Chapter 7
Regression

Thus far, we have looked at different approaches to modelling random variables with parametric distributions. Let us now address the very common situation where additional variables are associated with the distribution of our random variable. In other words, we have input variables x that influence the distribution of our output variable Y. To mediate this influence, we allow the parameters θ of the distribution of Y to depend upon x. More formally, we attempt to model the conditional distribution of Y given x

$$Y|x \sim f(y; \theta(x)),$$

where the parameter θ is influenced by x, possibly in a complex manner. This setting is generally called **regression** in statistics. Often, the mean value of Y is the parameter of interest and is modelled in response to the input variables x, i.e. $E(Y|x)$. However, other quantities, e.g. quantiles, can be considered as well. In the following chapter, we will present commonly used statistical regression models and end with an extended example to demonstrate their flexibility and usability.

From here on, we call the output variable Y the dependent variable and the input variable x the independent variable. In econometrics, Y is often called the endogenous variable and x the exogenous variable and in other strands of literature Y the response variable and x the covariate. We use these terms interchangeably in the subsequent chapters and have included Table 7.1 as a reference.

7.1 Linear Model

The problem of modelling a quantity x that influences the outcome Y is rather broad and regression models as described above are just one of the many possible approaches. Regression analyses have the advantage that they allow for model

Table 7.1 Alternative terms for Y and x in the regression setting

Y	x
Endogenous variable	Exogenous variable
Output variable	Input variable
Response variable	Covariate
Dependent variable	Independent variable
Label	Feature
Target	Covariable

interpretation. The regression coefficient β quantifies the influence of x on Y and allows its prediction when only x is known. We stress, however, that regression models are only one tool for modelling the influence of x on Y and a swathe of alternative techniques are available. This includes approaches like classification and regression trees, neural networks, support vector machines and so on. We will not cover these tools here but refer to the literature, e.g. Hastie et al. (2009).

7.1.1 Simple Linear Model

The classical linear model involves quantifying how the mean of the response variable Y depends on the covariate x. This is modelled as follows:

$$Y = \beta_0 + x\beta_x + \varepsilon. \tag{7.1.1}$$

The quantity ε is called the **error term** and the parameters are the intercept β_0 and the slope β_x. Model (7.1.1) is known as the linear regression model and even traces back to Gauß (1777–1855), who was the first to propose estimates of β_0 and β_x. The relationship between Y and x can be causal in nature, that is, one can assume that there is a direct influence of x on y. We will explicitly discuss this in Chap. 12. For now, we simply consider Model (7.1.1) as relating the distribution of Y to the independent variable x. To make this more explicit, we assume that covariate x is known and that the distribution of Y depends conditionally upon x. Secondly, we assume that the error term ε has zero mean, i.e. $E(\varepsilon|x) = 0$, which in turn implies that

$$E(Y|x) = \beta_0 + x\beta_x.$$

Hence, the conditional mean of Y is given by the linear predictor $\beta_0 + x\beta_x$. In other words, we have a linear relation between Y and x, which is disturbed by an error ε. These are the two essential assumptions in a regression model. The latter is often replaced by postulating normality of the error terms, that is, one assumes

$$\varepsilon|x \sim N(0, \sigma^2)$$

or equivalently

$$Y|x \sim N(\beta_0 + x\beta_x, \sigma^2).$$

Additionally, it is necessary that the distribution of ε does not depend on x, which is clearly stronger than simply assuming a vanishing mean $E(\varepsilon|x) = 0$. This also implies that the variance of ε (and hence of Y) does not depend on x. This lack of influence is called **variance homogeneity** or synonymously **(variance) homoskedasticity**. In contrast, if the variance depends on x, we call this **variance heterogeneity** or equivalently **(variance) heteroskedasticity**. We will demonstrate methods for addressing homoskedasticity later in this chapter.

An applied example of a regression model is given in Fig. 7.1, where x is the floor size of an apartment and Y the corresponding rent, with data taken from the Munich rental guide 2015. We see that the larger the apartment, the higher the rent. We also include the fitted linear line in the plot, where the intercept β_0 is 135.50 and

