

2.1 Introduction

All case studies that have been discussed in Chap. 1 have one main feature in common: We aim at modeling the effect of a given set of explanatory variables x_1, \dots, x_k on a variable y of primary interest. The variable of primary interest y is called *response* or *dependent variable* and the explanatory variables are also called *covariates*, *independent variables*, or *regressors*. The various models differ mainly through the type of response variables (continuous, binary, categorical, or counts) and the different kinds of covariates, which can also be continuous, binary, or categorical. In more complex situations, it is also possible to include time scales, variables to describe the spatial distribution or geographical location, or group indicators.

A main characteristic of regression models is that the relationship between the response variable y and the covariates is not a deterministic function $f(x_1, \dots, x_k)$ of x_1, \dots, x_k (as often is the case in classical physics), but rather shows random errors. This implies that the response y is a random variable, whose distribution depends on the explanatory variables. Galton's data set on heredity exemplified that, even though we know the exact height of the parents, we are unable to predict the exact height of their children. We can rather only estimate the *average size of the children* and the degree of dispersion from the mean value. Similar statements are valid for all problems discussed in Chap. 1. One main goal of regression is to analyze the influence of the covariates on the mean value of the response variable. In other words, we model the (conditional) expected value $E(y \mid x_1, \dots, x_k)$ of y depending on the covariates. Hence, the expected value is a function of the covariates:

$$E(y \mid x_1, \dots, x_k) = f(x_1, \dots, x_k).$$

It is then possible to decompose the response into

$$y = E(y \mid x_1, \dots, x_k) + \varepsilon = f(x_1, \dots, x_k) + \varepsilon,$$

where ε is the random deviation from the expected value. The expected value $E(y | x_1, \dots, x_k) = f(x_1, \dots, x_k)$ is often denoted as the *systematic component* of the model. The random deviation ε is also called *random or stochastic component*, *disturbance*, or *error term*. In many regression models, in particular in the classical linear model (see Sect. 2.2 and Chap. 2), it is assumed that the error term does not depend on covariates. This may not be true, however, in general. The primary goal of regression analysis is to use the data $y_i, x_{i1}, \dots, x_{ik}$, $i = 1, \dots, n$, to estimate the systematic component f , and to separate it from the stochastic component ε .

The most common class is the linear regression model given by

$$y = \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k + \varepsilon.$$

Here, the function f is linear so that

$$E(y | x_1, \dots, x_k) = f(x_1, \dots, x_k) = \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k$$

holds, i.e., the (conditional) mean of y is a linear combination of the covariates. Inserting the data yields the n equations

$$y_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_k x_{ik} + \varepsilon_i, \quad i = 1, \dots, n,$$

with unknown parameters or regression coefficients β_0, \dots, β_k . The linear regression model is especially applicable when the response variable y is continuous and shows an approximately normal distribution (conditional on the covariates). More general regression models are, e.g., required when either the response variable is binary, the effect of covariates is nonlinear, or if spatial or cluster-specific heterogeneity has to be considered. Starting from the classical linear regression model, the following sections of this chapter describe regression models of increasing flexibility and complexity. Examples taken from various fields of application provide an overview of their usefulness. A more detailed presentation of the different regression models, and especially the corresponding statistical inference techniques, will be given in the chapters to follow.

2.2 Linear Regression Models

2.2.1 Simple Linear Regression Model

Example 2.1 Munich Rent Index—Simple Linear Regression

We start by analyzing only the subset of apartments built after 1966. This sample is divided into three location strata: average, good, and top location. The left panel of Fig. 2.1 shows the scatter plot between the response variable net rent and the explanatory variable living area for apartments in average location. The scatter plot displays an approximate linear relationship between *rent* and *area*, i.e.,

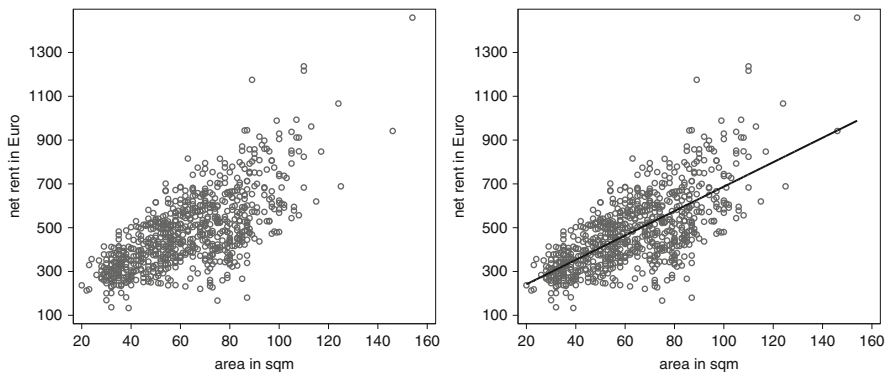


Fig. 2.1 Munich rent index: scatter plot between net rent and area for apartments in average location built after 1966 (*left panel*). In the *right panel*, a regression line is additionally included

$$rent_i = \beta_0 + \beta_1 \cdot area_i + \varepsilon_i. \quad (2.1)$$

The errors ε_i are random deviations from the regression line $\beta_0 + \beta_1 \cdot area$. Since systematic deviations from zero are already included in the parameter β_0 , $E(\varepsilon_i) = 0$ can be assumed. An alternative formulation of Eq. (2.1) is

$$E(rent | area) = \beta_0 + \beta_1 \cdot area.$$

This means that the expected net rent is a linear function of the living area. △

The example is a special case of the simple linear regression model

$$y = \beta_0 + \beta_1 x + \varepsilon,$$

where the expected value $E(y | x) = f(x)$ is assumed to be linear in the general relationship

$$y = f(x) + \varepsilon = E(y | x) + \varepsilon.$$

This implies that $E(y | x) = f(x) = \beta_0 + \beta_1 x$. More specifically, for the standard model of a simple linear regression, we assume

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, \dots, n, \quad (2.2)$$

with independent and identically distributed errors ε_i , such that

$$E(\varepsilon_i) = 0 \quad \text{and} \quad \text{Var}(\varepsilon_i) = \sigma^2.$$

The property of constant variance σ^2 across errors ε_i is also called *homoscedasticity*. In particular, this implies that the errors are independent of the covariates. When constructing confidence intervals and statistical tests, it is convenient if the

2.1 Standard Model of Simple Linear Regression

Data

$(y_i, x_i), i = 1, \dots, n$, with continuous variables y and x .

Model

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, \dots, n.$$

The errors $\varepsilon_1, \dots, \varepsilon_n$ are independent and identically distributed (i.i.d.) with

$$E(\varepsilon_i) = 0, \quad \text{Var}(\varepsilon_i) = \sigma^2.$$

We can interpret the estimated regression line $\hat{f}(x) = \hat{\beta}_0 + \hat{\beta}_1 x$ as an estimate $E(y|x)$ for the conditional expected value of y given the covariate value x . We can, thus, predict y through $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$.

additional assumption of Gaussian errors is reasonable:

$$\varepsilon_i \sim N(0, \sigma^2).$$

In this case, the observations of the response variable follow a (conditional) normal distribution with

$$E(y_i) = \beta_0 + \beta_1 x_i, \quad \text{Var}(y_i) = \sigma^2,$$

and the y_i are (conditionally) independent given covariate values x_i . The unknown parameters β_0 and β_1 are estimated according to the method of least squares (LS): the estimated values $\hat{\beta}_0$ and $\hat{\beta}_1$ are determined as the minimizers of the sum of the squared deviations

$$LS(\beta_0, \beta_1) = \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2$$

for given data $(y_i, x_i), i = 1, \dots, n$. Section 3.2 will present the method of least squares in detail. Inserting $\hat{\beta}_0, \hat{\beta}_1$ into the conditional mean, the estimated regression line

$$\hat{f}(x) = \hat{\beta}_0 + \hat{\beta}_1 x$$

results. The regression line is to be understood as an estimate $\widehat{E(y|x)}$ for the conditional mean of y given the covariate value x . Thus, the regression line can also be used to predict y for a given x . The predicted value of y is usually denoted by \hat{y} , i.e., $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$.

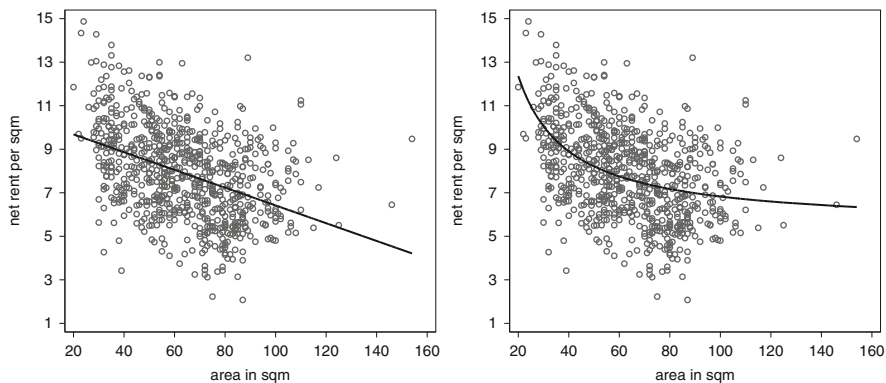


Fig. 2.2 Munich rent index: scatter plots between net rent per square meter and area. Included is the estimated effect \hat{f} of living area for a linear (*left*) and reciprocal (*right*) area effect

Example 2.2 Munich Rent Index—Simple Linear Regression

We illustrate the simple linear regression model using the data shown in Fig. 2.1 and the corresponding model (2.1). The data gives rise to doubts about the assumption of equal variances $\text{Var}(\varepsilon_i) = \text{Var}(y_i) = \sigma^2$ across observations since variability in rent seems to increase as living area increases. For the moment, we will ignore this problem, but Sect. 4.1.3 will illustrate how to deal with problems associated with unequal variances. See also Sect. 2.9.2 and Chap. 10 on quantile regression. According to the method of least squares, the parameter estimates for model (2.1) are $\hat{\beta}_0 = 130.23$ and $\hat{\beta}_1 = 5.57$ implying the estimated regression line

$$\hat{f}(\text{area}) = 130.23 + 5.57 \cdot \text{area}$$

illustrated in the right panel of Fig. 2.1. The slope parameter $\hat{\beta}_1 = 5.57$ can be interpreted as follows: If the living area increases by 1 m², the rent increases about 5.57 Euro on average.

If we choose the rent per square meter instead of the rent as response variable, the scatter plot illustrated in Fig. 2.2 (left) results. It is quite obvious that the relationship between rent per square meter and living area is nonlinear. This is also supported by the estimated regression line

$$\hat{f}(\text{area}) = 10.5 - 0.041 \cdot \text{area}.$$

The fit to the data is poor for small and large living area. A better fit can be achieved by defining the new explanatory variable

$$x = \frac{1}{\text{area}}$$

that yields a simple linear regression of the form

$$\text{rentsqm}_i = \beta_0 + \beta_1 x_i + \varepsilon_i = \beta_0 + \beta_1 \frac{1}{\text{area}_i} + \varepsilon_i. \quad (2.3)$$

With the help of the transformed covariate, Eq. (2.3) is again a simple linear regression, and we can still use the method of least squares to estimate the parameters β_0 and β_1 of the function

$$f(\text{area}) = \beta_0 + \beta_1 \cdot \frac{1}{\text{area}}.$$

We obtain

$$\hat{f}(\text{area}) = 5.44 + 138.32 \cdot \frac{1}{\text{area}}.$$

The corresponding curve in Fig. 2.2 (right) shows a better fit to the data. It reveals that on average the net rent per square meter declines nonlinearly as living area increases. A given living area, e.g., 30 m², corresponds to an estimated average rent per square meter of

$$\widehat{\text{rentsqm}} = 5.44 + 138.32 \cdot \frac{1}{\text{area}}.$$

If the living area increases by 1 m², the average rent decreases and is now given by

$$\widehat{\text{rentsqm}} = 5.44 + 138.32 \frac{1}{\text{area} + 1}.$$

Figure 2.2 (right) shows that the decline is nonlinear. It can be computed by inserting the specific values (e.g., 30 and 31 m²):

$$\widehat{\text{rentsqm}}(30) - \widehat{\text{rentsqm}}(31) = 138.32/30 - 138.32/31 \approx 0.15 \text{ Euro}.$$

An apartment of 60 m² shows a decline of the average rent per square meter by

$$\widehat{\text{rentsqm}}(60) - \widehat{\text{rentsqm}}(61) \approx 0.038 \text{ Euro}.$$

△

In general, the application of a linear regression model requires a relationship between the response and the covariate that is *linear in the coefficients* β_0 and β_1 . The regressor x - and also the response y -can be transformed to achieve linearity in the parameters, as has been illustrated in the above example. However, the question remains how to find an appropriate transformation for the covariate. Nonparametric regression models offer flexible and automatic approaches; see Sect. 2.5 for a first impression and Chap. 8 for full details.

2.2.2 Multiple Linear Regression

Example 2.3 Munich Rent Index—Rent in Average and Good Locations

We now add apartments in good location to the analysis. Figure 2.3 shows the data for rents in average and good locations. In addition to the estimated regression line for apartments in average location, there is another estimated regression line for apartments in a good location. Alternatively, both strata can be analyzed within a single model that shows parallel regression lines. This can be achieved through the model

$$\text{rent}_i = \beta_0 + \beta_1 \text{area}_i + \beta_2 \text{glocation}_i + \varepsilon_i. \quad (2.4)$$

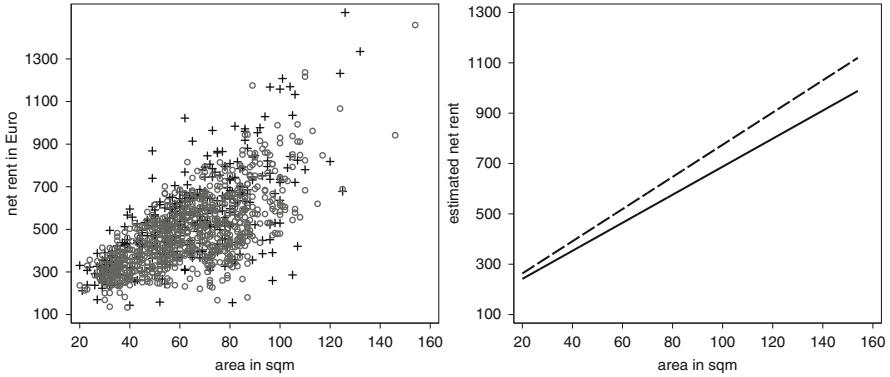


Fig. 2.3 Munich rent index: The *left panel* shows a scatter plot between net rent and area for apartments in average (*circles*) and good location (*plus signs*). The *right panel* displays separate regression lines for apartments in average (*solid line*) and good location (*dashed line*)

The variable *glocation* is a binary *indicator variable*

$$glocation_i = \begin{cases} 1 & \text{if the } i\text{th apartment is in good location,} \\ 0 & \text{if the } i\text{th apartment is in average location.} \end{cases}$$

The least squares method produces the estimated regression equation

$$\widehat{rent} = 112.69 + 5.85 \cdot area + 57.26 \cdot glocation.$$

Because of the 1/0 coding of *glocation*, we obtain the equivalent formulation

$$\widehat{rent} = \begin{cases} 112.69 + 5.85 \cdot area & \text{for average location,} \\ 169.95 + 5.85 \cdot area & \text{for good location.} \end{cases}$$

Figure 2.4 shows both parallel lines. The coefficients can be interpreted as follows:

- For apartments in a good and average location, the increase of living area by 1 m² leads to an average increase of rent of about 5.85 Euro.
- The average rent for an apartment in a good location is about 57.26 Euro higher than for an apartment of the same living area in an average location.

△

Model (2.4) is a special case of a multiple linear regression model for k regressors or covariates x_1, \dots, x_k :

$$y_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_k x_{ik} + \varepsilon_i,$$

where x_{ij} is the value of the j th covariate, $j = 1, \dots, k$, for the i th observation, $i = 1, \dots, n$. The covariates can be continuous, binary, or multi-categorical (after an appropriate coding, see below). Similar to the simple linear regression, x -variables can also be attained via transformation of original covariates. The same assumptions are made for the error variables ε_i in a multiple linear regression model as those in a simple linear regression model. In the case of Gaussian errors, the response variable

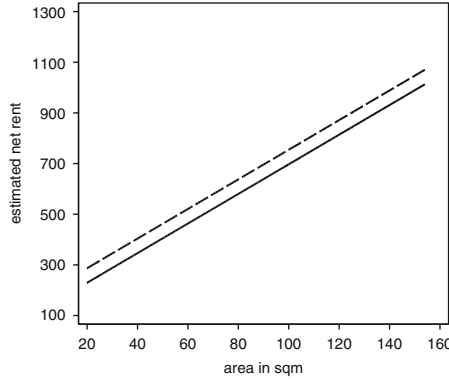


Fig. 2.4 Munich rent index: estimated regression lines for apartments in average (*solid line*) and good location (*dashed line*) according to model (2.4)

is (conditionally) independent and normally distributed given the covariates

$$y_i \sim N(\mu_i, \sigma^2),$$

with

$$\mu_i = E(y_i) = \beta_0 + \beta_1 x_{i1} + \dots + \beta_k x_{ik}.$$

The model is also called the *classical linear regression model*. A summary is given in Box 2.2. For notational convenience we omit here (and elsewhere) the dependence of expressions on the covariates, i.e., $E(y_i)$ is to be understood as an abbreviation for $E(y_i | x_{i1}, \dots, x_{ik})$.

