (Luce, 1959), which implies that the choice probabilities are *independent from irrelevant alternatives*.

McFadden (1986) calls the independence from irrelevant alternatives a blessing and a curse. An advantage is that if it holds, it makes it possible to infer choice behavior with multiple alternatives using data from simple experiments like paired comparisons. A disadvantage is that it is a rather strict assumption that may not hold for heterogeneous patterns of similarities encountered in economics. A famous problem that illustrates the case is the "red bus–blue bus" problem that has been used by McFadden (see, e.g., Hausman and Wise, 1978). Suppose a commuter has the initial alternatives of driving or taking a red bus with the odds given by

$$\frac{P_{\{\text{driving,red bus}\}}(\text{driving})}{P_{\{\text{driving,red bus}\}}(\text{red bus})} = 1.$$

Then an additional alternative becomes available, namely, a blue bus that is identical in all respects to the red bus, except color. If the logit model holds, it is seen from (8.12) that the odds of choosing the driving alternative over the red bus remain the same. Since the odds for the choice between the red and the blue bus are

$$\frac{P_{\{\text{red bus, blue bus}\}}(\text{red bus})}{P_{\{\text{red bus, blue bus}\}}(\text{blue bus})} = 1,$$

one obtains for any $B$ for all paired comparisons

$$\frac{P_B(\text{driving})}{P_B(\text{red bus})} = \frac{P_B(\text{driving})}{P_B(\text{blue bus})} = \frac{P_B(\text{red bus})}{P_B(\text{blue bus})} = 1$$

and therefore

$$P_{\{1,2,3\}}(\text{driving}) = P_{\{1,2,3\}}(\text{red bus}) = P_{\{1,2,3\}}(\text{blue bus}) = 1/3.$$

This is a counterintuitive result because the additional "irrelevant" alternative blue bus has decreased the choice probability of driving substantially. Problems of this type occur not only for

the logit model. The probit model based on iid distributions has a similar property as demonstrated by Hausman and Wise (1978). In fact, the same sort of counterintuitive results are found for all choice systems that share a property called *simple scalability* (Krantz, 1964). Simple scalability means that there exist scales $v_1, \ldots, u_k$ and functions $F_2, \ldots, F_k$ that determine the choice probability for $B = \{i_1, \ldots, i_m\}$ by

$$P_B(r) = F_m(v_{i_r}, \ldots, v_{i_m}),$$

where $F_m$ is strictly increasing in the first argument and strictly decreasing in the remaining $m - 1$ arguments. It is easily shown that simple scalability holds for the multinomial logit model.

Tversky (1972) shows that simple scalability is equivalent to order independence that holds whenever for all $r, s \in B \setminus C$ and $t \in C$

$$P_B(r) \geq P_B(s) \qquad \Leftrightarrow \qquad P_{C\cup\{r\}}(t) \leq P_{C\cup\{s\}}(t).$$

This is a weaker version of the independence of irrelevant alternatives, which implies that only the order of $P_B(r)$ and $P_B(s)$, and not necessarily their ratio, is independent of $B$.

The independence of irrelevant alternatives raises problems when one wants to combine results from different choice sets, which is frequently wanted in econometric applications. However, if the choice set is fixed, and each person faces the full set of alternatives, the counterintuitive results have no relevance. For the simultaneous treatment of different choice sets, the most widely used model in econometrics is McFadden's nested multinomial model, which is sketched in the following section.

## Pair Comparison Models

In pair comparison systems only two alternatives are compared at a time. Therefore, the family of subsets that is considered is given by $\mathcal{B} = \{\{i, j\}, i, j \in K\}\}$. The resulting pair comparison models are useful in psychometrics to scale stimuli or in marketing to scale the attractiveness of product brands. Moreover, it is often used in sport competitions when one wants to measure the ability of a team or a player. The most frequently used model is the logistic model

$$P_{\{r,s\}}(r) = \frac{\exp(u_r - u_s)}{1 + \exp(u_r - u_s)}, \tag{8.13}$$

which results from the assumption of a Gumbel distribution for the residual variation. It is also called the *BTL (Bradley-Terry-Luce) model*, with reference to Bradley and Terry (1952) and Luce (1959).