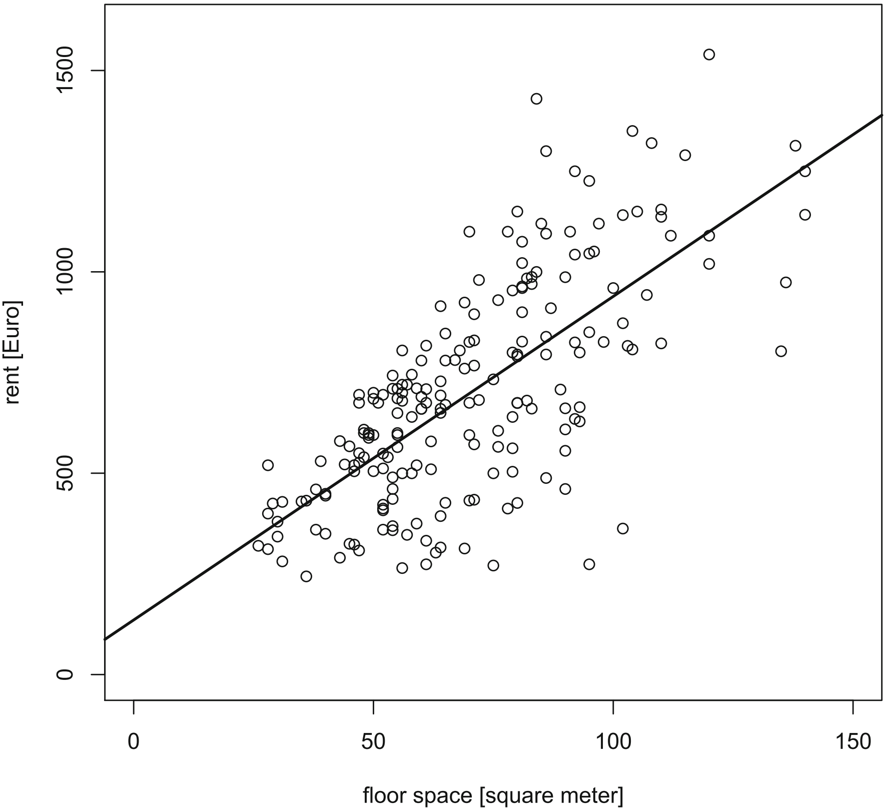

Fig. 7.1 Rent (in euros) for apartments with a given floor size. Taken from the data used to produce the Munich rental table 2015

the slope β_x is 8.04. This implies that the average rent increases by 8.04 euros per square metre.

Let us now turn our attention to estimating the parameters of our model that best fit the data. Suppose that data pairs $(y_i, x_i), i = 1, \ldots, n$, are available, as in Fig. 7.1. Assuming normality and independence for the error terms ε_i gives the log-likelihood (ignoring constant terms)

$$l(\beta_0, \beta_x, \sigma^2) = \sum_{i=1}^{n} \left\{ -\frac{1}{2} \log \sigma^2 - \frac{1}{2} \frac{(y_i - \beta_0 - x_i \beta_x)^2}{\sigma^2} \right\}. \tag{7.1.2}$$

Differentiating the log-likelihood with respect to β_0, β_x and σ^2 gives

$$\begin{aligned}
\frac{\partial l(\beta_0, \beta_x, \sigma^2)}{\partial \beta_0} &= \sum_{i=1}^{n} \frac{(y_i - \beta_0 - x_i \beta_x)}{\sigma^2} \\
\frac{\partial l(\beta_0, \beta_x, \sigma^2)}{\partial \beta_x} &= \sum_{i=1}^{n} x_i \frac{(y_i - \beta_0 - x_i \beta_x)}{\sigma^2} \\
\frac{l(\beta_y, \beta_x, \sigma^2)}{\partial \sigma^2} &= -\frac{n}{2\sigma^2} + \frac{1}{2} \sum_{i=1}^{n} \frac{(y_i - \beta_0 - x_i \beta_x)^2}{\sigma^4}.
\end{aligned}$$

Setting these to zero gives the Maximum Likelihood estimates

$$\begin{aligned}
\hat{\beta}_0 &= \frac{1}{n} \sum_{i=1}^{n} y_i - \frac{1}{n} \sum_{i=1}^{n} x_i \hat{\beta}_x \\
\hat{\beta}_x &= \sum_{i=1}^{n} x_i (y_i - \hat{\beta}_0) \Big/ \sum_{i=1}^{n} x_i^2 \\
\hat{\sigma}^2 &= \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - x_i \hat{\beta}_x)^2.
\end{aligned} \tag{7.1.3}$$

The estimates $\hat{\beta}_0$ and $\hat{\beta}_x$ are also known as least square estimates $\hat{\beta}_0$ and $\hat{\beta}_x$, which minimise the squared distance between prediction and true values $\sum_{i=1}^{n} (y_i - \beta_0 - x_i \beta_x)^2$. This is clear in Eq. (7.1.2), where exactly this expression must be minimised. The calculation of the estimates becomes much simpler if we rewrite the whole model in matrix notation. To do so, let

$$\begin{aligned} \boldsymbol{y} &= (y_1, \ldots, y_n)^T \in \mathbb{R}^{n\times 1} \\ \boldsymbol{X} &= \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \in \mathbb{R}^{n\times 2} \\ \boldsymbol{\varepsilon} &= (\varepsilon_1, \ldots, \varepsilon_n)^T \in \mathbb{R}^{n\times 1}. \end{aligned}$$

We can then write Model (7.1.1) in the form

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

where $\beta = (\beta_0, \beta_x)^T$ and $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^T$. Matrix $\boldsymbol{X}$ is also called the **design matrix**. Note that the column with entries "1" corresponds to including the intercept in the model and is the default setting. Given the data $\boldsymbol{y}$, the likelihood is written as

$$l(\beta, \sigma^2) = -\frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{X}\beta)^T(\boldsymbol{y} - \boldsymbol{X}\beta). \tag{7.1.4}$$