The following examples illustrate the flexible usage of a multiple linear regression model through appropriate transformation and coding of covariates.

Example 2.4 Munich Rent Index—Nonlinear Influence of Living Area

As in Example 2.2, we transform the living area to $x = \frac{1}{area}$ and formulate the linear model

$$rentsqm_i = \beta_0 + \beta_1 \cdot \frac{1}{area_i} + \beta_2 \cdot glocation_i + \varepsilon_i. \quad (2.5)$$

The estimated model for the average rent per square meter is

$$\widehat{rentsqm} = 5.51 + 134.72 \cdot \frac{1}{area} + 0.9 \cdot glocation.$$

Figure 2.5 shows both graphs for the average rent per square meter:

$$\widehat{rentsqm} = \begin{cases} 5.51 + 134.72 \cdot \frac{1}{area} & \text{for average location,} \\ 6.41 + 134.72 \cdot \frac{1}{area} & \text{for good location.} \end{cases}$$

The nonlinear effect of the living area can be interpreted as in Example 2.2.

△

2.2 Classical Linear Regression Model

Data

$(y_i, x_{i1}, \dots, x_{ik}), i = 1, \dots, n$, for a continuous variable y and continuous or appropriately coded categorical regressors x_1, \dots, x_k .

Model

$$y_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_k x_{ik} + \varepsilon_i, \quad i = 1, \dots, n.$$

The errors $\varepsilon_1, \dots, \varepsilon_n$ are independent and identically distributed (i.i.d.) with

$$E(\varepsilon_i) = 0, \quad \text{Var}(\varepsilon_i) = \sigma^2.$$

The estimated linear function

$$\hat{f}(x_1, \dots, x_k) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \dots + \hat{\beta}_k x_k$$

can be used as an estimator $\hat{E}(y|x_1, \dots, x_k)$ for the conditional expected value of y given the covariates x_1, \dots, x_k . As such it can be used to predict y , denoted as \hat{y} .

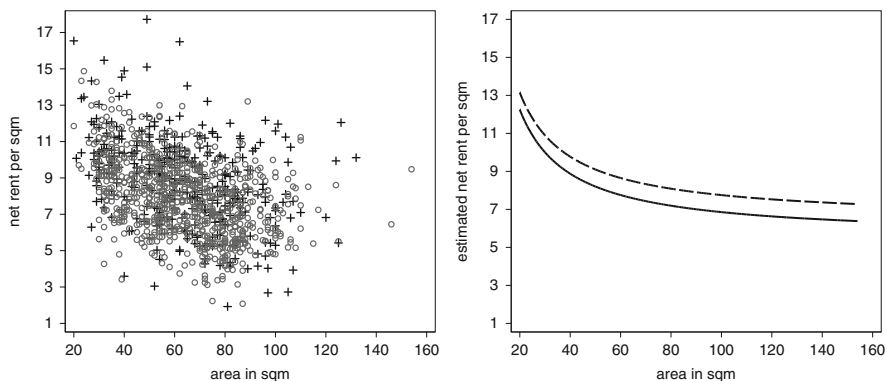


Fig. 2.5 Munich rent index: The *left panel* shows a scatter plot between net rent per square meter and area for apartments in average (*circles*) and good location (*plus signs*). The *right panel* shows estimated regression curves for apartments in normal (*solid line*) and good location (*dashed line*)

Examples 2.3 and 2.4 assume an additive effect of the location. Both models show that an apartment in a good location leads to an increase in rent (or rent per square meter) when compared to an apartment in average location with equal living area. In Example 2.3, the increase in rent is 57.26 Euro and in the previous example 0.9 Euro per square meter. The assumption of a solely additive effect in model

(2.4) implies the parallel lines in Fig. 2.4. However, comparing Figs. 2.3 and 2.4, the validity of this assumption is questionable. Including an interaction between the two covariates living area and location relaxes the assumption of parallel regression lines.

Example 2.5 Munich Rent Index—Interaction Between Living Area and Location

In order to include an interaction between living area and location in model (2.4), it is necessary to define an interaction variable by multiplying the covariates *area* and *glocation*

$$inter_i = area_i \cdot glocation_i.$$

It follows

$$inter_i = \begin{cases} 0 & \text{for average location,} \\ area_i & \text{for good location.} \end{cases}$$

We now extend the model (2.4) by adding the interaction effect $inter = area \cdot glocation$ to the two main effects and obtain

$$rent_i = \beta_0 + \beta_1 area_i + \beta_2 glocation_i + \beta_3 inter_i + \varepsilon_i. \quad (2.6)$$

Because of the definition of *glocation* and *inter*, an equivalent formulation of the model is given by

$$rent_i = \begin{cases} \beta_0 + \beta_1 area_i + \varepsilon_i & \text{for average location,} \\ (\beta_0 + \beta_2) + (\beta_1 + \beta_3) area_i + \varepsilon_i & \text{for good location.} \end{cases}$$

There is no interaction effect if $\beta_3 = 0$, and we retain the assumption of parallel lines with common slope β_1 as in model (2.4). If $\beta_3 \neq 0$, the effect of the living area, i.e., the slope of the line for apartments in a good location, changes by an amount of β_3 when compared to apartments in average location. In contrast to Fig. 2.3 (right), the least squares estimates for the regression coefficients are not obtained separately for the two locations, but rather simultaneously for model (2.6). We obtain

$$\hat{\beta}_0 = 130.23, \quad \hat{\beta}_1 = 5.57, \quad \hat{\beta}_2 = 5.20, \quad \hat{\beta}_3 = 0.82.$$

Figure 2.6 shows the estimated regression lines for apartments in average and good location. Whether or not an inclusion of an interaction effect is necessary can be statistically tested using the hypothesis

$$H_0 : \beta_3 = 0 \quad \text{against} \quad H_1 : \beta_3 \neq 0;$$

see Sect. 3.3.

△

As described in Example 1.1 (p. 5), in the entire data set, the location of apartments is given in three categories:

- 1 = average location,
- 2 = good location,
- 3 = top location.

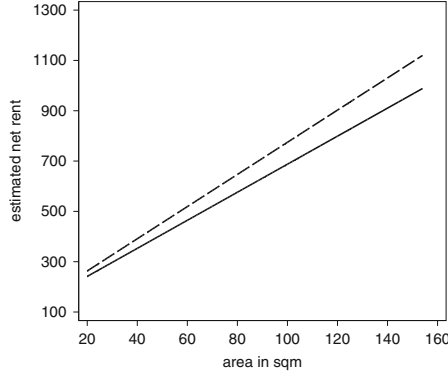


Fig. 2.6 Munich rent index: estimated regression lines for average (*solid line*) and good location (*dashed line*) based on the interaction model (2.6)

Since *location* is categorical and not continuous, it is not possible to include the effect of the location in the form of $\beta \cdot \text{location}$ in a linear regression model with the integer values 1, 2, or 3 for the location. The arbitrary coding of *location* would have considerable impact on the estimation results. The chosen coding automatically implies that the effect of apartments in a good location would be twice as high as in average location and the effect of apartments in top location would be three times as high. These relations change automatically with a different coding, especially if the distance between the arbitrarily coded covariate values is altered. For example, with a coding of 1, 4, and 9 for average, good, and top location, the effect would be four times or nine times as high for apartments in a good or top location as for apartments in average location, and the incremental impact varies when comparing average and good or good and top location. Further, not all categorical covariates have ordinal structure. Similar to the previous coding of *location* via *one* binary indicator variable expressed in Example 2.3, a coding using *two* binary variables is now necessary. In order to do so, one of the three location categories must be defined as the *reference category*. In the case of average location as the reference category, the two 1/0-indicator variables for good location and top location are defined as

$$glocation_i = \begin{cases} 1 & \text{if apartment } i \text{ is in good location,} \\ 0 & \text{otherwise,} \end{cases}$$

$$tlocation_i = \begin{cases} 1 & \text{if apartment } i \text{ is in top location,} \\ 0 & \text{otherwise.} \end{cases}$$

An apartment in the reference category (average location) is, thus, defined as $glocation = tlocation = 0$. The effect of each of these two binary variables is always directly interpreted in relation to the reference category in the regression model, as demonstrated in the next example. This type of 1/0 coding of a *multi-categorical variable* is also called dummy or indicator coding. In general, dummy coding is defined as follows for a variable x with c categories, $x \in \{1, \dots, c\}$: A

Table 2.1 Munich rent index: estimated coefficients in the multiple regression model

Variable	Estimated coefficient
$1/area$	137.504
$yearc$	-3.801
$yearc^2$	0.001
$glocation$	0.679
$tlocation$	1.519
$bath$	0.503
$kitchen$	0.866
$cheating$	1.870

reference category must be defined, e.g., c . The variable x can be then coded with $c - 1$ dummy variables x_1, \dots, x_{c-1} :

$$x_j = \begin{cases} 1 & \text{if } x = j, \\ 0 & \text{otherwise,} \end{cases} \quad j = 1, \dots, c - 1.$$

For the reference category c we obtain

$$x_1 = \dots = x_{c-1} = 0.$$

Section 3.1.3 describes the coding of categorical covariates in more detail.

Example 2.6 Munich Rent Index—Multiple Regression Model

For illustration, we now use the entire data set, including all explanatory variables mentioned in Example 1.1, in a multiple regression model for the response variable rent per square meter ($rentsqm$). The nonlinear effect of the living area is modeled via the transformed variable $1/area$ and the effect of location via dummy coding as described above. Since the effect of the year of construction may also be nonlinear, an additional quadratic polynomial is specified. We obtain the following model without interaction:

$$rentsqm_i = \beta_0 + \beta_1 \cdot (1/area_i) + \beta_2 yearc_i + \beta_3 yearc_i^2 + \beta_4 glocation_i + \beta_5 tlocation_i \\ + \beta_6 bath_i + \beta_7 kitchen_i + \beta_8 cheating_i + \varepsilon_i.$$

The binary regressors $bath$, $kitchen$, and $cheating$ (central heating system) are dummy coded, as shown in Table 1.2 (p. 6). Table 2.1 contains the estimated coefficients $\hat{\beta}_1$ to $\hat{\beta}_8$ for the regressors. Figure 2.7 shows the estimated nonlinear effects of living area and year of construction. The average effect plots (solid lines) are obtained by inserting different values for living area into the predicted rent per square meter

$$\widehat{rentsqm} = 3684.991 + 137.5044 \cdot (1/area) - 3.8007 yearc + 0.0098 yearc^2 \\ + 0.6795 glocation_i + 1.5187 tlocation_i + 0.5027 bath_i + 0.8664 kitchen \\ + 1.8704 cheating,$$

while the other covariates are held constant at their mean values (apart from year of construction). As expected, the effect on the net rent per square meter decreases nonlinearly with an increase of living area. For a detailed comparison of two apartments, e.g., with a

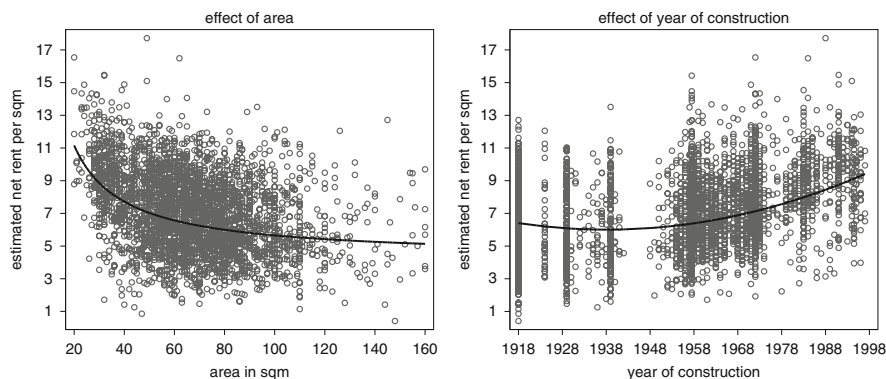


Fig. 2.7 Munich rent index: effects of area (*left*) and year of construction (*right*)

living area of 60 and 100 m², but with otherwise identical values for the year of construction, location, bath room, and central heating system indicators, we obtain a difference of $\hat{\beta}_1(1/60) - \hat{\beta}_1(1/100) = 137.504(1/60 - 1/100) = 0.92$ Euro for the average rent per square meter. The effect of year of construction is almost constant until 1945 and increases linearly thereafter. The effects of the indicator variables shown in Table 2.1 are interpreted as the difference in net rent per square meter compared to the reference category. The average rent per square meter increases, for example, by 0.68 Euro, if the apartment is in a good location (relative to one in average location).

△

2.3 Regression with Binary Response Variables: The Logit Model

The linear regression model is well suited for continuous response variables, which show—possibly after an appropriate transformation—an approximate normal distribution (conditional on the covariates). However, many applications have binary or more general categorical response variables.

Example 2.7 Patent Opposition

During the validation of a patent application, it is possible that objections are raised, e.g. see Example 1.3 (p. 8). The response variable patent opposition (*opp*) is binary and coded by

$$opp_i = \begin{cases} 1 & \text{opposition against patent } i, \\ 0 & \text{otherwise.} \end{cases}$$

The decision for an opposition against a patent may depend on various covariates. Some of these variables are continuous, for example, the year of the application (variable *year*), the number of citations (*ncit*), and the number of designated states (*ncountry*). Other covariates are binary, as given in Table 1.4 (p. 8).

△

The expected value of a binary variable y is given by

$$E(y) = P(y = 0) \cdot 0 + P(y = 1) \cdot 1 = P(y = 1).$$

The aim of a regression analysis with binary responses $y \in \{0, 1\}$ is to model the expected value $E(y)$ or in other words the probability

$$P(y = 1) = P(y = 1 \mid x_1, \dots, x_k) = \pi$$

in the presence of covariates. The classical linear regression model

$$y_i = P(y_i = 1) + \varepsilon_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_k x_{ik} + \varepsilon_i,$$

with $\varepsilon_i \sim N(0, \sigma^2)$ is not applicable for several reasons:

- In contrast to the left-hand side, the right-hand side is not binary.
- Even if the assumption of normality is relaxed for ε_i , the error variance $\text{Var}(\varepsilon_i)$ cannot be homoscedastic, i.e., $\text{Var}(\varepsilon_i) = \sigma^2$. Since y_i would have a Bernoulli distribution with $\pi_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_k x_{ik}$, it follows that

$$\text{Var}(y_i) = \pi_i(1 - \pi_i)$$

depends on the values of the covariates and the parameters β_0, \dots, β_k , and thus cannot have the same value σ^2 for all observations i .

- The linear model allows values $\pi_i < 0$ or $\pi_i > 1$ for $\pi_i = P(y_i = 1)$ which are impossible for probabilities.

These problems can be avoided by assuming the model

$$\pi_i = P(y_i = 1) = F(\beta_0 + \beta_1 x_{i1} + \dots + \beta_k x_{ik}),$$

where the domain of the function F is restricted to the interval $[0, 1]$. For reasons of interpretability it is sensible if we restrict ourselves to monotonically increasing functions F . Hence cumulative distribution functions (cdfs) are a natural choice for F . Choosing the logistic distribution function

$$F(\eta) = \frac{\exp(\eta)}{1 + \exp(\eta)}$$

yields the logit model

$$\pi_i = P(y_i = 1) = \frac{\exp(\eta_i)}{1 + \exp(\eta_i)}, \quad (2.7)$$

with the *linear predictor*

$$\eta_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_k x_{ik}.$$

Analogous to the linear regression model, the binary response variables y_i are assumed to be (conditionally) independent given the covariates $x_i = (x_{i1}, \dots, x_{ik})'$. Even though the predictor is linear, the interpretation changes compared to the linear model: If the value of the predictor η increases to $\eta + 1$, the probability for $y = 1$ increases in a *nonlinear* way from $F(\eta)$ to $F(\eta + 1)$. An alternative interpretation is obtained by solving the model equation (2.7) for η using the inverse function $\eta = \log\{\pi/(1 - \pi)\}$ of the logistic cdf $\pi = \exp(\eta)/\{1 + \exp(\eta)\}$. We obtain

$$\log\left(\frac{\pi_i}{1 - \pi_i}\right) = \log\left(\frac{P(y_i = 1)}{1 - P(y_i = 1)}\right) = \beta_0 + \beta_1 x_{i1} + \dots + \beta_k x_{ik} \quad (2.8)$$

or alternatively (because of $\exp(a + b) = \exp(a) \cdot \exp(b)$)

$$\frac{\pi_i}{1 - \pi_i} = \frac{P(y_i = 1)}{P(y_i = 0)} = \exp(\beta_0) \exp(\beta_1 x_{i1}) \cdot \dots \cdot \exp(\beta_k x_{ik}). \quad (2.9)$$

The left-hand side of Eq. (2.9), i.e., the ratio of the probabilities for $y = 1$ and $y = 0$, is referred to as *odds*. The left-hand side of Eq. (2.8), thus, corresponds to logarithmic odds (*log-odds*) for the outcome of $y = 1$ relative to $y = 0$. Here, we obtain a *multiplicative model* for the odds: A unit increase of the value x_{i1} of the covariate x_1 leads to a multiplication of the ratio (2.9) by the factor $\exp(\beta_1)$. Specifically,

$$\begin{aligned} \frac{P(y_i = 1 | x_{i1} + 1, \dots)}{P(y_i = 0 | x_{i1} + 1, \dots)} &= \exp(\beta_0) \exp(\beta_1(x_{i1} + 1)) \cdot \dots \cdot \exp(\beta_k x_{ik}) \\ &= \frac{P(y_i = 1 | x_{i1}, \dots)}{P(y_i = 0 | x_{i1}, \dots)} \exp(\beta_1). \end{aligned} \quad (2.10)$$

In the special case of a binary covariate x_1 the result is

$$\frac{P(y_i = 1 | x_{i1} = 1, \dots)}{P(y_i = 0 | x_{i1} = 1, \dots)} = \frac{P(y_i = 1 | x_{i1} = 0, \dots)}{P(y_i = 0 | x_{i1} = 0, \dots)} \exp(\beta_1). \quad (2.11)$$

This implies an increase of the odds $P(y_i = 1)/P(y_i = 0)$ for $\beta_1 > 0$, a decrease for $\beta_1 < 0$, and no change for $\beta_1 = 0$. For the log-odds (2.8) the usual interpretations of the parameters as in the classical linear regression model apply: if x_1 , say, increases by 1 unit, the log-odds change by β_1 . Since the assumptions for the linear regression model are not met, the parameters will not be estimated via the least squares method, but rather using the method of maximum likelihood (ML); see Sect. 5.1. A general introduction to likelihood-based inference is given in Appendix B.4.

