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 (2×2)-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.

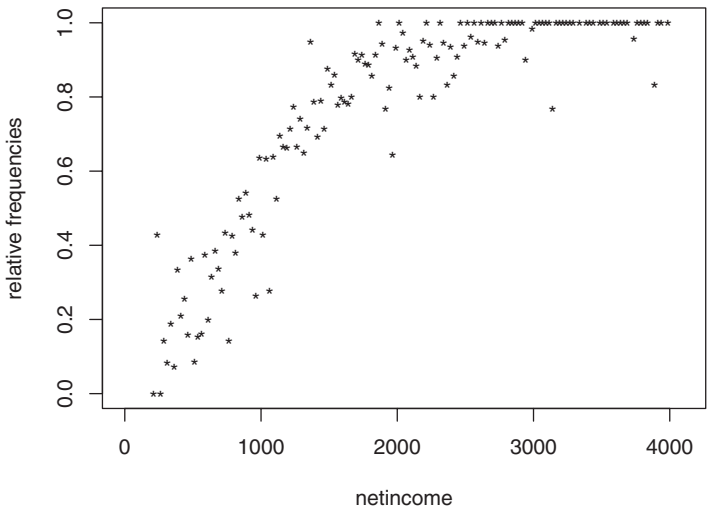

FIGURE 1.1: Car data, relative frequencies within intervals of length 50, plotted against net income in Euros.

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 μ (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 μ 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.

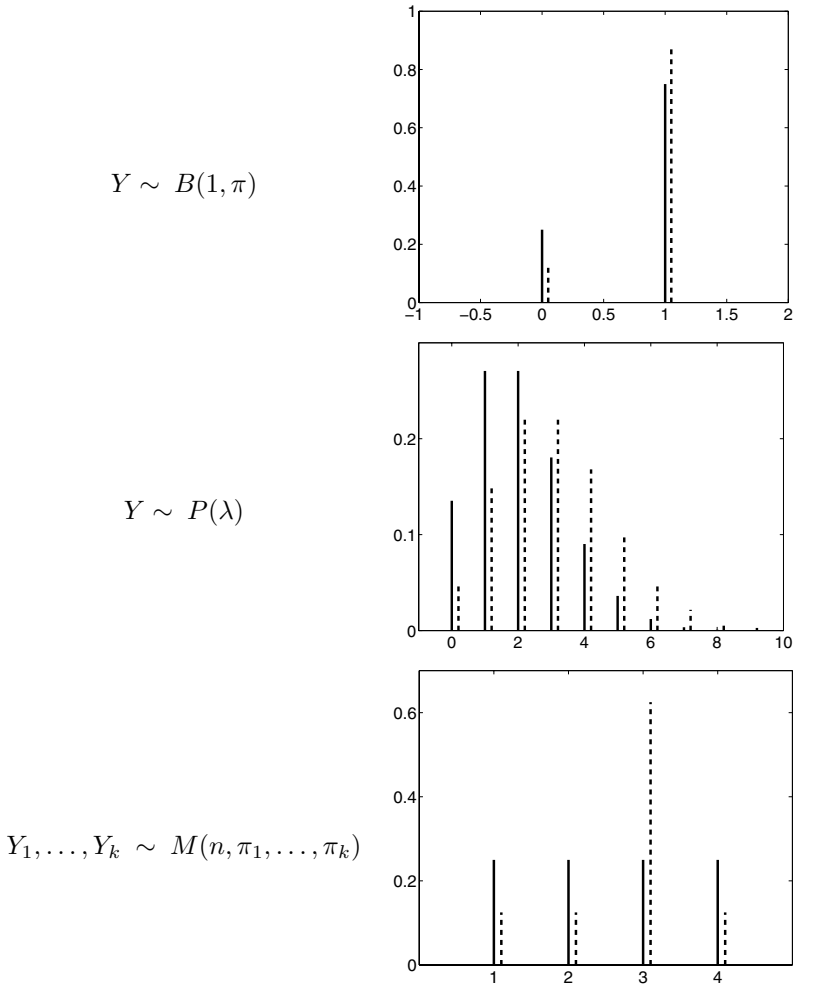

FIGURE 1.2: Binomial, Poisson, and multinomial distributions. Each panel shows two different distributions.

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.