Using matrix notation to get the derivative with respect to β gives us

$$\frac{\partial l(\beta, \sigma^2)}{\partial \beta} = \boldsymbol{X}^T(\boldsymbol{y} - \boldsymbol{X}\beta).$$

This gives the estimate

$$\hat{\beta} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}, \tag{7.1.5}$$

which is identical to the estimates in (7.1.1) but clearly has a more convenient form than the individual derivatives given above. The estimate (7.1.5) has a number of welcome properties, which can be easily derived. Firstly, the estimate is unbiased because

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

Secondly, as $Var(\boldsymbol{\varepsilon}) = \sigma^2 I_n$, where I_n is the diagonal matrix of dimension n, it can be seen that

$$Var(\hat{\beta}) = \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}. \tag{7.1.6}$$

It is also not difficult to show that $\sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ is equal to the inverse Fisher information matrix. Hence, the variance of the estimate is minimal. The estimate is also the best linear unbiased estimate (BLUE), i.e. the best unbiased linear estimator of the type $\hat{\beta} = AY$ with $A \in \mathbb{R}^{2\times n}$. This property is known as the **Gauß–Markov theorem**. The correlation of the estimates $\hat{\beta}_0$ and $\hat{\beta}_x$ is expressed in (7.1.6), where $\boldsymbol{X}^T\boldsymbol{X}$ is the 2×2 matrix

$$\boldsymbol{X}^T\boldsymbol{X} = \begin{pmatrix} n & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \end{pmatrix}.$$

This demonstrates that if $\sum x_i = 0$, then $\hat{\beta}_0$ and $\hat{\beta}_x$ are uncorrelated. Note that for a single covariate, we can always achieve this by redefining

$$\tilde{x}_i = x_i - \bar{x} \Leftrightarrow x_i = \tilde{x}_i + \bar{x},$$

such that the model becomes

$$Y_i = \beta_0 + x_i\beta_x + \varepsilon_i = \beta_0 + \tilde{x}_i\beta_x + \bar{x}\beta_x + \varepsilon_i = \tilde{\beta}_0 + \tilde{x}_i\beta_x + \varepsilon_i.$$

The intercept changes from β_0 to $\tilde{\beta}_0$, but the slope remains unchanged. For $\tilde{x}_i$, we find that $\sum_{i=1}^n \tilde{x}_i = \sum_{i=1}^n x_i - n\bar{x} = 0$, such that the Fisher information matrix for $\tilde{\beta}_0$ and β_x is given by

$$Var\begin{pmatrix} \hat{\tilde{\beta}}_0 \\ \hat{\beta}_x \end{pmatrix} = \sigma^2 \begin{pmatrix} n & 0 \\ 0 & \sum \tilde{x}_i^2 \end{pmatrix}^{-1}.$$

Furthermore, the matrix formulation enables us to make various extensions of the simple model (7.1.1). For instance, let x be a binary covariate of the form $x \in \{0, 1\}$. Because we did not make any assumptions about the values of the x variable, the different model formulation (7.1.1) remains valid. This also applies for (7.1.5).

7.1.2 Multiple Linear Model

Let us now assume that x is a discrete covariate which takes k different values labelled $\{1, \ldots, k\}$. Model (7.1.1) can then be rewritten as

$$Y = \beta_0 + 1_{\{x=1\}}\beta_1 + 1_{\{x=2\}}\beta_2 + \ldots + 1_{\{x=k-1\}}\beta_{k-1} + \varepsilon,$$

where $1_{\{.\}}$ is the indicator function, which takes value 1 if the statement in the brackets is true and zero if it is false. Note that we only need $k-1$ indicator variables, because the intercept covers the case where x takes value k. We can construct the matrix $\boldsymbol{X}$ as

$$\boldsymbol{X} = \begin{pmatrix} 1 & 1_{\{x_1=1\}} & \dots & 1_{\{x_1=k-1\}} \\ \vdots & \vdots & & \vdots \\ 1 & 1_{\{x_n=1\}} & \dots & 1_{\{x_n=k-1\}} \end{pmatrix}$$

with $\boldsymbol{y}$ and $\boldsymbol{\varepsilon}$ defined as above. This gives

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

where $\beta = (\beta_0, \beta_1, \dots, \beta_{k-1})^T$. Clearly, the estimate $\hat{\beta}$ is given by (7.1.5), but now has dimension k, i.e. $k-1$ coefficients β_1 to β_{k-1} plus the intercept β_0. The same model formulation applies if we have more than one covariate. Let, for instance, z be a second covariate, such that Model (7.1.1) extends to

$$Y = \beta_0 + x\beta_x + z\beta_z + \varepsilon.$$

We then define $\boldsymbol{X}$ as

$$\boldsymbol{X} = \begin{pmatrix} 1 & x_1 & z_1 \\ \vdots & \vdots & \vdots \\ 1 & x_n & z_n \end{pmatrix}.$$

Using data (y_i, x_i, z_i) again gives (7.1.5) as an estimate. This demonstrates the generality of the matrix formulation of the model, which does not need to be altered when including more covariates or parameters.