Example 2.8 Patent Opposition

Prior to analyzing the probability of patent opposition, we take a look at Fig. 2.8, which presents histograms and kernel density estimators for the continuous covariates number of patent claims (*nclaims*) and number of citations (*ncit*). The distributions of both variables show an extreme skewness to the right. The majority of the observations of *nclaims* are between 0 and 60, with only very few observations between 61 and the maximum value of

2.3 The Logit Model for Binary Response Variables

Data

$(y_i, x_{i1}, \dots, x_{ik})$, $i = 1, \dots, n$, for a binary response variable $y \in \{0, 1\}$ and for continuous or appropriately coded covariates x_1, \dots, x_k .

Model

For the binary response variables $y_i \in \{0, 1\}$ the probabilities $\pi_i = P(y_i = 1)$ are modeled by

$$\pi_i = \frac{\exp(\eta_i)}{1 + \exp(\eta_i)}$$

with the linear predictor

$$\eta_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_k x_{ik}.$$

An equivalent formulation is given by assuming the multiplicative model

$$\frac{P(y_i = 1)}{P(y_i = 0)} = \frac{\pi_i}{1 - \pi_i} = \exp(\beta_0) \cdot \exp(\beta_1 x_{i1}) \cdot \dots \cdot \exp(\beta_k x_{ik})$$

for the odds $\pi_i / (1 - \pi_i)$.

355. The variable *ncit* varies mainly between 0 and 15 with only a handful of observations between 15 and the maximum value of 40. Hence, it is impossible to make any reliable statements regarding the probability of patent opposition for observations with *nclaims* > 60 or *ncit* > 15. Consequently, these extreme cases are excluded from all analyses to follow. This example demonstrates the importance of the descriptive analysis of data prior to the application of more complex statistical tools.

We next divide the data into two groups: *biopharm* = 0 and *biopharm* = 1. For the subset *biopharm* = 0, i.e., the patents derived from the semiconductor/computer industry, a logit model

$$P(\text{opp}_i = 1) = \frac{\exp(\eta_i)}{1 + \exp(\eta_i)}$$

is estimated with the main effects linear predictor

$$\begin{aligned} \eta_i = & \beta_0 + \beta_1 \text{year}_i + \beta_2 \text{ncit}_i + \beta_3 \text{nclaims}_i + \beta_4 \text{ustwin}_i + \beta_5 \text{patus}_i \\ & + \beta_6 \text{patgsgr}_i + \beta_7 \text{ncountry}_i. \end{aligned}$$

Table 2.2 contains the estimated coefficients $\hat{\beta}_j$, $j = 0, \dots, 7$, together with the corresponding odds ratios $\exp(\hat{\beta}_j)$. In multiplicative form (2.9) we obtain

$$\frac{P(\text{opposition})}{P(\text{no opposition})} = \exp(201.74) \cdot \exp(-0.102 \cdot \text{year}_i) \cdot \dots \cdot \exp(0.097 \cdot \text{ncountry}_i).$$

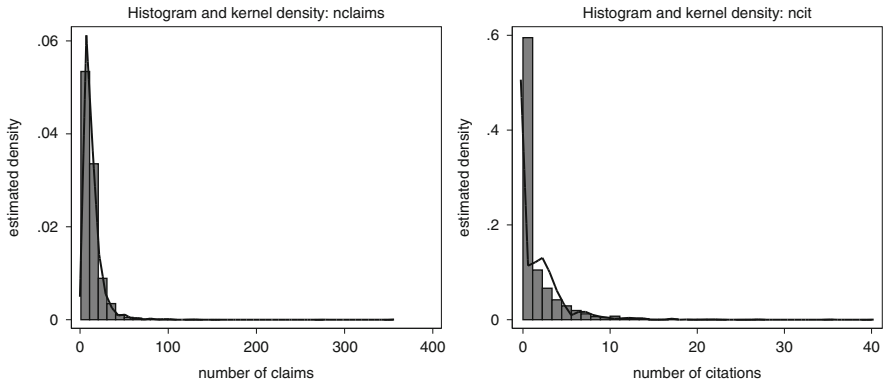


Fig. 2.8 Patent opposition: histogram and kernel density estimator for the continuous covariates *nclaims* (left) and *ncit* (right)

Table 2.2 Patent opposition: estimated coefficients and odds ratios for the logit model

Variable	Estimated coefficient	Estimated odds ratio
<i>intercept</i>	$\hat{\beta}_0 = 201.74$	
<i>year</i>	$\hat{\beta}_1 = -0.102$	$\exp(\hat{\beta}_1) = 0.902$
<i>ncit</i>	$\hat{\beta}_2 = 0.113$	$\exp(\hat{\beta}_2) = 1.120$
<i>nclaims</i>	$\hat{\beta}_3 = 0.026$	$\exp(\hat{\beta}_3) = 1.026$
<i>ustwin</i>	$\hat{\beta}_4 = -0.402$	$\exp(\hat{\beta}_4) = 0.668$
<i>patus</i>	$\hat{\beta}_5 = -0.526$	$\exp(\hat{\beta}_5) = 0.591$
<i>patgsgr</i>	$\hat{\beta}_6 = 0.196$	$\exp(\hat{\beta}_6) = 1.217$
<i>ncountry</i>	$\hat{\beta}_7 = 0.097$	$\exp(\hat{\beta}_7) = 1.102$

We observe, for instance, an increase in the odds of opposition against a patent from Germany, Switzerland, or Great Britain (*patgsgr* = 1) by the factor $\exp(0.196) = 1.217$ relative to a patent from the United States with the same values of the other covariates. A prediction of the odds $P(\text{opposition}) / P(\text{no opposition})$ for a new patent is obtained by inserting the observed covariate values into the estimated model. Similar to linear regression models, we have to decide whether the effect of a continuous covariate is linear or nonlinear. As an example, we model the effect of the number of countries (*ncountry*) using a cubic polynomial

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

The parameter estimates are given by

$$\hat{\beta}_7 = 0.3938 \quad \hat{\beta}_8 = -0.0378 \quad \hat{\beta}_9 = 0.0014.$$

Figure 2.9 shows the estimated polynomial and, for comparison, the linear effect (left panel). As before, the values of the remaining covariates are held fixed at their respective average values. Both the estimated regression coefficients and the visualized functions suggest that a linear effect of *ncountry* is sufficient in this case. This hypothesis can be formally tested using the statistical tests described in Sect. 5.1. The right panel of Fig. 2.9 shows

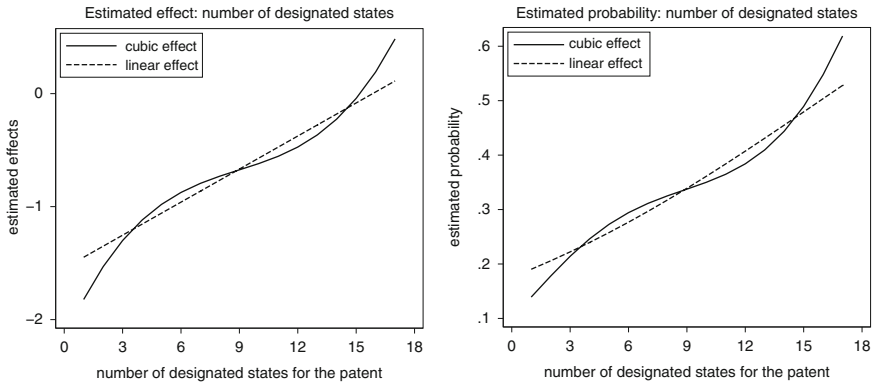


Fig. 2.9 Patent opposition: estimated linear and cubic effect of covariate *ncountry* (left panel) as well as estimated probabilities (right panel). For the probability plot, the values of the remaining covariates are held fixed at their respective mean values

the *estimated probabilities* π corresponding to the estimated *ncountry* effects. This is an alternative to the effect plots as the probability plots provide an intuition about the variability of the probability of patent opposition as the number of designated states increases and the remaining covariates are kept fixed at their mean values. Specifically, the graph is obtained by plotting $\hat{\pi}(\eta(\text{ncountry}))$ against *ncountry*. Thereby

$$\eta(\text{ncountry}) = \hat{\beta}_0 + \hat{\beta}_1 \overline{\text{year}} + \dots + \hat{\beta}_6 \overline{\text{patgsgr}} + \hat{\beta}_7 \text{ncountry} + \hat{\beta}_8 \text{ncountry}^2 + \hat{\beta}_9 \text{ncountry}^3,$$

with $\overline{\text{year}}, \dots, \overline{\text{patgsgr}}$ being the mean values of the remaining covariates. In our case the probability of patent opposition varies approximately between 0.15 and 0.6 as *ncountry* increases.

△

In addition to the logit model, other regression models for binary responses exist. Different models result when the logistic distribution function is replaced by an alternative distribution function. For instance, assuming $F = \Phi$, where Φ is the cdf of the standard normal distribution, yields the probit model (see Sect. 5.1 for more details).

In addition to binary response variables, other types of discrete response variables are possible in applications. For these applications, linear regression models are not appropriate. An example is a response y that represents counts $\{0, 1, 2, \dots\}$, e.g., the amount of damage events reported to an insurance company (see Example 2.12), or a multi-categorical response variable, e.g., with the categories poor, average, and good. Chapters 5 and 6 describe regression models for such discrete response variables in full detail.

2.4 Mixed Models

The regression models presented so far are particularly useful for the analysis of regression data resulting from cross-sectional studies, where the regression coefficients β_0, \dots, β_k are unknown population (“fixed”) parameters. Regression

Table 2.3 Hormone therapy with rats: number of observations per time point and dose group

Age (in days)	Control	Low	High	Total
50	15	18	17	50
60	13	17	16	46
70	13	15	15	43
80	10	15	13	38
90	7	12	10	29
100	4	10	10	24
110	4	8	10	22

problems also occur when analyzing longitudinal data, where a number of subjects or objects are repeatedly observed over time. In such a case, regression models for longitudinal data allow to model and estimate both the fixed population parameters and subject- or object-specific effects. The latter are called “random effects,” since they often belong to individuals who have been selected randomly from the population. Closely related to the random effects models with temporal structure are models for clustered data. Here, the response and covariates are collected repeatedly on several subjects, selected from primary units (clusters). An example for clusters are selected schools, in which certain tests for a subsample of students are conducted.

Mixed models include both the usual fixed population effects β_0, \dots, β_k and subject- or cluster-specific random effects in the linear predictor. Mixed modeling allows estimation and analysis on a subject-specific level, which is illustrated in the following example in the case of longitudinal data.

Example 2.9 Hormone Therapy with Rats

Researchers at the Katholieke Universiteit Leuven (KUL, Belgium) performed an experiment to examine the effect of testosterone on the growth of rats. A detailed description of the data and the scientific questions of the study can be found in Verbeke and Molenberghs (2000). A total of 50 rats were randomly assigned to either a control group or to one of two therapy groups. The therapy consisted of either a low or high dose of Decapeptyl, an agent to inhibit the production of testosterone in rats. The therapy started when the rats were 45 days old. Starting with the 50th day, the growth of the rat’s head was measured every tenth day via an X-ray examination. The distance (measured in pixels) between two well-defined points of the head served as a measure for the head height and was used as the response variable. The number n_i of repeated measures y_{ij} , $j = 1, \dots, n_i$, $i = 1, \dots, 50$, of the response was different for each rat. Only 22 rats in total had the complete seven measurements until the age of 110 days. Four rats were actually only measured once when they were 50 days old. Table 2.3 summarizes the resulting design of the study. Figure 2.10 shows the individual time series $\{y_{ij}, j = 1, \dots, n_i\}$ of rats $i = 1, \dots, 50$ separated for the three treatment groups.

To formulate regression models, we define the transformed age

$$t = \log(1 + (\text{age} - 45)/10)$$

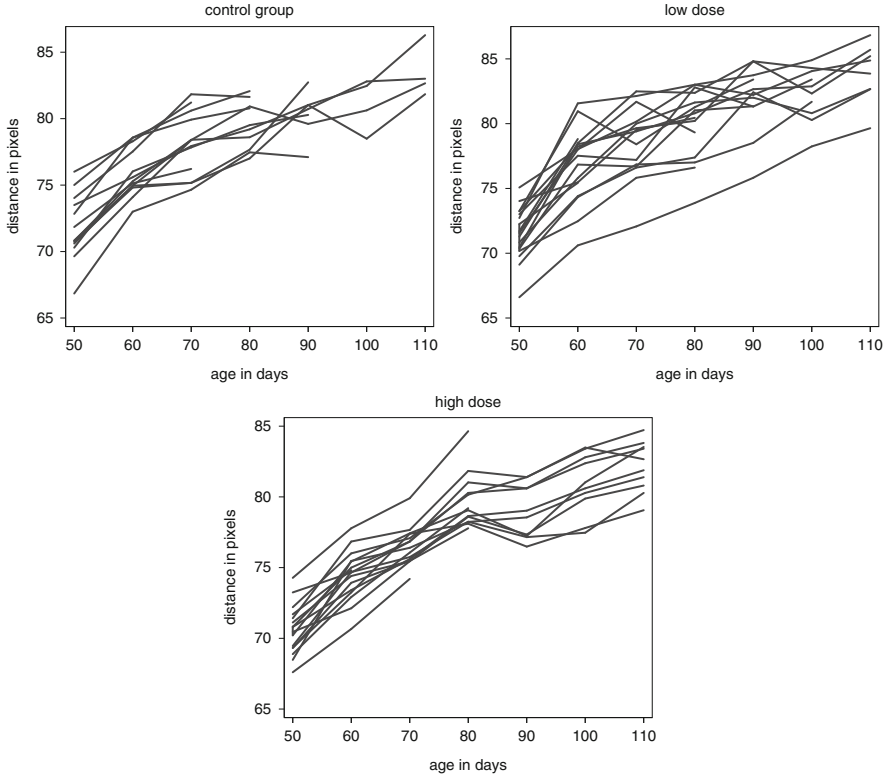


Fig. 2.10 Hormone therapy with rats: time series stratified for dose groups

as a covariate, analogous to Verbeke and Molenberghs (2000). The value $t = 0$ corresponds to the initiation of the treatment (age = 45 days). For the three therapy groups we define the indicator variables L , H , and C by

$$L_i = \begin{cases} 1 & \text{rat } i \text{ in low-dose group,} \\ 0 & \text{otherwise,} \end{cases}$$

$$H_i = \begin{cases} 1 & \text{rat } i \text{ in high-dose group,} \\ 0 & \text{otherwise,} \end{cases}$$

$$C_i = \begin{cases} 1 & \text{rat } i \text{ in control group,} \\ 0 & \text{otherwise.} \end{cases}$$

Using the transformed age t as time scale and $t = 0$ as the initiation of the treatment, we can formulate simple linear regression models according to the three groups:

$$y_{ij} = \begin{cases} \beta_0 + \beta_1 t_{ij} + \varepsilon_{ij} & i \text{ in low-dose group,} \\ \beta_0 + \beta_2 t_{ij} + \varepsilon_{ij} & i \text{ in high-dose group,} \\ \beta_0 + \beta_3 t_{ij} + \varepsilon_{ij} & i \text{ in control.} \end{cases}$$