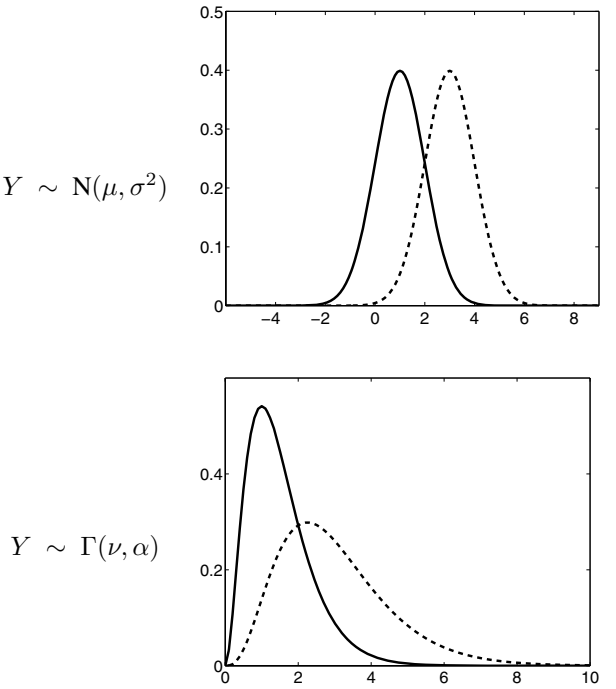

FIGURE 1.3: Normal and Gamma-distributions.

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).

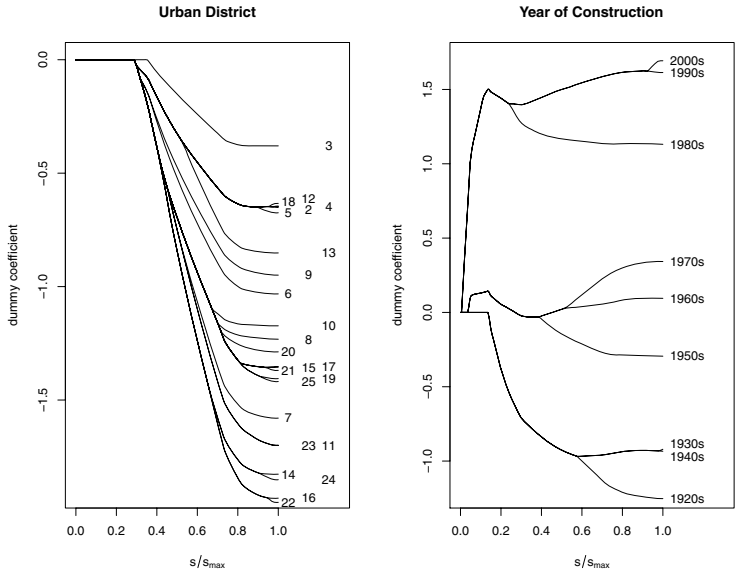

FIGURE 1.4: Effects of urban district and year of construction (in decades) on rent per square meter.

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 μ 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).

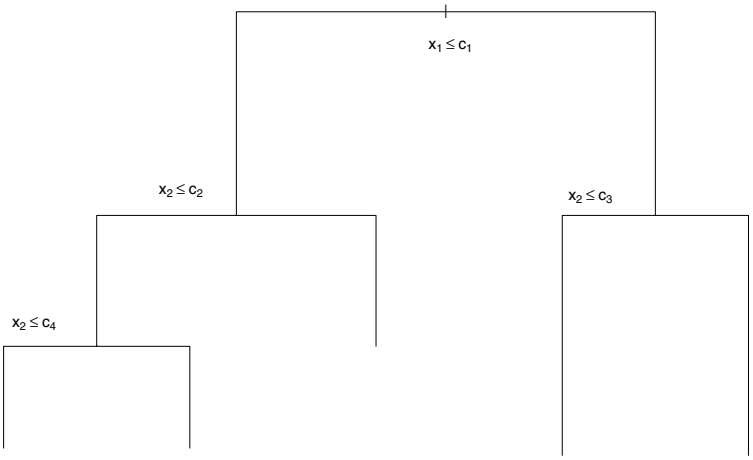

FIGURE 1.5: Example tree for two variables.

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 ε 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 β_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 β_A are dollars per year and β_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 β_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, β_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 β_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:

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

β_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 β_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 β_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 ε. 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 σ^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 σ^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 σ^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 σ^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 $\mathrm{var}(\tilde{r}_i) = \sigma^2$. 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

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

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 ith row; in $\boldsymbol{\mu}_{(i)}, \boldsymbol{y}_{(i)}$ the ith 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 ith 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 $\mathrm{var}(r_{(i)}) = \sigma^2 + \sigma^2\boldsymbol{x}_i^T(\boldsymbol{X}_{(i)}^T\boldsymbol{X}_{(i)})^{-1}\boldsymbol{x}_i = \sigma^2(1+h_{(i)})$, 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 σ^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 σ^2, which does not depend on the ith 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 jth 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 χ^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 χ^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 β_i and -1 corresponds to β_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 χ^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 β that minimizes the least-squares criterion.

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