Given the structure of $\hat{\beta}$, we see that the Maximum Likelihood estimate $\hat{\beta}$ fulfils the distributional property

$$\hat{\beta} \sim N(\beta, \sigma^2(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X})^{-1}). \tag{7.1.7}$$

For the existence of the estimator $\hat{\beta}$, we need $(\boldsymbol{X}^T\boldsymbol{X})$ to be invertable, which holds if $\boldsymbol{X}$ has full rank.

Note that σ^2 also needs to be estimated and $\hat{\sigma}^2$ in (7.1.3) can easily be written in matrix notation. It appears, however, that the estimate should be bias corrected. Note that the Maximum Likelihood estimate fulfils

$$\hat{\sigma}^2 = (\boldsymbol{y} - \boldsymbol{X}\hat{\beta})^T(\boldsymbol{y} - \boldsymbol{X}\hat{\beta})/n.$$

The observable values $\hat{\varepsilon} = y - \boldsymbol{X}\hat{\beta}$ are called fitted residuals, such that $\hat{\sigma}^2 = \hat{\varepsilon}^T\hat{\varepsilon}/n$.

We define with

$$\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$$

the **hat matrix**, which is idempotent, meaning that $\boldsymbol{H}\boldsymbol{H} = \boldsymbol{H}$ and $\boldsymbol{H^T} = \boldsymbol{H}$. Given the structure of the estimate $\hat{\beta}$, we can then write $\boldsymbol{y} - \boldsymbol{X}\hat{\beta} = (\boldsymbol{I_n} - \boldsymbol{H})\boldsymbol{y}$ with $\mathbf{I}_n$ as n dimensional identity matrix. If we now replace the observed values $\boldsymbol{y}$ with random variables $\boldsymbol{Y}$, we can calculate the mean value of $\hat{\sigma}^2$. In fact, using linear algebra rules for the trace $tr(.)$ of a matrix, we get

$$\begin{aligned} E((\boldsymbol{Y} - \boldsymbol{X}\hat{\beta})^T(\boldsymbol{Y} - \boldsymbol{X}\hat{\beta})) &= E(\boldsymbol{Y}^T(\boldsymbol{I}_n - \boldsymbol{H})^T(\boldsymbol{I}_n - \boldsymbol{H})\boldsymbol{Y}) \\ &= E(\mathrm{tr}(\boldsymbol{Y}^T(\boldsymbol{I}_n - \boldsymbol{H})\boldsymbol{Y})) \\ &= E(\mathrm{tr}((\boldsymbol{I}_n - \boldsymbol{H})\boldsymbol{Y}\boldsymbol{Y}^T)) \\ &= \mathrm{tr}((\boldsymbol{I}_n - \boldsymbol{H})E(\boldsymbol{Y}\boldsymbol{Y}^T)) \\ &= \mathrm{tr}((\boldsymbol{I}_n - \boldsymbol{H})(\boldsymbol{X}\beta\beta^T\boldsymbol{X}^T + \sigma^2\boldsymbol{I}_n)) \\ &= \sigma^2\mathrm{tr}(\boldsymbol{I}_n - \boldsymbol{H}), \end{aligned}$$

where the final simplification follows as $(\boldsymbol{I}_n - \boldsymbol{H})\boldsymbol{X}\beta\beta^T\boldsymbol{X}^T = 0$. Because

$$\begin{aligned} \mathrm{tr}(\boldsymbol{I}_n - \boldsymbol{H}) &= n - \mathrm{tr}(\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T) \\ &= n - \mathrm{tr}((\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{X}) = n - p, \end{aligned}$$

we find the variance estimate to be biased, where p is the number of columns in $\mathbf{X}$. It is therefore common to replace the variance estimate with its unbiased counterpart

$$s^2 = \frac{(\boldsymbol{y} - \boldsymbol{X}\hat{\beta})(\boldsymbol{y} - \boldsymbol{X}\hat{\beta})}{n-p}.$$

If we now also replace σ^2 in (7.1.7) with its unbiased estimate s^2, we obtain a multivariate t-distribution for $\hat{\beta}$ with $n - p$ degrees of freedom

$$\hat{\beta} \sim t_{n-p}(\beta, s^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}). \tag{7.1.8}$$

However, if n is large, it is reasonable to use the normal distribution described in (7.1.7), even if σ^2 is estimated with s^2.

The multiple linear regression model described by the linear equation

$$E(Y|x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p$$

is a very useful and flexible tool for modelling associations between Y and regressor variables $x_1, \ldots, x_p$. The regression coefficients β_k can be interpreted as follows. If x_k is increased by 1, then the expectation of Y is increased by β_k, if all other covariates remain fixed. This is a typical "ceteris paribus" interpretation. The coefficient β_k describes the pure association when adjusting for all other covariates.

Furthermore, we make no assumptions about the nature of the variables $x_1, \ldots, x_p$ and therefore have many alternative model structures at our disposal. These include

- indicator variables $x_k \in \{0, 1\}$ to model the association between nominal variables and Y;
- transformed variables, i.e. defining for example $x_k = \log(z)$, which indicates a logarithmic relationship between z and Y;
- using polynomials or splines, e.g. $x_1 = z, x_2 = z^2, x_3 = z^3$ for polynomial regression;
- using products of variables, e.g. use x_1, x_2 and $x_3 = x_1x_2$. In this case, the term x_1x_2 is called an **interaction term**.