For $t = 0$, all three treatment groups have $E(y_{ij}) = \beta_0$, i.e., β_0 is the *population mean* at treatment initiation. The coefficients β_1 , β_2 , and β_3 correspond to the different slopes associated with the variable t , i.e., the effects of the (transformed) age in the three linear models. The three models can be combined into the single model

$$y_{ij} = \beta_0 + \beta_1 L_i \cdot t_{ij} + \beta_2 H_i \cdot t_{ij} + \beta_3 C_i \cdot t_{ij} + \varepsilon_{ij}, \quad (2.12)$$

with 1/0-indicator variables L , H , and C for the three groups. Similar to β_0 , the parameters β_1 , β_2 , and β_3 are *population effects*, which do not capture any individual differences between the rats. However, Fig. 2.10 reveals some obvious differences of the individual curves in the intercept, as well as possible differences in the slope. Moreover, the variability within the individual curves is notably less than the total variation of the data in any of the three group-specific scatter plots. In particular, these findings show that the observations are partly *correlated*, whereas so far we have always assumed independence among observations. While the observations between subjects can still be assumed independent, observations within rats are clearly correlated. The inclusion of subject-specific information will consider the correlation and therefore improve the quality of the estimates. In order to incorporate subject-specific effects, we extend the regression models mentioned above and obtain

$$y_{ij} = \begin{cases} \beta_0 + \gamma_{0i} + (\beta_1 + \gamma_{1i})t_{ij} + \varepsilon_{ij} & i \text{ in low-dose group,} \\ \beta_0 + \gamma_{0i} + (\beta_2 + \gamma_{1i})t_{ij} + \varepsilon_{ij} & i \text{ in high-dose group,} \\ \beta_0 + \gamma_{0i} + (\beta_3 + \gamma_{1i})t_{ij} + \varepsilon_{ij} & i \text{ in control group,} \end{cases}$$

or equivalently

$$y_{ij} = \beta_0 + \gamma_{0i} + \beta_1 L_i \cdot t_{ij} + \beta_2 H_i \cdot t_{ij} + \beta_3 C_i \cdot t_{ij} + \gamma_{1i} \cdot t_{ij} + \varepsilon_{ij}. \quad (2.13)$$

The model contains subject-specific deviations γ_{0i} from the population mean β_0 as well as subject-specific deviations γ_{1i} from the population slopes β_1 , β_2 , and β_3 . In contrast to the “fixed” effects $\beta = (\beta_0, \beta_1, \beta_2, \beta_3)'$, the subject-specific effects $\gamma_i = (\gamma_{0i}, \gamma_{1i})'$ are considered as random effects, because the rats have been randomly selected from a population. We assume that the random effects are independent and identically distributed and follow a normal distribution, i.e.,

$$\gamma_{0i} \sim N(0, \tau_0^2), \quad \gamma_{1i} \sim N(0, \tau_1^2). \quad (2.14)$$

Without loss of generality the expected values can be set to zero, because the population mean values are already included in the fixed effects β . At first sight, a more natural approach to consider subject-specific effects is the inclusion of *subject-specific dummy variables* without a random effects distribution. In principal, such an approach is possible for a small or moderate number of subject-specific effects. However, since usually a large or even huge number of dummy variables are necessary for the subject-specific effects, the resulting estimates would be highly unstable. As we will see in full detail in Chap. 7, the random effects distribution Eq. (2.14), in particular, the common variances τ_0^2 and τ_1^2 across subjects, *stabilizes* estimation.

For the errors ε_{ij} , we make the same assumptions as in the classical linear model, i.e., the errors are independent and identically distributed as

$$\varepsilon_{ij} \sim N(0, \sigma^2). \quad (2.15)$$

Note, however, that correlation within individuals is considered through the subject-specific effects. Since the model (2.13) includes fixed effects as in the classical linear regression

2.4 Linear Mixed Models for Longitudinal and Clustered Data

Data

For each of the $i = 1, \dots, m$ subjects or clusters, n_i repeated observations

$$(y_{ij}, x_{ij1}, \dots, x_{ijk}), \quad j = 1, \dots, n_i,$$

for a continuous response variable y and continuous or appropriately coded covariates x_1, \dots, x_k are given.

Model

For the linear mixed model, we assume

$$y_{ij} = \beta_0 + \beta_1 x_{ij1} + \dots + \beta_k x_{ijk} + \gamma_{0i} + \gamma_{1i} u_{ij1} + \dots + \gamma_{qi} u_{ijq} + \varepsilon_{ij},$$

$i = 1, \dots, m, j = 1, \dots, n_i$. Thereby, the β_0, \dots, β_k are fixed *population effects* and $\boldsymbol{\gamma}_i = (\gamma_{0i}, \gamma_{1i}, \dots, \gamma_{qi})'$ are *subject- or cluster-specific effects*. We assume that the random effects $\boldsymbol{\gamma}_i$ are independent and identically distributed according to a (possibly multivariate) normal distribution.

model (2.12), as well as random effects $\gamma_{0i}, \gamma_{1i}, i = 1, \dots, 50$, it is called a *linear mixed model* or a regression model with both fixed and random effects. \triangle

In the case of longitudinal data, some of the covariates x_{ij1}, \dots, x_{ijk} can be time-varying, as, e.g., the transformed age in the rats example. They can also be time-constant; examples are the indicator variables L_i, H_i , and C_i . For cluster data, this means that in cluster i the covariate value depends on object j or alternatively that the covariate contains only cluster-specific information.

A general notation for linear mixed models for longitudinal and cluster data is given by

$$y_{ij} = \beta_0 + \beta_1 x_{ij1} + \dots + \beta_k x_{ijk} + \gamma_{0i} + \gamma_{1i} u_{ij1} + \dots + \gamma_{qi} u_{ijq} + \varepsilon_{ij}, \quad (2.16)$$

where $i = 1, \dots, m$ is the individual or cluster index and $j = 1, \dots, n_i$ indicates the j th measurement for individual or cluster i . In the case of repeated measurements over time, the observed (not necessarily equally spaced) time points for individual i are denoted by $t_{i1} < \dots < t_{ij} < \dots < t_{in_i}$. The fixed parameters β_0, \dots, β_k in Eq. (2.16) measure population effects, while the random parameters $\boldsymbol{\gamma}_i = (\gamma_{0i}, \gamma_{1i}, \dots, \gamma_{qi})'$ describe subject- or cluster-specific effects. The additional design variables u_{ij1}, \dots, u_{ijq} often consist of some of the covariates x_{ij1}, \dots, x_{ijk} , as the transformed age t_{ij} in Example 2.9.

In most situations, similar assumptions are made for the random errors as in the classical linear regression model, i.e., the ε_{ij} are independently and identically (normally) distributed with $E(\varepsilon_{ij}) = 0$ and $\text{Var}(\varepsilon_{ij}) = \sigma^2$. It is also possible to model correlations between the errors ε_{ij} , $j = 1, \dots, n_i$, of repeated observations within individuals or clusters; see Chap. 7. Such correlated errors are necessary if there is extra correlation not taken into account by the subject-specific effects. For the random effects, it is also often assumed that the γ_{li} , $l = 0, \dots, q$, are independent and identically distributed according to separate normal distribution, as in Example 2.9. Again, more general formulations with correlated random effects are possible. Then the vector of individual random effects $\boldsymbol{\gamma}_i$ is assumed to be i.i.d. according to a multivariate normal distribution with possibly non-diagonal covariance matrix; see Chap. 7 for details.

Analyses with mixed models for longitudinal or cluster data have the following advantages:

- Correlations between observations of the same individual or cluster are taken into account (at least to a certain extent).
- Subject-specific effects can serve as surrogates for unobserved covariate effects, which either have been measured insufficiently or not measured at all. Since the observations differ due to such unobserved covariates, there is an implied unobserved heterogeneity.
- The inclusion of subject-specific information often leads to more precise estimates for the fixed effects, i.e., less variable estimators, when compared to standard regression models. In any case, mixed models ensure that inference regarding the regression coefficients is correct in the sense that we obtain correct standard errors, confidence intervals, and tests.
- Mixed models stabilize the estimators of random effects by assuming a common random effects distribution.
- The estimated individual curves further allow for individual-specific predictions, which are not available in standard regression models.

Statistical inference for fixed and random effects, as well as for the error and random effects variances, is accomplished using likelihood approaches or Bayesian inference as outlined in Chap. 7.

Example 2.10 Hormone Therapy with Rats

First consider model (2.13), which comprises subject-specific deviations γ_{0i} from the overall population intercept β_0 and the subject-specific slope parameters γ_{1i} . We estimate the fixed effects, the variance parameters σ^2 , τ_0^2 , and τ_1^2 , and the random effects using inference techniques described in Chap. 7 and implemented in function `lmer` of the R package `lme4`. Table 2.4 contains the estimated fixed effects and the variance parameters. Since the estimated value $\hat{\tau}_1^2$ for $\text{Var}(\gamma_{1i})$ is very small, we also consider a simpler model that does not contain the subject-specific terms $\gamma_{1i}t_{ij}$. The results can be found again in Table 2.4. As we can see, the estimates are all very similar.

For the simpler model without random slope, Fig. 2.11 shows a kernel density estimator for the estimated values $\hat{\gamma}_{0i}$, $i = 1, \dots, 50$ together with a superimposed normal density and a normal quantile plot in a separate graph. We do not find any serious deviations from the assumed normal distribution.

△

Table 2.4 Hormone therapy with rats: estimation results for the mixed model (2.13) and the simplified model without individual-specific slope parameter

	Parameter	Model (2.13) estimated value	Simplified model estimated value
Intercept	β_0	68.607	68.607
Low-dose	β_1	7.505	7.507
High-dose	β_2	6.874	6.871
Control	β_3	7.313	7.314
$\text{Var}(\gamma_{0i})$	τ_0^2	3.739	3.565
$\text{Var}(\gamma_{1i})$	τ_1^2	<0.001	
$\text{Var}(\varepsilon_{ij})$	σ^2	1.481	1.445

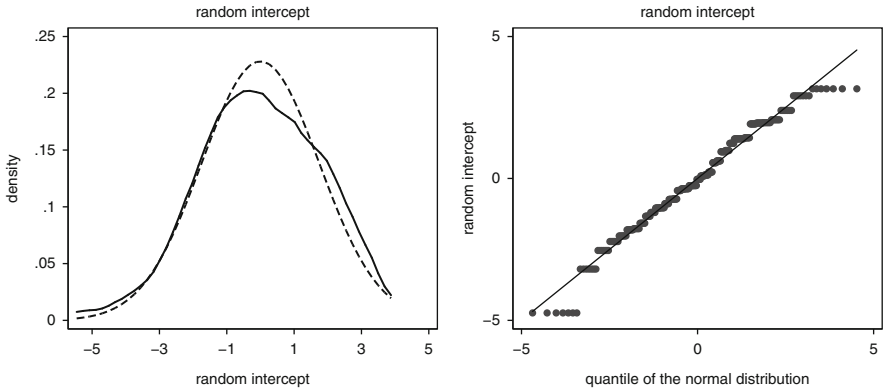


Fig. 2.11 Hormone therapy with rats: The *left plot* shows for the random intercept a kernel density estimator (*solid line*) and the density of an adapted normal distribution (*dashed line*). The *right panel* displays a normal quantile plot

In Chap. 7, we present generalizations of the basic linear mixed models, including mixed models for binary and discrete response variables. This more general group of models can also be used as a basis for inference in nonparametric and semiparametric regression models outlined in the next section and presented in detail in Chaps. 8 and 9.

2.5 Simple Nonparametric Regression

Figure 2.2 (p. 25) shows the scatter plot of the two variables *rentsqm* and *area* for the rent index data discussed in Sect. 2.2. The scatter plot reveals that the living area has a nonlinear effect on the net rent per square meter. Consequently, in Example 2.2, the effect of living area was modeled nonlinearly by

$$f(\text{area}) = \beta_0 + \beta_1/\text{area}. \tag{2.17}$$

Figure 1.8 (p. 17) presents the scatter plot between the Z-score (as a measure of chronic undernutrition) of a child in Zambia and the age of the child (see Example 1.2 on p. 5). Again, we observe that the Z-score depends on the age of the child in a nonlinear way.

In fact, in most of the various applications presented in Chap. 1, nonlinear effects are present. It is often very difficult to model these with ad hoc parametric approaches as for instance in Eq. (2.17). Moreover, in most cases other transformations are reasonable, e.g., $f(area) = \beta_0 + \beta_1 \log(area)$ or $f(area) = \beta_0 + \beta_1 (area)^{\frac{1}{2}}$. In more complex applications with more continuous regressors, searching for suitable transformations becomes very difficult or intractable even for very experienced researchers.

Non- and semiparametric regression models allow for flexible estimation of nonlinear effects. They do not require any restrictive assumptions regarding a certain parametric functional form. In the case of just one continuous covariate x , the *standard model for nonparametric regression* is defined as

$$y_i = f(x_i) + \varepsilon_i. \quad (2.18)$$

For the error variable ε_i , the same assumptions as in the simple linear regression model (2.2) are made.

The function f is assumed to be sufficiently smooth, but no specific parametric form is specified. It is estimated in a data-driven way through nonparametric approaches. Chapter 8 describes several techniques of how to estimate the unknown function f . To give the reader a first impression of nonparametric regression models, Figs. 2.12 and 2.13 demonstrate an easily comprehensible estimation concept using the Zambia malnutrition data. For illustration purposes, we restrict our analysis to the observations of a specific district in Zambia; see Fig. 2.12a. The goal is to find an estimator $\hat{f}(c_age)$ for the (nonlinear) relationship between the Z-score and the age of a child. Figure 2.12a shows that a simple regression line cannot produce a satisfactory fit. However, we find that a linear model is *locally* justified, i.e., if the analysis is restricted to appropriately defined intervals; see Fig. 2.12b, c.

Based on these observations, we obtain the following approach that is known as the *nearest neighbor estimator*:

1. Determine a number of values

$$c_age_1 < c_age_2 < \dots < c_age_m$$

within the support of c_age for which estimates $\hat{f}(c_age_j)$, $j = 1, \dots, m$, of f are to be computed.

2. When estimating f at c_age_j , use a predetermined number of observations in the *neighborhood* of c_age_j . In Fig. 2.12b, c the nearest 70 observations, either to the right or left, of $c_age_j = 11$ (panel b) and $c_age_j = 28$ (panel c) were used.

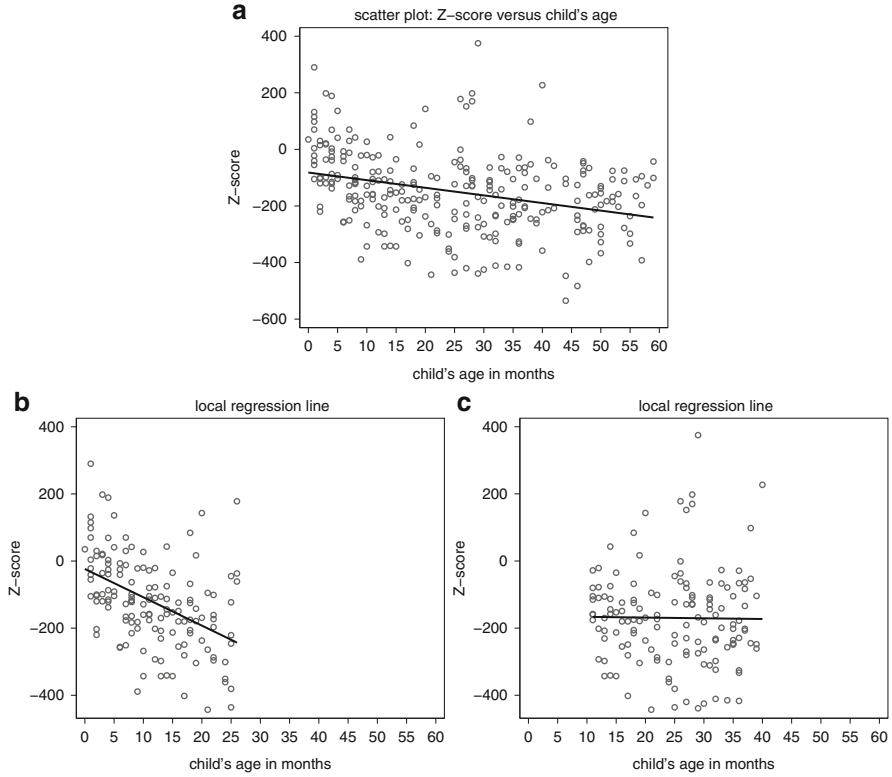


Fig. 2.12 Malnutrition in Zambia: illustration of a global and local regression. (a) shows a global regression line for the relationship between Z-score and child's age. (b) and (c) show local regression lines based on a subset of the data

3. Estimate a local regression line based on the observations taken in step 2. to obtain the estimate $\hat{f}(c_age_j) = \hat{\beta}_0 + \hat{\beta}_1 c_age_j$. Note that for every value of c_age_j , a separate regression line is estimated. This implies that the regression coefficients $\hat{\beta}_0$ and $\hat{\beta}_1$ vary according to c_age .
4. Combine the obtained estimates $\hat{f}(c_age_1), \dots, \hat{f}(c_age_m)$ and visualize the estimated curve.

An illustration of the nearest neighbor estimator is given in Fig. 2.13.

