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.

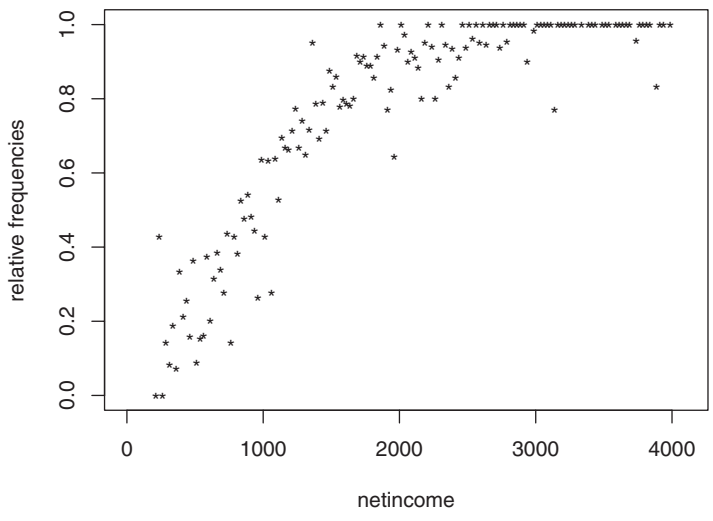


FIGURE 1.1: Car data, relative frequencies within intervals of length 50, plotted against net income in Euros.

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.

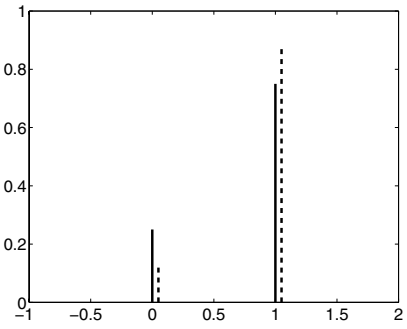
Sometimes the assumption of a specific distribution, even if it reflects the type of data collected, is too strong to explain the variability in responses satisfactorily. In practice, one often finds that count data and relative frequencies are more variable than is to be expected under the Poisson and 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.

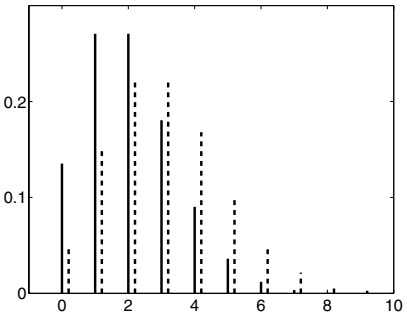
Structuring the Influential Term

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.

$Y \sim B(1, \pi)$



$Y \sim P(\lambda)$



$Y_1, \ldots, Y_k \sim M(n, \pi_1, \ldots, \pi_k)$

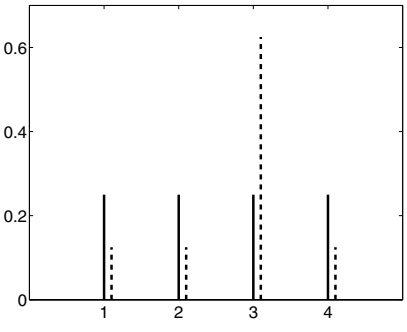


FIGURE 1.2: Binomial, Poisson, and multinomial distributions. Each panel shows two different distributions.

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.

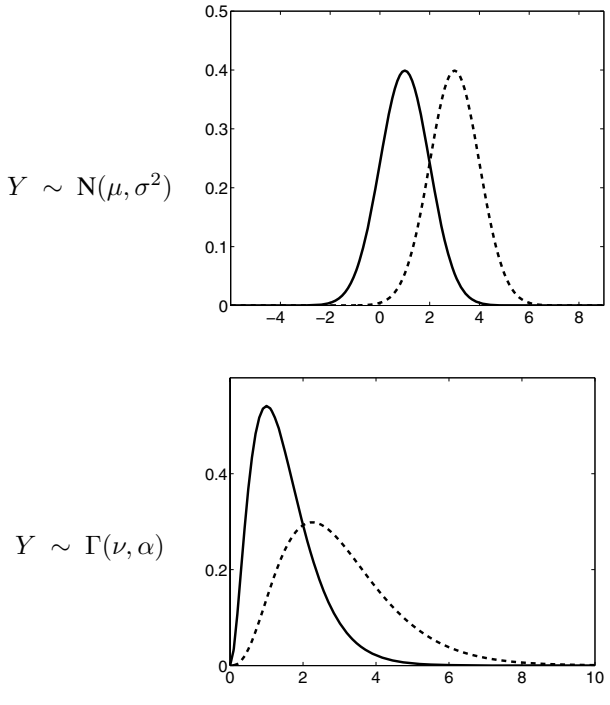


FIGURE 1.3: Normal and Gamma-distributions.

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).

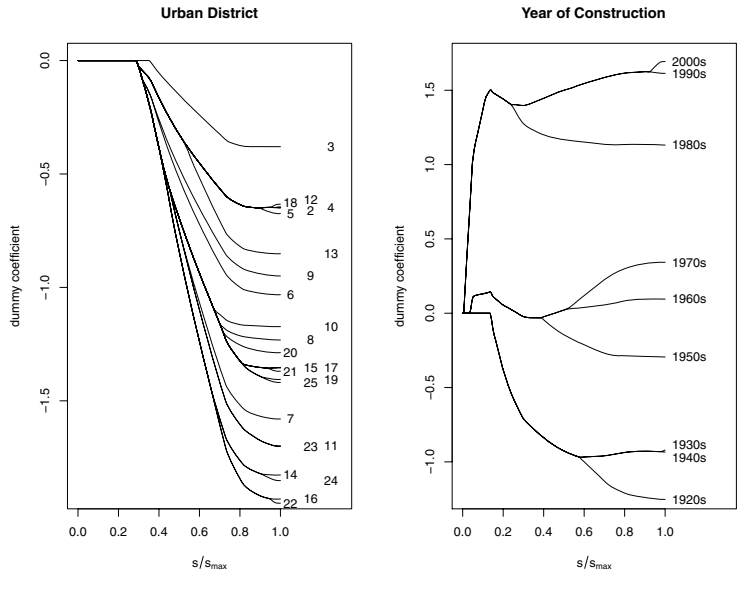


FIGURE 1.4: Effects of urban district and year of construction (in decades) on rent per square meter.

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

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