Regression is an essential tool in statistics and there are many extensions, even beyond the material discussed in this chapter. We refer to Fahrmeir et al. (2015) for an extensive discussion of the field.

Example 26 In a clinical trial, patients suffering from high blood pressure were treated psychologically. Patients were randomly assigned to three groups: the control group (S1) did not receive psychological treatment. The other two groups of patients received either 1 or 2 therapy sessions (S2) or more than 2 therapy sessions (S3). Blood pressure was measured at the beginning and at the end of the study. We use the following linear regression model

$$Y_i = \beta_0 + BPS_i\beta_1 + 1_{\{i \in S2\}}\beta_2 + 1_{\{i \in S3\}}\beta_3 + \varepsilon_i \tag{7.1.9}$$

to model the relationship between the final blood pressure Y and the type of therapy. The variables $1_{\{i \in S2\}}$ and $1_{\{i \in S3\}}$ indicate whether person i belongs to group S2 or group S3, respectively. The control group is the reference and β_2 and β_3 can be interpreted as the difference in blood pressure between the control group and their respective groups, for a given starting blood pressure. The covariate BPS_i gives the blood pressure at onset, which should have an effect on the final blood pressure Y. The fitted regression coefficients are given in Table 7.2. The intercept is difficult to interpret, but we can see that the blood pressure at onset has a significant effect. This

Table 7.2 Regression estimates and standard errors

	Estimate	Std. error	p-Value
β_0	76.4	9.8	<0.001
β_1	0.31	0.08	<0.001
β_2	−4.7	1.9	0.114
β_3	−6.07	1.8	0.0014

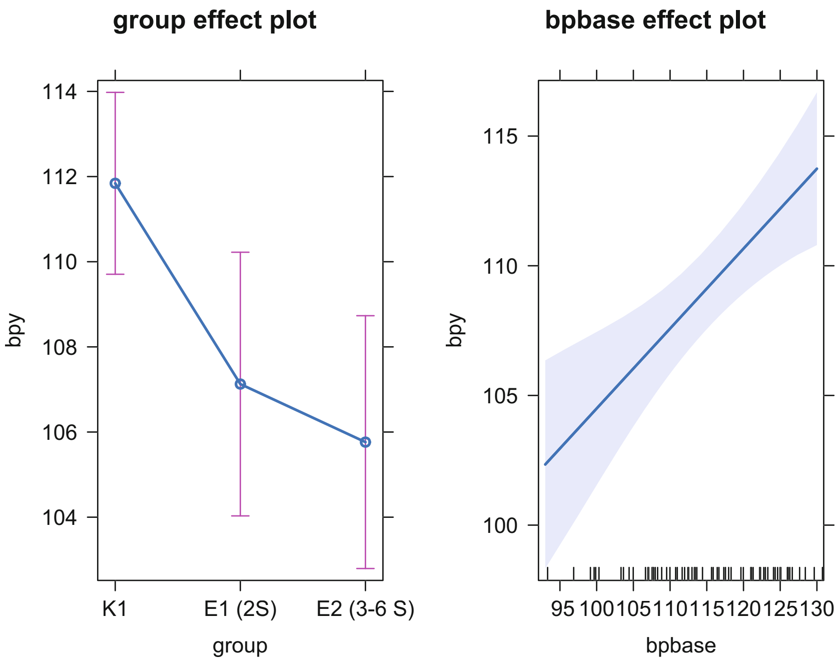

Fig. 7.2 Visualisation of the fitted effects

can be seen from the given p-value, which tests $H_0 : \beta_1 = 0$ with the test statistic in (7.1.8). As a simple rule of thumb, the standardised estimate

$$\frac{\hat{\beta}_1}{\sqrt{\widehat{Var}(\hat{\beta}_1)}} \tag{7.1.10}$$

can also be used, which asymptotically follows a standard normal distribution. Hence, if the absolute value of the ratio (7.1.10) is larger than the $1 - \alpha/2$ quantile of the $N(0, 1)$ distribution, this indicates a rejection of the hypothesis $H_0 : \beta_1 = 0$, with significance level α. The result shows a significant reduction in blood pressure in both treatment groups, meaning that the estimated effect in the treatment group S3 is significant. The results are visualised in Fig. 7.2.

▷

7.1.3 *Bayesian Inference in the Linear Model*

Thus far, we have treated β as an unknown model parameter, which justifies the use of Maximum Likelihood estimation. We could also take a Bayesian view and impose a prior distribution on the parameters β and σ^2. Taking this perspective will also give us an insight into the similarities between Bayesian estimation and Maximum Likelihood estimation.