Figure 2.12 also suggests another approach. Instead of estimating a *global* regression line, the domain of the covariate, in our example the child's age, could first be subdivided into several non-overlapping intervals. In a second step, a separate regression line could then be estimated using the data in each interval. This procedure is illustrated in Fig. 2.14a. Here the range of values was divided into the three intervals $[0, 19)$, $[19, 39)$ and $[39, 59]$. In contrast to the global regression line, we obtain a satisfactory fit to the data. There is, however, a flaw: the separate regression lines induce discontinuities at the interval boundaries. An obvious solution is to impose the additional constraint that the function is globally

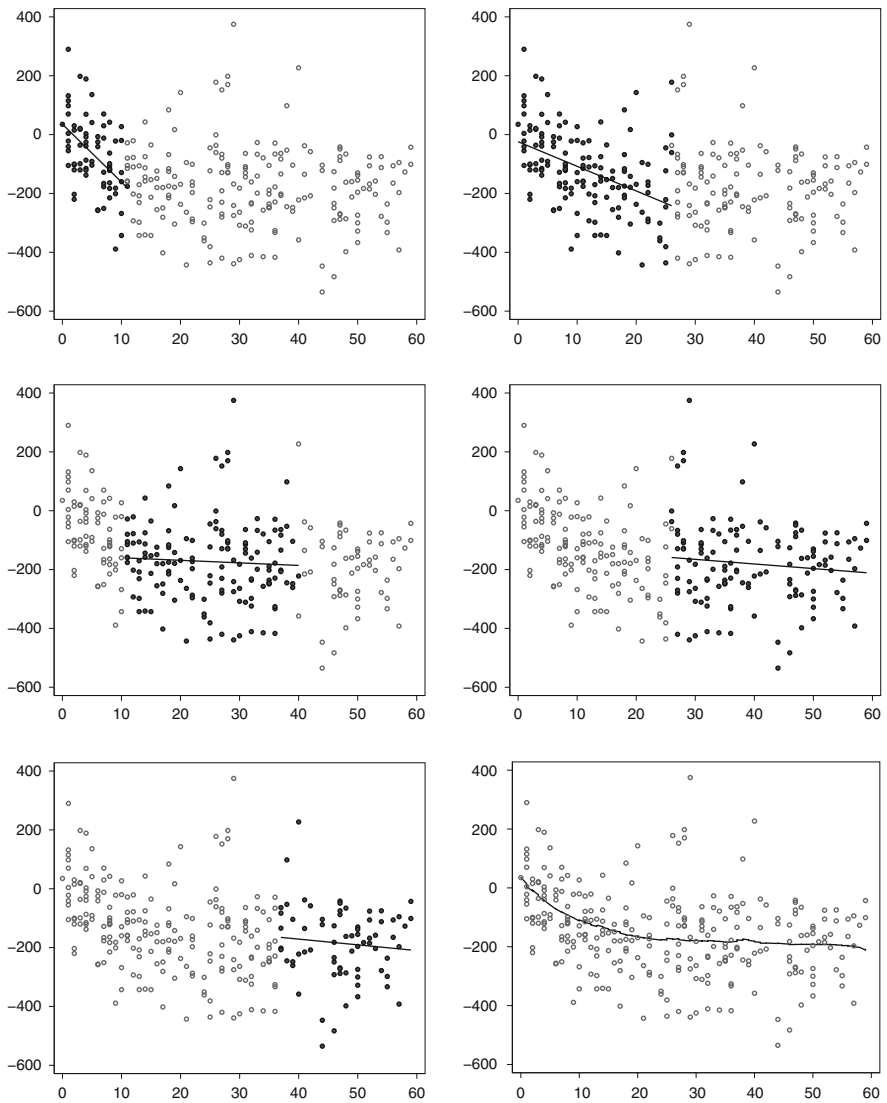


Fig. 2.13 Malnutrition in Zambia: illustration for a nearest neighbor estimator based on the nearest 70 observations (either to the *right* or *left*) of the relationship between Z-score and child's age

continuous. This implies that the regression lines merge continuously at the interval boundaries. Taking this requirement into consideration, we obtain the estimates shown in Fig. 2.14b, which is a special case of *polynomial splines*. Splines are piecewise polynomials, which fulfill certain smoothness conditions at the interval boundaries (also called the knots of the spline). Flexible regression based on splines

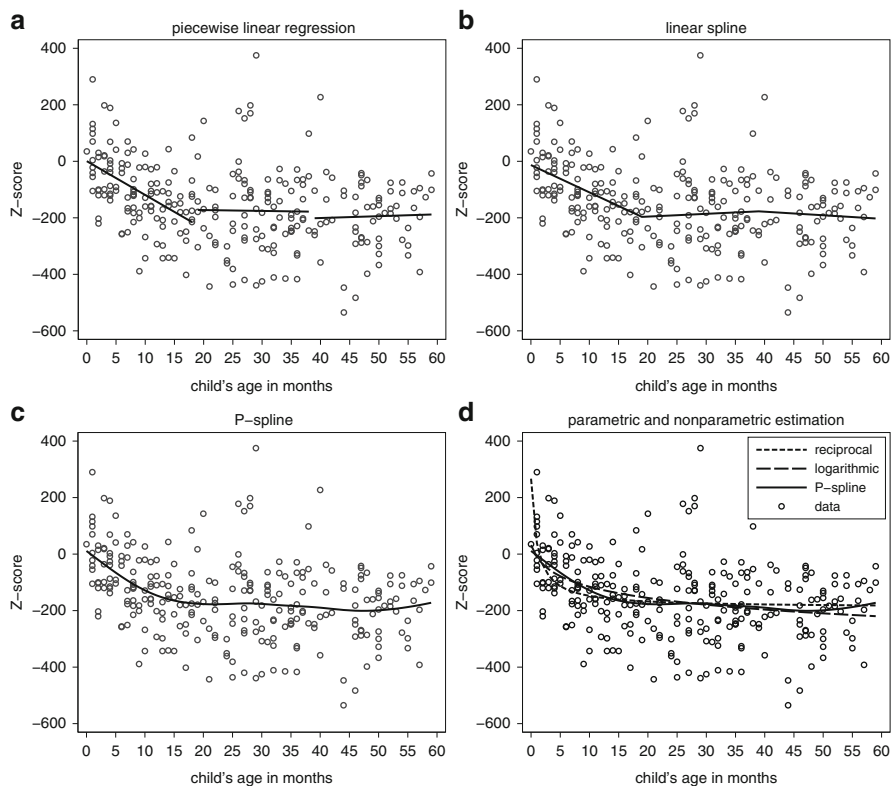


Fig. 2.14 Malnutrition in Zambia: piecewise linear regression [panel (a)], linear spline [panel (b)], and cubic P-spline [panel (c)] for estimating the relationship between Z-score and child's age. Panel (d) shows a comparison of parametric and nonparametric regressions

will play a major role in Chaps. 8 and 9. There we will discuss the most important questions and problems regarding spline regression, for example, how can splines be represented mathematically? How many knots are required to get a satisfactory fit? Where should we place the knots? Figure 2.14c gives a first flavor of the capabilities of spline estimators. It shows a flexible fit to the data based on so-called *P(enalized)-splines*.

On the basis of a nonparametric fit to the data, it is reasonable to search for a simpler parametric functional form that conserves the key features of the function. In this sense we can also understand nonparametric and semiparametric regression as a means of exploratory data analysis that helps to find satisfactory parametric forms. Figure 2.14d shows a comparison of the nonparametric fit with two parametric specifications given by

$$zscore = \beta_0 + \beta_1 \log(age + 1) + \varepsilon$$

2.5 Standard Nonparametric Regression Model

Data

(y_i, x_i) , $i = 1, \dots, n$, with continuous response variable y and continuous covariate x .

Model

$$y_i = f(x_i) + \varepsilon_i, \quad i = 1, \dots, n.$$

We do not assume a simple parametric form for function f . We rather assume certain smoothness characteristics, for example, continuity or differentiability. The same assumptions as in the classical linear regression model apply for errors ε_i .

and

$$zscore = \beta_0 + \beta_1/(age + 1) + \varepsilon.$$

Despite similar functional forms, the nonparametric estimated curve still fits better, particularly in the range of 0–20 months.

2.6 Additive Models

In most applications, as in the examples on the Munich rent index and on malnutrition in Zambia, a moderate (or even large) number of continuous or categorical covariates are available.

Example 2.11 Malnutrition in Zambia

The continuous covariates are *c_age* (age of the child), *c_breastf* (duration of breastfeeding), *m_bmi* (mother's body mass index), *m_height* (mother's height), and *m_agebirth* (mother's age at birth). As in the linear regression model, the categorical covariates *m_education* (mother's education), *m_work* (mother's professional status), and *region* (place of residence) must be dummy coded. Category 2 = "primary school" is chosen as the reference category for the education level. The dummy variables *m_education1*, *m_education3*, and *m_education4* correspond to the education levels "no education," "secondary education," and "higher education," respectively. For the place of residence, we choose the region Cop-perbelt (*region* = 2) as reference category. The variables *region1*, *region3*, ..., *region9* serve as dummy variables for the remaining regions.

For some of the continuous covariates nonlinear effects on the Z-score can be expected (see the scatter plots in Figs. 1.8 and 1.9). For this reason, we specify the *additive model*

$$zscore = f_1(c_age) + f_2(m_bmi) + f_3(m_agebirth) + f_4(m_height) + \beta_0 + \beta_1 m_education1 + \dots + \beta_{11} region9 + \varepsilon \quad (2.19)$$

rather than a linear model. For the moment, the duration of breast-feeding ($c_breastf$) is excluded from the model, since it is highly correlated with the child's age (c_age). See the case study in Sect. 9.8 for more details on modeling the effects of the correlated covariates $c_breastf$ and c_age .

The interpretation of the intercept β_0 and the regression coefficients β_1, β_2, \dots of the categorical covariates $m_education$, m_work , and $region$ is identical to the linear regression model.

Similar to the possibly nonlinear function f in the basic nonparametric regression model (2.18), the functions f_1, f_2, f_3 , and f_4 remain unspecified and are also estimated in a nonparametric way, together with the regression coefficients β_0, β_1, \dots . Even though the model is no longer linear due to the nonlinear effects f_1, \dots, f_4 , it remains additive. As there are no interaction terms between the covariates, it is called an additive main effects model.

Additive models exhibit the following *identification problem*: If we change, e.g., $f_1(c_age)$ to $\tilde{f}_1(c_age) = f_1(c_age) + a$ by adding an arbitrary constant a , and at the same time change β_0 to $\tilde{\beta}_0 = \beta_0 - a$ by subtracting a , the right-hand side of Eq. (2.19) remains unchanged. Hence, the level of the nonlinear function is not identified, and we are forced to impose additional identifiability conditions. This is done, for instance, by imposing the constraints

$$\sum_{i=1}^n f_1(c_age_i) = \dots = \sum_{i=1}^n f_4(m_height_i) = 0,$$

i.e., each nonlinear function is centered around zero. In Fig. 2.15, the estimated functions are constrained in this way. Visualization of the estimated curves as is done in Fig. 2.15 is also the best way to interpret the estimated effects. The effect of the child's age can be interpreted as follows: The average Z-score decreases linearly as the child gets older until 18 months, then stabilizes. There is slight evidence that children older than three years may even show a slight improvement in nutrition condition. Figure 2.15 shows also 80 % and 95 % confidence intervals for the estimated effects. They can be understood as measures for the uncertainty of the effects. In case of the age effect the confidence intervals become wider as age increases. To a large extent the width of confidence intervals reflects the distribution of covariates. Densely populated areas of the covariate domain typically show narrower confidence intervals than sparsely populated areas. For the covariate age the number of observations (slightly) decreases as age increases.

In Sect. 9.8, we will interpret the remaining effects within the scope of a detailed case study.

△

The general form of an additive model (without interaction) is

$$y_i = f_1(z_{i1}) + \dots + f_q(z_{iq}) + \beta_0 + \beta_1 x_{i1} + \dots + \beta_k x_{ik} + \varepsilon_i, \quad (2.20)$$

with the same assumptions for the error term as in the linear regression model. The smooth functions f_1, \dots, f_q represent the (main) effects of the continuous covariates z_1, \dots, z_q and are estimated using nonparametric techniques; see Chap. 9. The covariates x_1, \dots, x_k are categorical or continuous having linear effects. Additive main effect models of the form Eq. (2.20) can be expanded with the inclusion of interaction terms. For two continuous covariates z_1 and z_2 , this can be achieved by adding a smooth two-dimensional function $f_{1,2}(z_1, z_2)$. Compared to (2.20) this leads to the extended predictor

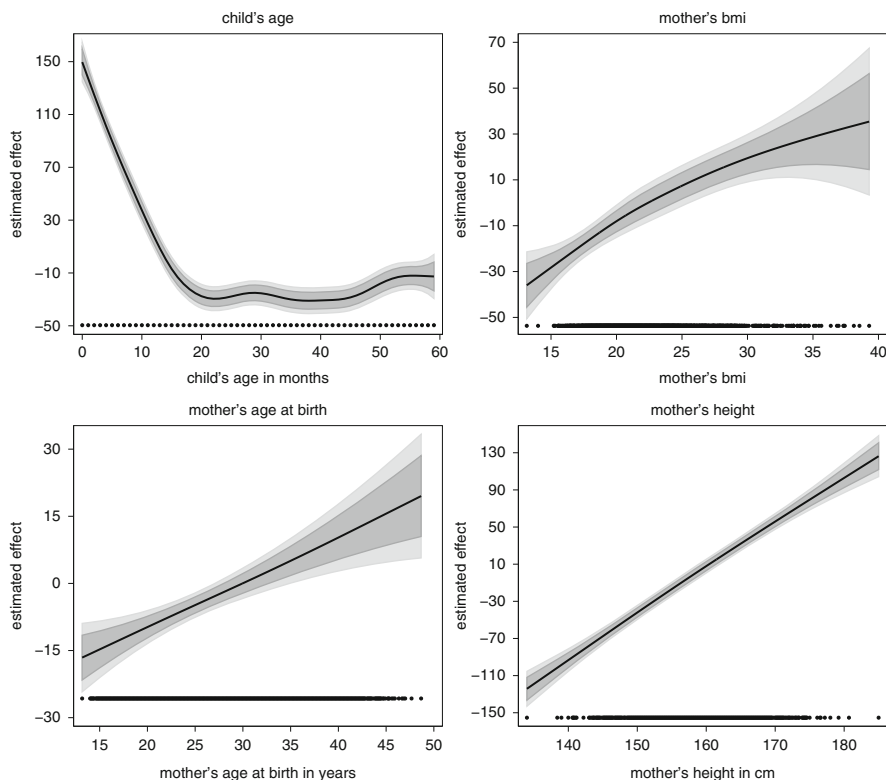


Fig. 2.15 Malnutrition in Zambia: estimated nonlinear functions including 80% and 95% pointwise confidence intervals. The *dots* in the lower part of the figures show the distribution of covariate values. Estimation has been carried out using `remlreg` objects of the software package `BayesX`

$$\eta_i = f_1(z_{i1}) + f_2(z_{i2}) + f_{1,2}(z_{i1}, z_{i2}) + \dots$$

Hence, the interaction effect $f_{1,2}$ modifies the main effects f_1 and f_2 of both covariates. Estimation of the smooth surface $f_{1,2}$ results from the extension of nonparametric techniques for one-dimensional functions to the bivariate case; see Sect. 8.2 for details. An interaction between a continuous covariate z_1 and a binary covariate x_1 is modeled by extending the predictor to

$$\eta_i = f_1(z_{i1}) + \dots + f_q(z_{iq}) + \beta_0 + f_{x_1}(z_{i1})x_{i1} + \dots$$

The interaction term $f_{x_1}(z_1)x_1$, with a smooth function f_x , can be interpreted as a varying effect of x_1 over the domain of z_1 . Models with parametric interactions are covered in detail in section on “Interactions Between Covariates” of Sect. 3.1.3, while Sect. 9.3 focuses on models with nonparametric interactions.

2.6 Standard Additive Regression Models

Data

$(y_i, z_{i1}, \dots, z_{iq}, x_{i1}, \dots, x_{ik}), i = 1, \dots, n$, with y and x_1, \dots, x_k similar to those in linear regression models and additional continuous covariates z_1, \dots, z_q .

Model

$$y_i = f_1(z_{i1}) + \dots + f_q(z_{iq}) + \beta_0 + \beta_1 x_{i1} + \dots + \beta_k x_{ik} + \varepsilon_i.$$

For the errors ε_i the same assumptions as in the classical linear regression model are made. The functions $f_1(z_1), \dots, f_q(z_q)$ are assumed to be “smooth” and represent nonlinear effects of the continuous covariates z_1, \dots, z_q .

A possible approach to estimate additive models is via an iterative procedure, called backfitting, with the simple smoothers (nearest neighbor, splines, etc.) as building blocks. Details will be given in Chap. 9.

2.7 Generalized Additive Models

Nonlinear effects of continuous covariates can also occur in regression models for binary and other non-normal response variables. Similar to the additive models presented in the previous section, it is often preferable to allow for flexible nonparametric effects of the continuous covariates rather than assuming restrictive parametric functional forms. Approaches for flexible and data-driven estimation of nonlinear effects become even more important for non-normal responses, as graphical tools (e.g., scatter plots) are often not applicable to get an intuition about the relationship between responses and covariates.

Example 2.12 Vehicle Insurance

We first illustrate the usage of generalized additive models with the analysis of vehicle insurance data for Belgium in 1997; see Denuit and Lang (2005) for a complete description of the data.