If $u_r > u_s$, the probability that stimuli $r$ is preferred over $s$ (or that player $r$ will win against player $s$) increases with the difference of attractiveness (or ability) $u_r - u_s$. Since ties are not allowed, in the simple model the probability is $0.5$ if $u_r = u_s$. The advantage of the model is that one obtains the attractiveness of stimuli on a one-dimensional scale, and estimates can be used to predict the future outcome. When one assumes that the comparisons are independent, simple logit models apply for estimation. The set of stimuli can be treated as a factor and the design matrix specifies the differences between the alternatives. Of course a reference alternative has to be chosen, for example, by setting $u_1 = 0$.

The model is easily extended to incorporate an order effect. By specifying

$$P_{\{r,s\}}(r) = \frac{\exp(\alpha + u_r - u_s)}{1 + \exp(\alpha + u_r - u_s)}, \tag{8.14}$$

one obtains $P_{\{r,s\}}(r) = \exp(\alpha)/1 + \exp(\alpha)$ for the preference of $r$ over $s$. In sports competitions, $\alpha$ represents the home advantage; when stimuli are rated it refers to the order in which the stimuli are presented. Moreover, category-specific variables can be included in the model; for literature see Section 8.11.

### Example 8.7: Paired Comparison

Rumelhart and Greeno (1971) asked 234 college students for their preferences concerning nine famous persons. For each of the 36 pairs the subjects were instructed to choose the person with whom they would rather spend an hour discussing a topic of their choosing. Table 8.8 shows the data, and Table 8.9 gives the estimates of the fitted BTL model. One might want to distinguish between the effect of the profession and the effect of the person by including profession in the predictor. Let the scale value be structured as $u_r = w_{r1}\gamma_r + w_{r2}\gamma_2 + \delta_r$, where $w_{r1} = 1$ if person $r$ is a politician (0 otherwise), and $w_{r2} = 1$ if person $r$ is a sportsman (0 otherwise). Actor is the reference category and $\delta_r$ is the additional effect of the person with reference $\delta_2 = 0$ for politicians, $\delta_5 = 0$ for sportsmen, and $\delta_9 = 0$ for actors. Table 8.10 shows that sportsmen are distinctly preferred over actors, but not politicians. □

TABLE 8.8: Pair comparison referring to politicians Harold Wilson (1), Charles de Gaulle (2), and Lyndon B. Johnson (3); sporstmen Johnny Unitas (4), Carl Yastrazemski (5), and A. Foyt (6); and actors Brigitte Bardot (7), Elizabeth Taylor (8), and Sophia Loren (9).

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | Σ |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | – | 159 | 163 | 175 | 183 | 179 | 173 | 160 | 142 | 1334 |
| 2 | 75 | – | 138 | 164 | 172 | 160 | 156 | 122 | 122 | 1109 |
| 3 | 71 | 96 | – | 145 | 157 | 140 | 138 | 122 | 120 | 989 |
| 4 | 59 | 70 | 89 | – | 176 | 115 | 124 | 86 | 61 | 780 |
| 5 | 51 | 62 | 77 | 58 | – | 77 | 95 | 72 | 61 | 553 |
| 6 | 55 | 74 | 94 | 119 | 157 | – | 134 | 92 | 71 | 796 |
| 7 | 61 | 78 | 96 | 110 | 139 | 100 | – | 67 | 48 | 699 |
| 8 | 74 | 112 | 112 | 148 | 162 | 142 | 167 | – | 87 | 1004 |
| 9 | 92 | 112 | 114 | 173 | 173 | 163 | 186 | 147 | – | 1160 |

TABLE 8.9: Estimated coefficients for pair comparison.

| | Persons | $u_i$ | Standard Deviation |
|---|---|---|---|
| | 1 WI | -0.382 | 0.066 |
| politicians | 2 GA | 0.106 | 0.064 |
| | 3 JO | 0.350 | 0.064 |
| | 4 UN | 0.772 | 0.065 |
| sportsmen | 5 YA | 1.260 | 0.067 |
| | 6 FO | 0.739 | 0.065 |
| | 7 BB | 0.940 | 0.065 |
| actors | 8 ET | 0.319 | 0.066 |
| | 9 SL | 0.000 | 0.064 |