To begin with, we need to choose a prior. We will assume, for simplicity, a flat prior for β, that is, $f_\beta(\beta) = \text{const}$. This prior is degenerate, meaning that $\int f_\beta(\beta)d\beta = \infty$ and is, therefore, not a proper density. However, as discussed previously, prior distributions may be degenerate but still lead to proper posterior distributions, which happens to be the case here. For the prior of the variance, we assume that $f_\sigma(\sigma^2) \propto \sigma^{-2}$, which is a flat prior on $\log \sigma^2$ and also degenerate. The posterior is given by

$$f_{\beta,\sigma^2}(\beta, \sigma^2|\boldsymbol{y}) \propto (\sigma^2)^{-\frac{n}{2}+1} \exp\left(-\frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{X}\beta)^T(\boldsymbol{y} - \boldsymbol{X}\beta)\right).$$

Let us now determine the conditional posterior of β given σ^2, i.e. $\beta|\sigma^2, \boldsymbol{y}$. We rewrite $\boldsymbol{y} - \boldsymbol{X}\beta$ as $(\boldsymbol{y} - \boldsymbol{X}\hat{\beta}) - (\boldsymbol{X}\beta - \boldsymbol{X}\hat{\beta})$ with $\hat{\beta}$ as Maximum Likelihood estimate as defined above. This gives

$$\begin{aligned}(\boldsymbol{y} - \boldsymbol{X}\beta)^T(\boldsymbol{y} - \boldsymbol{X}\beta) = (\boldsymbol{y} - \boldsymbol{X}\hat{\beta})^T(\boldsymbol{y} - \boldsymbol{X}\hat{\beta}) + (\hat{\beta} - \beta)^T\boldsymbol{X}^T\boldsymbol{X}(\hat{\beta} - \beta)\\ - \underbrace{2(\boldsymbol{y} - \boldsymbol{X}\hat{\beta})^T\boldsymbol{X}(\hat{\beta} - \beta)}_{0}.\end{aligned}$$

The final component vanishes, with the definition of $\hat{\beta}$ in (7.1.5). This can be seen because

$$\begin{aligned}(\boldsymbol{y} - \boldsymbol{X}\hat{\beta})^T\boldsymbol{X}(\hat{\beta} - \beta) =&(\boldsymbol{y} - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y})^T(\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} - \boldsymbol{X}\beta)\\ =&\boldsymbol{y}^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})\boldsymbol{X}^T\boldsymbol{y} - \boldsymbol{y}^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}-\\ &\boldsymbol{y}^T\boldsymbol{X}\beta + \boldsymbol{y}^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{X}\beta\\ =&0.\end{aligned}$$

Defining $s^2 = (\boldsymbol{y} - \boldsymbol{X}\hat{\beta})(\boldsymbol{y} - \boldsymbol{X}\hat{\beta})/(n-p)$ as above, with p as the dimension of β, the posterior can then be rewritten as

$$f_{\beta,\sigma^2}(\beta, \sigma^2|\boldsymbol{y}) \propto(\sigma^2)^{-(\frac{n-p}{2}+1)} \exp\left\{-\frac{n-p}{2\sigma^2}s^2\right\}(\sigma^2)^{-\frac{p}{2}} \exp\left\{-\frac{1}{2\sigma^2}(\beta - \hat{\beta})^T\boldsymbol{X}^T\boldsymbol{X}(\beta - \hat{\beta})\right\}. \tag{7.1.11}$$

This shows that, conditional on σ^2, the parameter vector β has a normal posterior, i.e.

$$\beta|\sigma^2, \boldsymbol{y} \sim N(\hat{\beta}, (\boldsymbol{X}^T\boldsymbol{X})^{-1}\sigma^2). \tag{7.1.12}$$

Note that (7.1.12) is the same as the distribution of the Maximum Likelihood parameter estimate (7.1.7), but with estimate and parameter reversed. We will take a closer look at this interesting phenomenon in a moment.

The marginal posterior for σ^2 can be obtained by integrating out β in the latter component of (7.1.11), which is proportional to $|\sigma^2\mathbf{X}^T\mathbf{X}|^{\frac{1}{2}} \propto (\sigma^2)^{-\frac{p}{2}}$. Consequently, the last two multiplicative components in (7.1.11) cancel out and no longer depend upon σ^2. The first two components are proportional to an inverse gamma distribution (see Chap. 5), such that

$$\sigma^2|\boldsymbol{y} \sim \text{Inv Gamma}\left(\frac{n-p}{2}, \frac{s^2(n-p)}{2}\right).$$

These statements hold for the given prior distributions. We refer to Box and Tiao (1973) for more details on the above calculations. Note that the posterior factorises to

$$f_{\beta,\sigma^2}(\beta, \sigma^2|\boldsymbol{y}) = f(\beta|\sigma^2, \boldsymbol{y})f(\sigma^2|\boldsymbol{y})$$

and integrating out σ^2 gives the marginal posterior distribution for β. It can be shown that this is, in fact, a p-dimensional t-distribution such that

$$\beta|y \sim t_{n-p}(\hat{\beta}, s^2(X^TX)^{-1}).$$

We have reached a rather interesting relation between the parameter posterior and the ML estimate. The posterior of β is equal to the distribution of the estimate $\hat{\beta}$, just with β and $\hat{\beta}$ reversed. We observed the same similarity in the previous section, when we looked at the mean of the normal distribution.