The calculation of vehicle insurance premiums is based on a detailed statistical analysis of the risk structure of the policyholders. An important part of the analysis is the modeling of the claim frequency, which generally depends on the characteristics of the policyholder and the vehicle type. Typical influencing factors of the claim frequency are the policyholder's age (*age*), the age of the vehicle (*age_v*), the engine capacity measured in horsepower (*hp*), and the claim history of the policyholder. In Belgium, the claim history is measured

2.7 Poisson Additive Model

A Poisson additive model $y_i \sim \text{Po}(\lambda_i)$ is defined via the rate

$$\lambda_i = \text{E}(y_i) = \exp(\eta_i)$$

and the additive predictor

$$\eta_i = f_1(z_{i1}) + \dots + f_q(z_{iq}) + \beta_0 + \beta_1 x_{i1} + \dots + \beta_k x_{ik}.$$

Poisson additive models are a special case of generalized additive models for non-normal responses (Chap. 9).

with the help of a 23-step bonus malus score (*bm*). The higher the score, the worse is the insurant's claims history. The statistical analysis is based on regression models with the claim frequency (within 1 year) as the response variable. Since the claim frequency is restricted to the discrete values 0, 1, 2, ..., regression models for continuous response variables are not appropriate.

△

For count data, the Poisson distribution is often assumed for the response, i.e., $y \sim \text{Po}(\lambda)$, with $\lambda = \text{E}(y)$ as the expected number of claims; see Definition B.4 in Appendix B.1 for the Poisson distribution. Our goal is to model the expected number of claims λ as a function of the covariates. Similar to binary responses, the obvious choice of $\lambda = \eta$ with a linear or additive predictor η is problematic, since we cannot guarantee that the estimated expected claim frequency $\hat{\lambda}$ is positive. We therefore assume $\lambda = \exp(\eta)$ in order to ensure a positive expected claim frequency. When using a linear predictor, we obtain a multiplicative model for the expected claim frequency that leads us to a similar interpretation as we already obtained in the logit model:

$$\lambda = \exp(\beta_0 + \beta_1 x_1 + \dots + \beta_k x_k) = \exp(\beta_0) \cdot \exp(\beta_1 x_1) \cdot \dots \cdot \exp(\beta_k x_k).$$

A unit increase in one of the covariates, e.g., x_1 , leads to a change of the expected claim frequency by a factor of $\exp(\beta_1)$. For an additive predictor, we obtain

$$\lambda = \exp(f_1(z_1) + \dots + f_q(z_q) + \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k).$$

Depending on the form of the nonlinear function, the expected count increases or decreases with a unit increase in a covariate. Moreover, in contrast to the purely linear predictor, the change is also dependent on the value of the covariate (because of the nonlinearity). Typically, an increase in x_1 , e.g., from 20 to 21, causes a

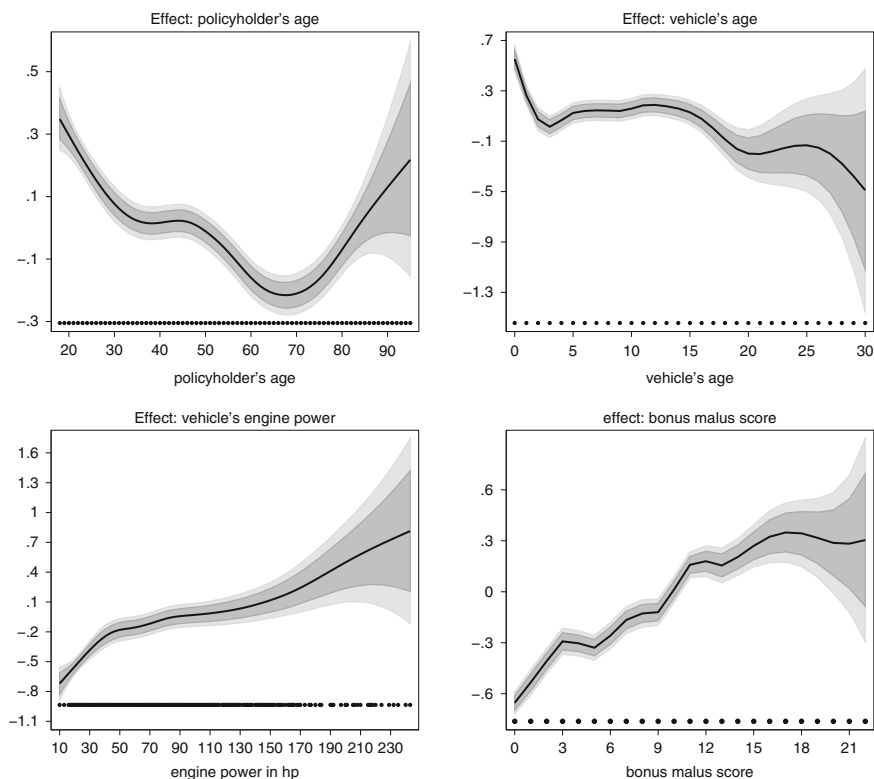


Fig. 2.16 Vehicle insurance: estimated nonlinear functions including 80 % and 95 % pointwise confidence intervals. The dots in the lower part of the figures show the distribution of covariate values. Estimation has been carried out using `remlreg` objects of the software package `BayesX`

different change in the expected count when compared to that of an increase, e.g., from 30 to 31.

Example 2.13 Vehicle Insurance—Additive Model

We model the claim frequencies of the Belgian insurance data using an additive predictor with possibly nonlinear functions of the variables *age*, *age_v*, *hp*, and *bm*:

$$\eta_i = f_1(\text{age}_i) + f_2(\text{age_v}_i) + f_3(\text{hp}_i) + f_4(\text{bm}_i) + \beta_0 + \beta_1 \text{gender}_i + \dots$$

The dots indicate that the predictor may contain other categorical covariates in addition to the continuous variables, e.g., *gender*. We estimated the nonlinear functions and the regression coefficients using the methods presented in detail in Chap. 9. Figure 2.16 shows the estimates $\hat{f}_1, \dots, \hat{f}_4$ for the insurance data. The function related to policyholder's age is notably nonlinear. Initially, the effect on the expected frequency is almost linear until the policyholder reaches the age of 40, then the effect remains nearly constant for several years until the age of 50, then decreases until approximately 70, followed by a rapid increase for

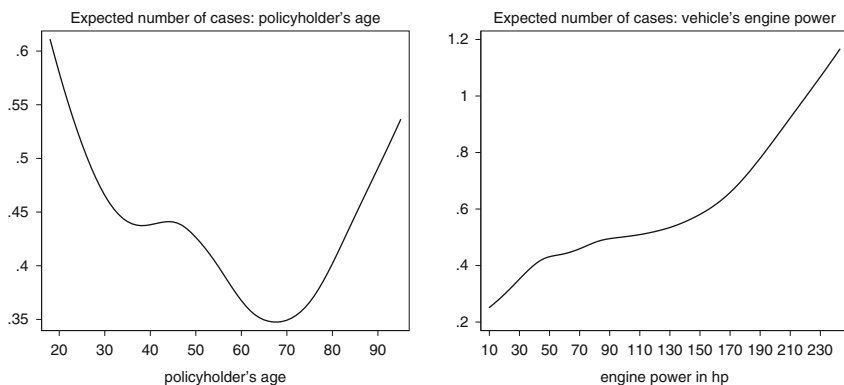


Fig. 2.17 Vehicle insurance: expected number of cases per year among 1,000 policyholders depending on the policyholder's age and vehicle's engine power. The effects of the remaining covariates are held fixed at their mean values

older policyholders. Since we do not have much data on the elderly policyholders, we must be careful with interpretation. This is also reflected by the wide confidence intervals.

Since the expected claim frequency $\lambda = \exp(\eta)$ is a nonlinear function of the covariates, it is not easy to decipher the effect of a particular covariate on λ . It is therefore advisable to plot the estimated expected claim frequency $\hat{\lambda}$ against the covariates. To do so, we plot the estimated rate $\hat{\lambda}$ separately against every continuous covariate while keeping the effects of the remaining covariates fixed at their mean value. See Fig. 2.17 which demonstrates such plots with the covariates *age* and *hp*. Since the expected frequencies are quite low, we plotted $1000\hat{\lambda}$, i.e., the expected claim frequency per year among 1,000 policyholders, rather than the estimated rate $\hat{\lambda}$ itself. For instance, we have plotted

$$1000 \exp \left(\hat{f}_1(\text{age}) + \hat{f}_2(\overline{\text{age-v}}) + \hat{f}_3(\overline{hp}) + \hat{f}_4(\overline{bm}) + \beta_0 + \beta_1 \overline{gender} + \dots \right)$$

against *age* with $\overline{\text{age-v}}$, \overline{hp} , \overline{bm} , and \overline{gender} being the respective covariate means. We observe that the expected number of insurance cases varies between 0.35 and 0.6 per 1,000 policyholders for the *age* variable and between 0.2 and 1.2 for *hp*.

△

2.8 Geoadditive Regression

In addition to the values $(y_i, x_{i1}, \dots, x_{ik}, z_{i1}, \dots, z_{iq})$, $i = 1, \dots, n$, of the response and covariates, many applications contain small-scale geographical information, for example, the residence (address), zip code, location, or county for the individual or unit. For the examples discussed so far, this applies to the data regarding the Munich rent index, malnutrition in Zambia, vehicle insurance, and the health status of trees. In these applications, it is often important to appropriately include geographic information into the regression models in order to capture spatial heterogeneity not covered by the other covariates.

Example 2.14 Malnutrition in Zambia—Geoadditive Model

Example 2.11 (p. 49) already included regional effects using dummy variables for the regions. The corresponding region effects were estimated as categorical “fixed” effects. This conventional approach has two disadvantages. First, information regarding regional closeness or the neighborhood of regions is not considered. Second, if we wish to use small scale information about the district the mother resides, the analogous district-specific approach is difficult or even impossible. Including a separate fixed effect dummy variable for every district results in a model with a large number of parameters, causing a high degree of estimation inaccuracy. Hence, it is better to understand the geographic effect of the variable *district* as an unspecified function $f_{geo}(district)$ and to consider the geographical distance of the districts appropriately when modeling and estimating f_{geo} . A typical assumption is that the regression parameters of two neighboring districts sharing a common boundary should be “similar in size” (a more precise definition of this concept is given in Sect. 8.2). Conceptually, this is very similar to the nonparametric estimation of a smooth function f of a continuous covariate, as, for example, $f(c_age)$. Hence, we reanalyze the data with the following *geoadditive model*:

$$\begin{aligned} zscore = & f_1(c_age) + f_2(m_bmi) + f_3(m_agebirth) + f_4(m_height) \\ & + f_{geo}(district) + \beta_0 + \beta_1 m_education + \dots + \beta_4 m_work + \varepsilon. \end{aligned}$$

We used `remlreg` objects of the software *BayesX* for estimation. In comparison to Example 2.11, the linear part of the predictor no longer contains any region-specific dummy variables. Figure 2.18 shows the map of Zambia, which is divided into color-coded district-specific effects. The geographic effect can now be interpreted similar to a nonlinear effect of a continuous covariate. An effect of, e.g., 40 implies an average Z-score increase of 40 points relative to a district with a zero effect.

The district-specific pattern shows that geographic or spatial effects do not have much in common with the administrative borders. Other causes must be responsible for the spatial effects. In the current situation, the visible north–south divide is due to climatic differences, with a much better nutrition situation in the north. The climatic conditions in the south are worse than in the north, since the southern regions have a much lower altitude compared to the northern ones. In this sense geoadditive models can be understood as an exploratory tool for data analysis: The estimated spatial effects may help to identify geographic covariates that explain the geographic variations.

△

In general, *geoadditive regression* is useful, if in addition to the response variable and continuous or categorical covariates, a *location variable* s_i , $i = 1, \dots, n$, is observed for every unit i . This location variable s can be a location index, as in Example 2.14, with a finite domain $s \in \{1, \dots, S\}$, comprising, e.g., counties, districts etc. In addition, neighborhood information, on the basis of a geographical map or a graph, is available. In other applications, for example, the data on the health status of trees (see Example 1.4), s is a continuous variable containing precise information about the position or the location through geographic coordinates. For flexibly modeling the function f_{geo} , several alternative approaches are available. The choice of a particular model in part depends on whether the location variable is discrete or continuous; see Sect. 8.2 for details.

Geoadditive regression analyses can also be conducted for non-normally distributed responses, especially binary, categorical, or discrete response variables, as in the analysis of the health status of trees or the claim frequency of vehicle insurance policies. In these cases, we expand the predictor η_i in additive logit

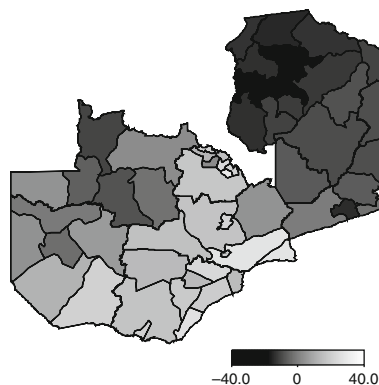


Fig. 2.18 Malnutrition in Zambia: estimated spatial effect

or Poisson models or in generalized additive models to a so-called *geoadditive predictor*

$$\eta_i = f_1(z_{i1}) + \dots + f_q(z_{iq}) + f_{geo}(s_i) + \beta_0 + \beta_1 x_{i1} + \dots + \beta_k x_{ik}.$$

Example 2.15 Vehicle Insurance

It is known that claims associated with vehicle insurance can widely vary across geographic areas. For this reason, many insurance companies report geographically heterogeneous insurance premiums, i.e., the insurance rates differ depending on the policyholder's residence. A realistic modeling of the claim frequency, thus, requires an adequate consideration of the spatial heterogeneity of claim frequencies. In order to do so, we extend the additive predictor of Example 2.13 to

$$\eta_i = f_1(age_i) + f_2(age_v) + f_3(hp_i) + f_4(bm_i) + f_{geo}(district_i) + \dots,$$

where $f_{geo}(district)$ represents a spatial district-specific effect. Figure 2.19 shows the estimated geographic effect obtained using `remlreg` objects of the software `BayesX`: the darker the color, the higher is the estimated effect. The hatched areas mark districts for which we do not have any observations. In comparison to a region with an effect of zero, an effect of approximately 0.3 implies an increase of the expected frequency by a factor of $\exp(0.3) = 1.35$. We find three areas where the expected claims frequencies are clearly higher: the metropolitan areas around Brussels in the center, Antwerp in the North, and Liège in the East of Belgium. The sparsely populated regions in the South show on average lower frequencies.

△

In the following example, we will look at an application taken from survival analysis.

Example 2.16 Survival Analysis of Patients Suffering from Leukemia

The goal of this application is the analysis of covariate effects on the survival time of patients who are diagnosed with a specific type of leukemia. The geographic variation of the survival time is of particular interest, as it might give us information about other risk

Fig. 2.19 Vehicle insurance:
estimated spatial effect

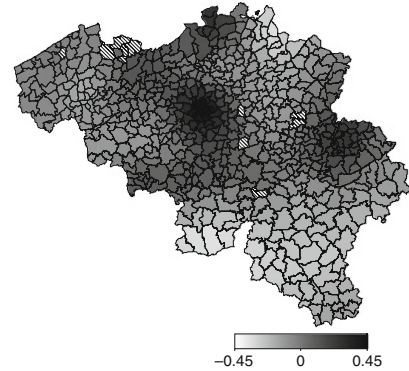
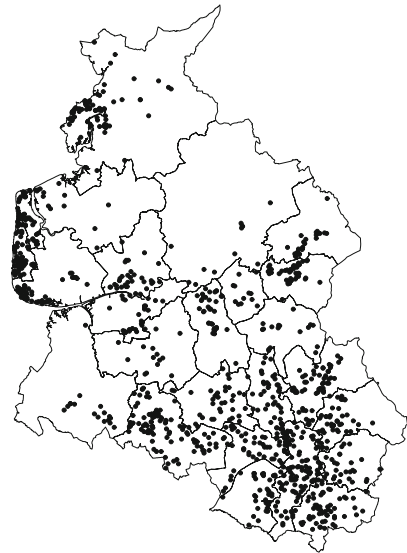


Fig. 2.20 Leukemia data:
spatial distribution of
observations in Northwest
England. Every point
corresponds to one
observation



factors that so far are unknown. The geographical effect may also be closely related to the quality of the health care system in a certain area.

The application studies the survival time of 1,043 patients from the Northwest of England who were diagnosed with acute myeloid leukemia during 1982 through 1998. The data are taken from the British Northwest leukemia register. In addition to the survival time of the patients, there is also information about the following covariates: gender (1 = female, 0 = male), patient's age at the time of diagnosis (*age*), the amount of leucocytes (*lc*), and the Townsend Index (*ti*) that specifies a poverty index of a patient's residential district. A higher value of the Townsend Index reflects a poorer residential district. Geographic information is also included for each patient, as we know the exact coordinates (longitude and latitude) of the patient's residence in Northwest England. Moreover, the patient's residence can be assigned to the particular district of Northwest England. Figure 2.20 shows the geographic distribution of the observations. Approximately 16% of the patients were censored, i.e., they survived the end of the study.

2.8 Geoadditive Models

Data

In addition to the continuous response variable y , the continuous covariates z_1, \dots, z_q , and the remaining covariates x_1, \dots, x_k , there is information about the geographic location s available.