We may now draw inference about β from either the Frequentist or the Bayesian perspective. The former takes β as a fixed, but unknown, parameter, while the latter looks at its posterior distribution. In fact, taking the above flat prior with $f_\beta(\beta)$ and $f_\sigma(\sigma^2) = \sigma^{-2}$ (which is again flat for $\log(\sigma^2)$), we have the same distribution for $\beta - \hat{\beta}$, regardless of whether we consider β as random and $\hat{\beta}$ as fixed (the Bayesian view) or, conversely, $\hat{\beta}$ as random and β as fixed (the Frequentist view).

We can use the example in Fig. 7.1, which related the rent of an apartment to its floor size, to help explain how the model can be interpreted. The parameter estimates $\hat{\beta}$ are listed in Table 7.3. The intercept is estimated with $\hat{\beta}_0 = 135.47$ and the fitted

Table 7.3 Parameter estimates and variance estimates for rental data

	Estimate	Std. error	t value	p value p-Value
(Intercept)	135.47	43.29	3.13	0.002
fspc	8.04	0.58	13.76	< 2e−16

slope with $\hat{\beta}_x = 8.04$. Hence, one could say that the average rent increase per square metre is in the order of $8.04\,€/qm^2$. The Fisher matrix is given by

$$\hat{\sigma}^2(\boldsymbol{X}^T\boldsymbol{X})^{-1} = \begin{pmatrix} 1873.97 & -23.95 \\ -23.65 & 0.34 \end{pmatrix}.$$

Assuming β_0 and β_x are parameters with estimates $\hat{\beta}_0$ and $\hat{\beta}_x$ gives the standard errors listed in the second column above, e.g. $\sqrt{1873.97} = 43.29$. The third column gives the standardised value of the ratio $\hat{\beta}_0/\sqrt{Var(\hat{\beta}_0)}$ and $\hat{\beta}_x/\sqrt{Var(\hat{\beta}_x)}$. The fourth column lists the p-values that result from testing hypotheses $H_0 : \beta_0 = 0$ and $H_0 : \beta_x = 0$. Clearly, we can conclude that floor space significantly influences the rent of an apartment (which is no surprise).

Let us use the above results to derive confidence intervals or credibility regions in the Bayesian case. Note that

$$\hat{\delta}_0 = \hat{\beta}_0 - \beta \quad \text{and} \quad \hat{\delta}_x = \hat{\beta}_x - \beta$$

follow a multivariate t-distribution with $n-2$ degrees of freedom. The corresponding isolines of this density are visualised in the left-hand plot of Fig. 7.3. We may now derive a credibility region R such that

$$P(\hat{\delta} \in R) = 0.95$$

with $\hat{\delta} = (\hat{\delta}_0, \hat{\delta}_x)$. This region is indicated by the solid ellipse in Fig. 7.3, which is in fact both a confidence region and a posterior credibility region. Hence, $\hat{\beta}_0 - \beta$ can take values between approximately -100 and 100, while $\hat{\beta}_x - \beta_x$ can take values

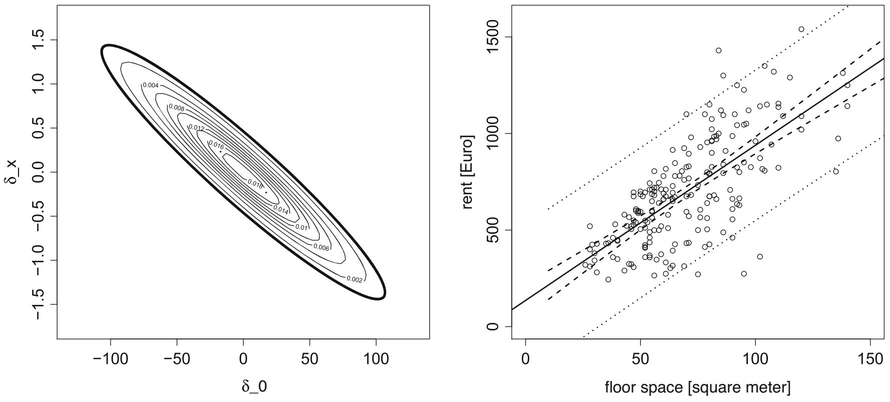

Fig. 7.3 Left plot: isolines of multivariate t-distribution resulting from rental guide data. Right plot: confidence region or credibility region for linear influence of floor space on rent (α=0.05, dashed lines) and prediction interval ($1-\alpha$=0.95, dotted lines)

between approximately -1.5 and 1.5. Taking values (δ_0, δ_x) within the confidence region R gives different intercepts and slopes for the trend line. This is mirrored in the right-hand plot of Fig. 7.3, where the dashed lines represent the bounds of R, the confidence region of δ_0 and δ_x. They also represent the confidence bounds on β_0 and β_x.