Model

$$y_i = f_1(z_{i1}) + \dots + f_q(z_{iq}) + f_{geo}(s_i) + \beta_0 + \beta_1 x_{i1} + \dots + \beta_k x_{ik} + \varepsilon_i.$$

We make the same assumptions for the error variable ε_i as in the classical linear regression model. The unknown smooth functions f_1, \dots, f_q, f_{geo} and the parametric effects are to be estimated on the basis of the given data.

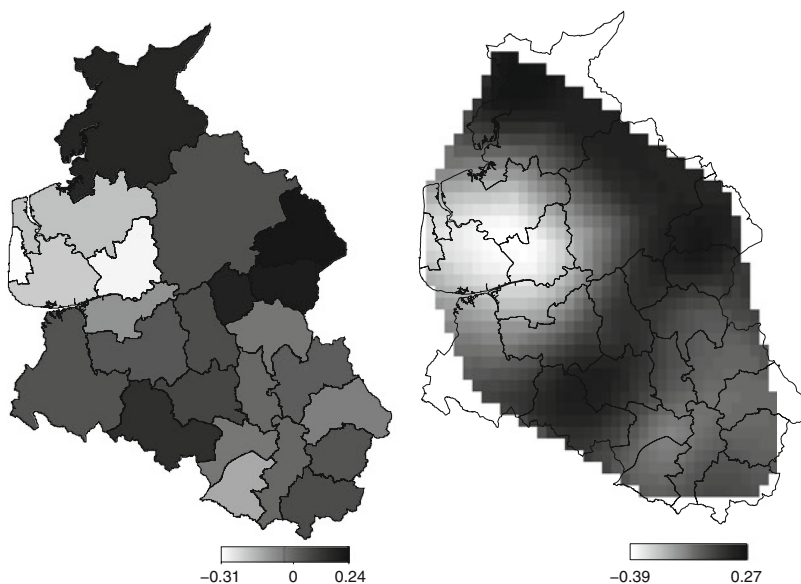


Fig. 2.21 Leukemia data: estimated spatial effect based on districts (*left*) and exact coordinates of the observations (*right*)

In order to estimate the effect of the covariates on the survival time T_i of an individual i , we use *hazard rate models*. The hazard rate $\lambda_i(t)$ of the survival time for individual i is defined as the limit

$$\lambda_i(t) = \lim_{\Delta t \rightarrow 0} \frac{P(t \leq T_i \leq t + \Delta t \mid T_i \geq t)}{\Delta t}.$$

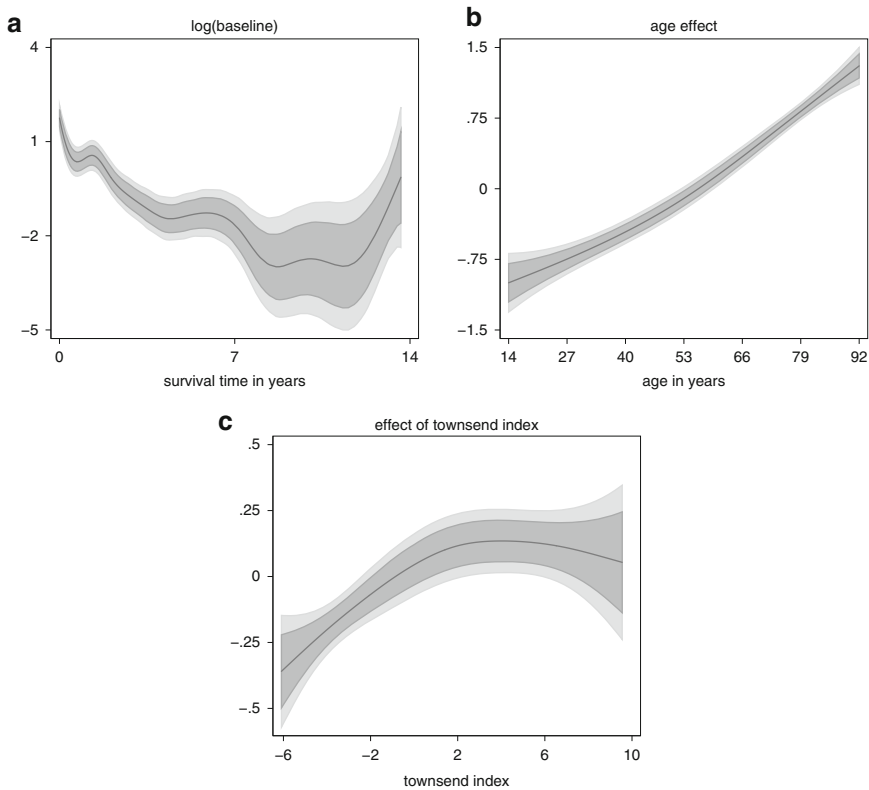


Fig. 2.22 Leukemia data: estimated nonlinear covariate effects with 80 % and 95 % pointwise confidence intervals

The hazard rate $\lambda_i(t)$ therefore characterizes the conditional probability of survival in the interval $[t, t + \Delta t]$, given the individual survived until time t , relative to the interval length Δt . In our application, we use a geoadditive hazard rate model, which links the hazard rate $\lambda_i(t)$ with a geoadditive predictor over the exponential function (similar to the Poisson model of Examples 2.13 and 2.15):

$$\lambda_i(t) = \exp[g(t) + f_1(\text{age}_i) + f_2(\text{ti}_i) + f_{\text{geo}}(s_i) + \beta_0 + \beta_1 \text{lc}_i + \beta_2 \text{gender}_i].$$

This model can be viewed as a generalization of the popular Cox model with simple linear predictors. The model contains nonparametric effects of the continuous covariates *age* and *ti*, linear effects of *lc* and *gender*, as well as a spatial effect, which can either be defined by being based on the exact coordinates or on the districts. We will outline different possibilities of how to model various types of spatial effects in Sect. 8.2. Furthermore, the model also has a time-dependent component $g(t)$, which models the temporal variation of the mortality risk from the time of diagnosis. We refer to the function $g(t)$ as the *log baseline hazard rate* and $\lambda_0(t) = \exp[g(t)]$ as the *baseline hazard rate*.

The following results are based on `remlreg` objects of the software `BayesX`. Figure 2.21 shows the estimated spatial effects and displays obvious geographic variation of the mortality risk. Presumably the geographic effects are surrogates for unobserved covariates, which could to some extent explain the geographic variation. Figure 2.22 shows

the estimated functions $g(t)$, $f_1(\text{age})$, and $f_2(\text{ti})$. The log-baseline hazard rate reflects a decreasing nonlinear trend in mortality risk, up to approximately eight years after the first diagnosis, then followed by an increasing trend. The effect of age has a monotone, almost linear trend, whereas the effect of the Townsend Index indicates that the mortality risk increases in poorer areas (corresponding to higher values in the index) and then remains relatively constant. The estimated effect of the number of leucocytes lc is positive with $\hat{\beta}_1 = 0.003$, but apparently very low. Only when lc is large, the effect of $\hat{\beta}_1 lc_i$ becomes important in size. The estimated effect of gender is very small with $\hat{\beta}_2 = 0.073$; upon further testing, we conclude that gender has little effect on the hazard rate. Δ

Even though regression models for the analysis of survival times play an important role in many fields, this book will not give a detailed presentation, but Sect. 5.8 lists references for further reading. The methodology for the presented example is described in Kneib and Fahrmeir (2007) and Fahrmeir and Kneib (2011).

2.9 Beyond Mean Regression

In the models considered so far, we have restricted ourselves to modeling the (conditional) mean of the response y in dependence of covariates. For example, in the multiple linear regression model of Sect. 2.2.2, we assume independent and normally distributed responses $y_i \sim N(\mu_i, \sigma^2)$, $i = 1, \dots, n$, where the expected value μ_i depends linearly on the covariates in the form

$$\mu_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_k x_{ik}.$$

Other parameters of the response distribution (in case of the normal distribution the variance σ^2) are explicitly assumed to be independent of covariates. In a number of applications, this assumption might not be justified as we will illustrate through the data on the Munich rent index.

Example 2.17 Munich Rent Index—Heterogeneous Variances

Consider Fig. 2.23 which shows scatter plots between the net rent in Euro and the living area (left panel) and year of construction (right panel). Additionally included are estimated regression lines between the response and the covariates. At least the scatter plot between net rent and living area suggests that, additional to the expected value μ , also the variance σ^2 of net rents depends on the covariates. We observe increasing variability as living area increases. Δ

In the next two sections, we present regression models that allow the modeling of other parameters of the response distribution in dependence of covariates, in addition to the expected value. In Sect. 2.9.1 we introduce models with normally distributed responses where the mean *and* the variance depend on covariates. In Sect. 2.9.2 we even go one step further by dropping the normality assumption and modeling the *quantiles* of the response distribution in dependence of covariates.

Another important class of regression models beyond the mean are *hazard regression models* for durations or lifetimes, which are briefly considered in Example 2.16. In fact, it can be shown that a complete specification of the hazard



Fig. 2.23 Munich rent index: scatter plots of rents in Euro versus living area (*left panel*) and year of construction (*right panel*) together with estimated regression lines

rate implies a complete specification of the distribution of a lifetime in dependence of covariates. We do not further discuss hazard rate regression in this book, but refer to the literature cited in Sect. 5.8.

2.9.1 Regression Models for Location, Scale, and Shape

A straightforward approach that extends the multiple linear regression model to cope with variances depending on covariates is to assume $y_i \sim N(\mu_i, \sigma_i^2)$, where in addition to the means

$$\mu_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_k x_{ik},$$

the standard deviations (alternatively the variances)

$$\sigma_i = \alpha_0 + \alpha_1 x_{i1} + \dots + \alpha_k x_{ik} \quad (2.21)$$

depend linearly on the covariates. Similar to logit or probit models and Poisson regression, assumption (2.21) is problematic as it does not guarantee positive standard deviations. Therefore, we replace Eq. (2.21) by

$$\sigma_i = \exp(\alpha_0 + \alpha_1 x_{i1} + \dots + \alpha_k x_{ik}) = \exp(\alpha_0) \exp(\alpha_1 x_{i1}) \cdots \exp(\alpha_k x_{ik}), \quad (2.22)$$

to ensure that the standard deviations are positive. For notational simplicity, we assume exactly the same set of covariates for the expected values μ_i as for the standard deviations σ_i . Of course, this limitation can easily be dropped in practice to allow for different covariates in the mean and the variance equation.

Example 2.18 Munich Rent Index—Linear Model for Location and Scale

We take the data on the Munich rent index and assume the model $rent_i \sim N(\mu_i, \sigma_i^2)$ with

$$\mu_i = \beta_0 + \beta_1 area_i + \beta_2 yearc_i, \quad \sigma_i = \exp(\alpha_0 + \alpha_1 area_i + \alpha_2 yearc_i),$$

for the expected value and standard deviation of the net rents. For simplicity, we restrict ourselves to the two covariates living area and year of construction. Using the R package `gamlss` we obtain the estimates

$$\hat{\mu}_i = -4617.6889 + 5.1847 \cdot area_i + 2.4162 \cdot yearc_i$$

and

$$\hat{\sigma}_i = \exp(8.5235 + 0.0141 area_i - 0.0023 yearc_i).$$

The results for the mean can be interpreted in the usual way as outlined in Sect. 2.2.2:

- Increasing the living area by 1 m² leads to an average increase of the net rent of about 5.18 Euro.
- Modern flats are on average more expensive than older flats. Every year increases the average net rent by 2.42 Euro.

Of course, this interpretation is only meaningful if the chosen linear model is justified (apart from the question whether the net rent per square meter is more appropriate than the plain net rent as a response variable). Figure 2.23 suggests that the linearity assumption is at least problematic for the effect of the year of construction.

Interpretation of the results for the standard deviation is slightly more complicated due to the nonlinearity induced by the exponential link but is similar to Poisson regression (see Sect. 2.7):

- A unit increase of the living area increases the standard deviation by a (small) factor of $\exp(0.014094) = 1.0141938$. This is in line with our observation from the scatter plot in Fig. 2.23 which shows increased variability of net rents as the living area increases.
- A unit increase of the year of construction decreases the standard deviation of net rents by the factor $\exp(-0.002347) = 0.99765575$ which is again close to unity. This estimate is not easily verified through the scatter plot in Fig. 2.23.

△

It is straightforward to generalize mean and variance estimation in Gaussian regression models to nonlinear covariate effects as in additive or geoadditive models. An additive model for location and scale is obtained by generalizing the equations for the mean and standard deviation to

$$\mu_i = f_1(z_{i1}) + \dots + f_q(z_{iq}) + \beta_0 + \beta_1 x_{i1} + \dots + \beta_k x_{ik}$$

and

$$\sigma_i = \exp(g_1(z_{i1}) + \dots + g_q(z_{iq}) + \alpha_0 + \alpha_1 x_{i1} + \dots + \alpha_k x_{ik}).$$

Here, g_1, \dots, g_k are additional smooth functions of the covariates z_1, \dots, z_q . Additive modeling is even more important for the standard deviation as the type of effect (linear or nonlinear) of a certain covariate is much harder (if not impossible) to detect through graphical aids as with mean regression. For instance, the scatter plot between the net rent and the year of construction in the right panel of Fig. 2.23 does not provide clear guidance how to model the effect of year of construction on the standard deviation of the net rent. The following example shows additive models for location and scale in action.

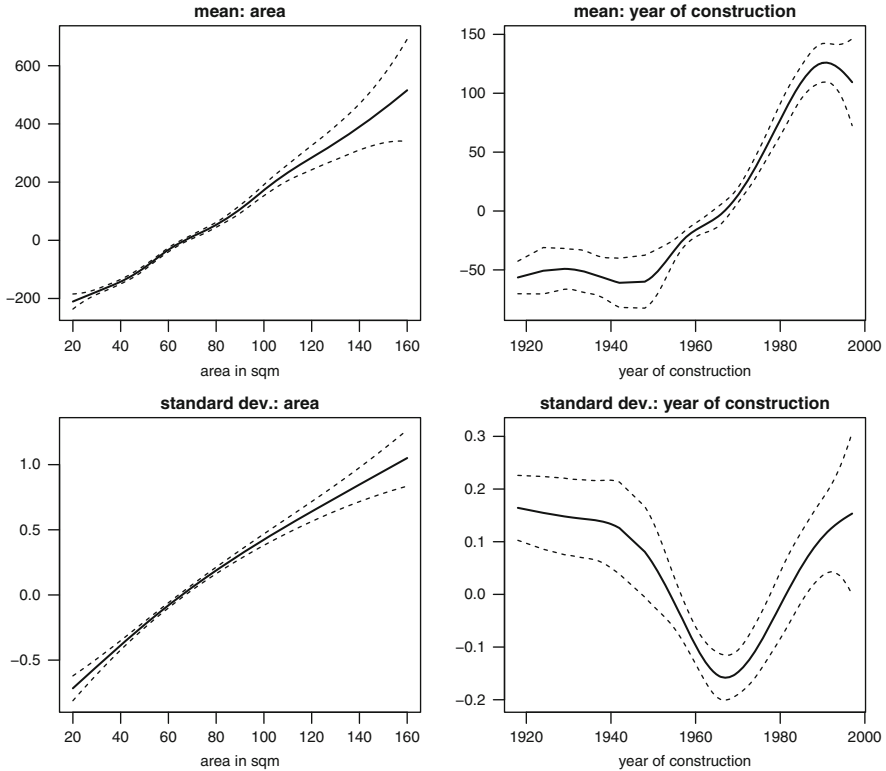


Fig. 2.24 Munich rent index: estimated effects of living area and year of construction for the mean and standard deviation

Example 2.19 Munich Rent Index—Additive Model for Location and Scale



We continue the previous example and assume now possibly nonlinear effects of living area and year of construction in the equations for the mean and the standard deviation:

$$\mu_i = f_1(\text{area}_i) + f_2(\text{year}_i) + \beta_0 \quad \sigma_i = \exp(g_1(\text{area}_i) + g_2(\text{year}_i) + \alpha_0).$$

The resulting estimates including pointwise confidence intervals have been obtained using the R package `gamlss` and are provided in Fig. 2.24. We see that the linear effects of Example 2.18 are (ex post) justified for the area effects but not for the effects of year of construction. The upper left panel of Fig. 2.24 confirms our previous finding that larger flats are on average more expensive than smaller flats with more or less linearly increasing rents. The effect of the year of construction is almost constant until the post-World War II era indicating that flats built before 1945 with otherwise identical living area are on average equally expensive. After World War II, the rents increase almost linearly as the year of construction increases. The effect of living area on the standard deviation is again almost linear implying increasing variability of net rents as the living area increases. The effect of the year of construction is approximately U-shaped with lower variability of net rents in the

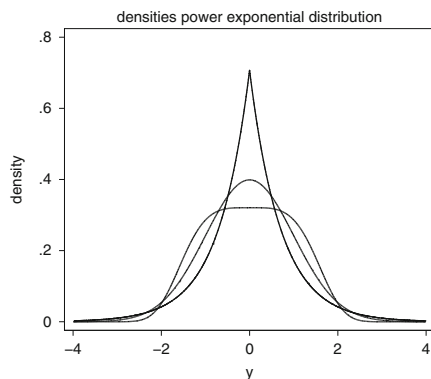


Fig. 2.25 Some densities of the power exponential distribution for $\nu = 1, 2, 4$

1960s and 1970s. This can be explained by a boom in construction building in these years, with flats having comparably homogeneous (typically poor) quality. \triangle

So far we have modeled the mean and the standard deviation of responses as a function of covariates. This type of modeling is a special case of an even more general approach for linear or additive modeling of location, scale, and shape. Generalized additive models for location, scale, and shape (GAMLSS) have been proposed by Rigby and Stasinopoulos (2005). Meanwhile the approach has been fully developed including professional software and inference; see Rigby and Stasinopoulos (2009) and the GAMLSS homepage <http://gamlss.org/> for a full introduction. The approach and corresponding software are able to deal with a huge variety of continuous and discrete distributions for regression modeling. An example is the so-called power exponential distribution whose probability density contains, additional to the mean $\mu \in \mathbb{R}$ and variance $\sigma^2 > 0$, a shape parameter $\nu > 0$ controlling the shape of the density. The probability density is given by

$$f(y) = \frac{\nu \exp\left(-\left|\frac{z}{c}\right|^\nu\right)}{2c\Gamma(1/\nu)},$$

where $c^2 = \Gamma(1/\nu)\Gamma(3/\nu)^{-1}$ is a constant depending on the shape parameter ν , $z = (y - \mu)/\sigma$, and Γ is the gamma function. Compared to the normal distribution, the parameter ν gives the density some extra flexibility to control the shape (to a certain extent). Figure 2.25 shows the density of the power exponential distribution for the three choices $\nu = 1, 2, 4$ of the shape parameter and fixed mean $\mu = 0$ and variance $\sigma^2 = 1$. For $\nu = 2$ we obtain the normal distribution as a special case. Using GAMLSS, we are able to assign additive predictors for each of the three parameters μ , σ^2 , and ν of the power exponential distribution. It is beyond the scope of this book to cover GAMLSS modeling in full detail. However, the GAMLSS literature is readily accessible once the material on additive models and extensions described in Chaps. 8 and 9 has been studied.

2.9.2 Quantile Regression

The GAMLSS framework allows to model the most important characteristics of the response distribution as a function of covariates. However, we still rely on a specific parametric probability distribution like the normal or power exponential distribution. In contrast, quantile regression aims at directly modeling the *quantiles* of the response distribution in dependence of covariates without resorting to a specific parametric distribution family. For $0 < \tau < 1$ let q_τ be the τ -quantile of the response distribution, e.g., $q_{0.75}$ is the 75 % quantile. Then in linear quantile regression we assume

$$q_{\tau,i} = \beta_{\tau,0} + \beta_{\tau,1}x_{i1} + \dots + \beta_{\tau,k}x_{ik},$$

i.e., the quantile q_τ of the response distribution is a linear combination of the covariates as in the multiple linear regression model. Generalizations to additive or geoadditive predictors are conceptually straightforward (although estimation is truly a challenge). The response distribution is implicitly determined by the estimated quantiles q_τ provided that quantiles for a reasonable dense grid of τ -values are estimated. In contrast to the GAMLSS framework, a specific *parametric* distribution is not specified a priori which makes quantile regression a distribution-free approach. The following example gives a flavor of the capabilities of quantile regression. Full details are given in the last chapter of the book (but note that large portions of Chap. 10 on quantile regression are accessible immediately after reading the parts on the classical linear model in Chap. 3 and Sect. 4.1).

Example 2.20 Munich Rent Index—Quantile Regression

We take the rent index data and estimate a linear effect of living area on 11 quantiles q_τ , $\tau = 0.05, 0.1, \dots, 0.9, 0.95$, of the net rent in Euro, i.e.,

$$q_{\tau,i} = \beta_{\tau,0} + \beta_{\tau,1} \cdot \text{area}_i.$$

The top left panel of Fig. 2.26 shows a scatter plot of the net rents versus living area together with estimated quantile regression lines. From top to bottom the lines correspond to the 95 %, 90 %, 80 %, ..., 10 %, and 5 % quantiles. The results are based on the R package `quantreg`. We observe a clear change of the slope (and to a lesser extent also the intercept) of the regression lines with the quantile τ . For higher quantiles, the regression lines are comparably steep indicating a strong effect of the living area on the respective quantile. Note that higher quantiles correspond to the high-price segment of the rent market. As τ decreases, the slopes of the regression lines decrease more and more. For the lowest quantiles, corresponding to the low-price segment of the rent market, the regression lines are almost parallel to a constant line. That is, in the low-price segment, the rents increase very slowly.

We finally point out that estimates for the quantiles of the response distribution can be obtained also on the basis of the linear models based on normally distributed responses. Assuming the simple linear model $\text{rent}_i = \beta_0 + \beta_1 \text{area}_i + \varepsilon_i$ with $\varepsilon_i \sim N(0, \sigma^2)$ implies $\text{rent}_i \sim N(\beta_0 + \beta_1 \text{area}_i, \sigma^2)$ and the quantiles $q_{\tau,i} = \beta_0 + \beta_1 \text{area}_i + \sigma \cdot z_\tau$ where z_τ is the τ -quantile of the $N(0, 1)$ distribution. Thus, assuming a simple linear model with normal errors implies quantile curves that are *parallel* to each other. Assuming a model with linear predictors for location and scale, i.e.,

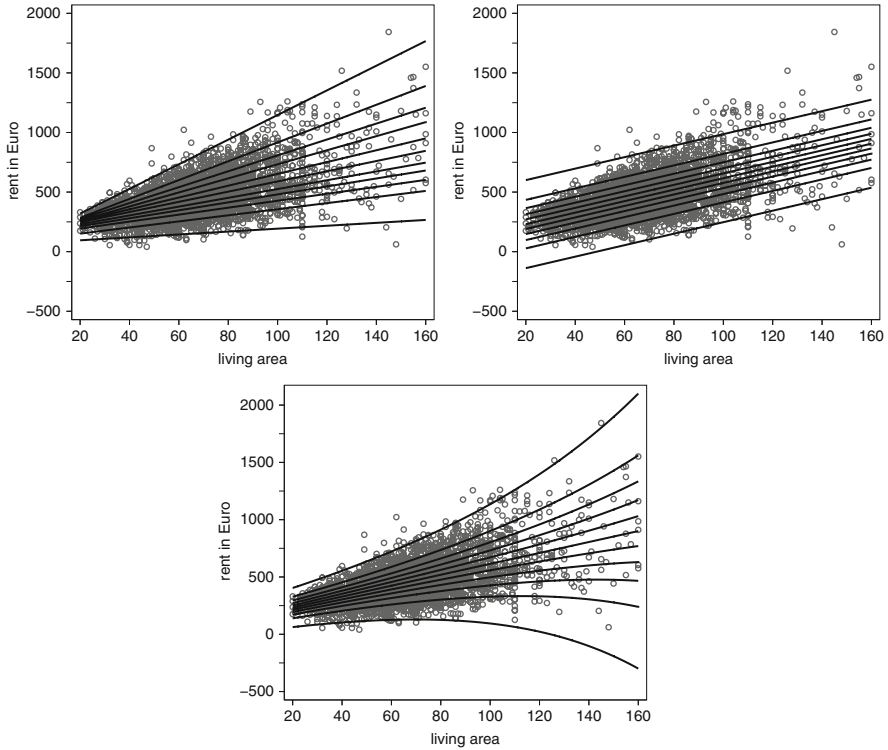


Fig. 2.26 Munich rent index: scatter plots of the rents in Euro versus living area together with linear quantile regression fits for 11 quantiles (*top left panel*), quantiles determined from a classical linear model (*top right panel*), and quantiles determined from a linear model for location and scale (*bottom panel*)

$$rent_i \sim N(\beta_0 + \beta_1 area_i, \sigma_i^2), \quad \sigma_i = \exp(\alpha_0 + \alpha_1 area_i),$$

results in the quantiles

$$q_{\tau,i} = \mu_i + \sigma_i z_\tau = \beta_0 + \beta_1 area_i + \exp(\alpha_0 + \alpha_1 area_i) z_\tau,$$

which are no longer parallel to each other because the standard deviations of the rents depend on the living area. For comparison with completely distribution-free quantile regression, the estimated quantiles based on linear models with normal errors are also included in Fig. 2.26. The top right panel shows estimated quantiles in the simple linear model; the bottom panel displays results if the standard deviation is additionally modeled in dependence of the living area. While the parallel quantile lines of the simple linear model are clearly not adequate, the linear model for location and scale shows reasonable estimated quantiles that are not too far away from the distribution-free estimated quantile curves in the top left row. The largest differences can be observed for very large and low quantiles (95 %, 90 %, 5 %, 10 %). Our comparison shows that parametric regression models, in our case normal regression models for location *and* scale, may well be an alternative to completely distribution-free quantile regression. Particularly promising are the models from the GAMLSS family of regression models.

△

2.10 Models in a Nutshell

We summarize the regression models of this chapter in concise form and indicate in which chapters they are described in more detail. In this way, the common general structure of all models will also become more transparent.

2.10.1 Linear Models (LMs, Chaps. 3 and 4)

- *Response:* Observations y_i are continuous with

$$y_i = \eta_i + \varepsilon_i, \quad i = 1, \dots, n.$$

Errors $\varepsilon_1, \dots, \varepsilon_n$ are i.i.d. with

$$E(\varepsilon_i) = 0, \quad \text{Var}(\varepsilon_i) = \sigma^2.$$

- *Mean:*

$$E(y_i) = \beta_0 + \beta_1 x_{i1} + \dots + \beta_k x_{ik} = \eta_i^{\text{lin}}.$$

- *Predictor:*

$$\eta_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_k x_{ik} = \eta_i^{\text{lin}}.$$

2.10.2 Logit Model (Chap. 5)

- *Response:* Observations $y_i \in \{0, 1\}$ are binary and independently $B(1, \pi_i)$ distributed.
- *Mean:*

$$E(y_i) = P(y_i = 1) = \pi_i = \frac{\exp(\eta_i^{\text{lin}})}{1 + \exp(\eta_i^{\text{lin}})}.$$

- *Predictor:*

$$\eta_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_k x_{ik} = \eta_i^{\text{lin}}.$$

2.10.3 Poisson Regression (Chap. 5)

- *Response:* Observations $y_i \in \{0, 1, 2, \dots\}$ are count data, indicating how often some event of interest has been observed in a certain period of time. In a Poisson model it is assumed that the y_i are independently $\text{Po}(\lambda_i)$ distributed.

- *Mean:*

$$E(y_i) = \lambda_i = \exp(\eta_i^{lin}).$$

- *Predictor:*

$$\eta_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_k x_{ik} = \eta_i^{lin}.$$

2.10.4 Generalized Linear Models (GLMs, Chaps. 5 and 6)

- *Response:* Observations y_i are continuous, categorical, or count data. Depending on the measurement scale and distributional assumptions, they are (realizations of) independent Gaussian, binomial, Poisson, or gamma random variables.

- *Mean:*

$$E(y_i) = \mu_i = h(\beta_0 + \beta_1 x_{i1} + \dots + \beta_k x_{ik}) = h(\eta_i^{lin}),$$

where h is a (known) response function, such as $h(\eta) = \exp(\eta)/(1 + \exp(\eta))$ in a logit model.

- *Predictor:*

$$\eta_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_k x_{ik} = \eta_i^{lin}.$$

- *Remark:* Generalized linear models are a broad class of models, with linear models, logit models, and Poisson models as special cases. Extensions to categorical responses are presented in Chap. 6.

2.10.5 Linear Mixed Models (LMMs, Chap. 7)

- *Response:* Observations y_{ij} , $i = 1, \dots, m$, $j = 1, \dots, n_i$ are continuous with

$$y_{ij} = \eta_{ij} + \varepsilon_{ij}.$$

They are structured in form of longitudinal or clustered data for m individuals or clusters, with n_i observations per individual or cluster. For errors ε_{ij} , we usually make the same assumptions as for linear models. More general error assumptions, taking correlations within individual- or cluster-specific observations into account, are possible.

- *Mean:*

$$\begin{aligned} E(y_{ij}) &= \beta_0 + \beta_1 x_{ij1} + \dots + \beta_k x_{ijk} + \gamma_{0i} + \gamma_{1i} u_{ij1} + \dots + \gamma_{qi} u_{ijq} \\ &= \eta_{ij}^{lin} + \gamma_{0i} + \gamma_{1i} u_{ij1} + \dots + \gamma_{qi} u_{ijq}. \end{aligned}$$

The individual- or cluster-specific random effects γ_{li} , $l = 0, \dots, q$, are assumed to be i.i.d. Gaussian random variables. Alternatively the vector $\boldsymbol{\gamma}_i = (\gamma_{0i}, \dots, \gamma_{qi})'$ is i.i.d. multivariate Gaussian with possibly non-diagonal covariance matrix.

- *Predictor:*

$$\eta_{ij} = \eta_{ij}^{lin} + \gamma_{0i} + \gamma_{1i}u_{ij1} + \dots + \gamma_{qi}u_{ijq}.$$

- *Remark:* LMM with correlated random effects, as well as generalized linear mixed models (GLMMs), will also be considered in Chap. 7.

2.10.6 Additive Models and Extensions (AMs, Chaps. 8 and 9)

- *Response:* Observations y_i are continuous with

$$y_i = \eta_i + \varepsilon_i.$$

For errors ε_i , the same assumptions are made as for linear models.

- *Mean:*

$$E(y_i) = f_1(z_{i1}) + \dots + f_q(z_{iq}) + \eta_i^{lin} = \eta_i^{add}.$$

- *Predictor:*

$$\eta_i = f_1(z_{i1}) + \dots + f_q(z_{iq}) + \eta_i^{lin} = \eta_i^{add}.$$

- *Remark:* Additive models can be extended to include interactions, spatial effects, and random effects. For interactions the predictor is extended to

$$\eta_i = \eta_i^{add} + f_1(z_1, z_2) + \dots$$

or

$$\eta_i = \eta_i^{add} + f(z_1)x_1 + \dots$$

In geoaddivitive models the predictor is extended to

$$\eta_i = \eta_i^{add} + f_{geo}(s_i)$$

with the spatial effect $f_{geo}(s)$ of the location variable s . Incorporation of random effect results in the predictor

$$\eta_{ij} = \eta_{ij}^{add} + \gamma_{0i} + \gamma_{1i}u_{ij1} + \dots$$

The additive model then becomes an additive mixed model (AMM), generalizing linear mixed models.

2.10.7 Generalized Additive (Mixed) Models (GA(M)Ms, Chap. 9)

- *Response:* Observations y_i are continuous, categorical, or count data. Depending on the measurement scale and distributional assumptions, they are (realizations of) independent Gaussian, binomial, multinomial, Poisson, or gamma random variables.

- *Mean:*

$$E(y_i) = \mu_i = h(\eta_i^{add})$$

with (known) response function h .

- *Predictor:*

$$\eta_i = f_1(z_{i1}) + \dots + f_q(z_{iq}) + \eta_i^{lin} = \eta_i^{add}.$$

- *Remark:* Interactions, spatial effects, and random effects can be included as for additive (mixed) models.

2.10.8 Structured Additive Regression (STAR, Chap. 9)

- *Response:* Observations y_i are continuous, categorical, or count data. Depending on the measurement scale and distributional assumptions, they are (realizations of) independent Gaussian, binomial, multinomial, Poisson, or gamma random variables.

- *Mean:*

$$E(y_i) = \mu_i = h(\eta_i)$$

with response function h .

- *Predictor:*

$$\eta_i = f_1(v_{i1}) + \dots + f_q(v_{iq}) + \eta_i^{lin}.$$

The arguments v_1, \dots, v_q are scalar or multivariate variables of different type, constructed from the covariates. Correspondingly, the functions f_1, \dots, f_q are of different type. Some examples are:

$$\begin{aligned} f_1(v_1) &= f(z_1), & v_1 &= z_1, & \text{nonlinear effect of } z_1 \\ f_2(v_2) &= f_{geo}(s), & v_2 &= s, & \text{spatial effect of the location variable } s \\ f_3(v_3) &= f(z)x, & v_3 &= (z, x), & \text{effect of } x \text{ varying over the domain of } z \\ f_4(v_4) &= f_{1,2}(z_1, z_2), & v_4 &= (z_1, z_2), & \text{nonlinear interaction between } z_1 \text{ and } z_2 \\ f_5(v_5) &= \gamma_i u, & v_5 &= u, & \text{random effect of } u. \end{aligned}$$

- *Remark:* Structured additive regression reflects the fact that the predictor includes effects of different type in structured additive form. All model classes discussed so far are special cases of STAR models.

2.10.9 Quantile Regression (Chap. 10)

- *Response:* Observations y_i are continuous and independent with generally unspecified distribution.
- *Quantiles:* Quantile regression models the quantiles q_τ , $0 < \tau < 1$, of the response distribution using a linear or additive predictor. The most general

predictor is a STAR predictor, i.e.,

$$q_{\tau,i} = f_{\tau,1}(v_{i1}) + \dots + f_{\tau,q}(v_{iq}) + \eta_{\tau,i}^{lin}$$

with variables v_j and functions f_j as in Sect. 2.10.8.

Regression

Models, Methods and Applications

Fahrmeir, L.; Kneib, Th.; Lang, S.; Marx, B.D.

2013, XIV, 698 p., Hardcover

ISBN: 978-3-642-34332-2