Note that this confidence interval represents the uncertainty of both the Frequentist and the Bayesian approaches on the *linear relation* between floor space and rent. It does not mirror the range that the actual rent might take. This quantity is called the **prediction interval** and, in this case, would depict the range of possible rents for a given floor space. With the prediction interval, we aim to predict the rent of an apartment which is not in the database. Assume for a given floor space x_0, we want to predict the rent. We know that the rent is given by

$$Y = \beta_0 + x_0\beta_x + \varepsilon. \tag{7.1.13}$$

If we replace the parameters in the model above with their estimates, this gives

$$\hat{Y} = \hat{\beta}_0 + x_0\hat{\beta}_x + \varepsilon, \tag{7.1.14}$$

where clearly the error term ε is unobserved. Taking the expectation gives

$$E(\hat{Y}|x_0) = \hat{\beta}_0 + x_0\hat{\beta}_x.$$

We can also derive the variance of the prediction, which is given by

$$Var(\hat{Y}|x) = Var(\hat{\beta}_0 + x\hat{\beta}_x) + \sigma^2.$$

The first component is the estimation variability, which is expressed in the confidence interval, as previously discussed. Because the (new) residual ε is independent of the observed values, the variance decomposes additively to the estimation variance and the residual variance. In the Bayesian formulation, this looks very similar, but now we do not replace β_0 and β_x in (7.1.13) with estimates but work with the posterior distribution directly. Consequently, the variance in the rent of a new apartment is $Var(\hat{Y}|x) = Var(\beta_0 + x\beta_x|\boldsymbol{y}) + Var(\varepsilon|\boldsymbol{y})$, where we condition on the data $\boldsymbol{y}$. In both cases, the variance decomposes additively, as the inference on β_0 and β_x is based on the data, i.e. the data points seen in Fig. 7.1. We include the prediction interval as a dotted line in the right-hand plot of Fig. 7.3. Both the confidence and prediction intervals are shown for a $(1 - \alpha)$ level with $\alpha = 0.05$.

7.2 Weighted Regression

This section will cover how the idea of weighting can be used in linear regression. Let us first assume that the variance of Y_i and hence the variance of the error terms ε_i are different for each observation. This is usually called variance heterogeneity, which we denote as

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

σ_i^2 can now depend on some of the covariates, i.e. $\sigma_i^2 = \sigma^2(x_i)$. This could, for instance, be modelled as $\sigma_i = \exp(x_i \gamma)$. In this case, we get a regression model where both the mean *and* the variance depend upon the covariates. Estimation can then be carried out with the standard Maximum Likelihood approximation, see Kneib (2013). We will not fully exploit the flexibility of this model here, but just consider σ_i as known to an individual multiplicative constant, i.e. $\sigma_i^2 = a_i \sigma^2$, where a_i is known. Note that the model is not identifiable, which means we need to put an extra constraint on a_i to have a unique representation. The common setting is that $\sum_{i=1}^n a_i = n$, such that the variance homogeneous model results from the special case $a_i = 1$ for $i = 1, \ldots, n$. The log-likelihood is then given by

$$\begin{aligned} l(\beta, \sigma^2) &= -\frac{n}{2} \log \sigma^2 - \frac{1}{2} \sum_{i=1}^{n} \frac{(y_i - x_i \beta)^2}{a_i \sigma^2} \\ &= -\frac{n}{2} \log \sigma^2 - \frac{1}{2\sigma^2} (\boldsymbol{y} - \boldsymbol{X}\beta)^T \boldsymbol{W} (\boldsymbol{y} - \boldsymbol{X}\beta), \end{aligned}$$

where $\boldsymbol{W} = diag(\frac{1}{a_1}, \ldots, \frac{1}{a_n})$. The ML estimate for β is given by the weighted least squares estimate

$$\hat{\beta} = (\boldsymbol{X}^T \boldsymbol{W} \boldsymbol{X})^{-1} (\boldsymbol{X}^T \boldsymbol{W} \boldsymbol{y}). \tag{7.2.1}$$

It is easy to see that $\hat{\beta}$ is unbiased, because $E(Y|X) = \boldsymbol{X}\beta$, which inserted in (7.2.1) shows $E(\hat{\beta}) = \beta$. Moreover, the variance is given by

$$\begin{aligned} Var(\hat{\beta}) &= (\boldsymbol{X}^T \boldsymbol{W} \boldsymbol{X})^{-1} (\boldsymbol{X}^T \boldsymbol{W} Var(\boldsymbol{Y}) \boldsymbol{W} \boldsymbol{X}) (\boldsymbol{X}^T \boldsymbol{W} \boldsymbol{X})^{-1} && (7.2.2) \\ &= \sigma^2 (\boldsymbol{X}^T \boldsymbol{W} \boldsymbol{X})^{-1}, && (7.2.3) \end{aligned}$$

as the variance of the observations is $Var(\boldsymbol{Y}) = \sigma^2 \boldsymbol{W}$. In this case, we can rely on the inverse Fisher matrix as variance estimate for $\hat{\beta}$.

A different form of weighting occurs if we focus on the mean of Y, which often occurs in **survey weighting**. This becomes an issue when the data themselves are biased, for instance, by the over-representation of particular groups of individuals or when observations are not identically distributed. To make this clear, let us look at the following simplified example. Let, as above, Y be the rent of an apartment of floor size x. The data come from two different groups: 50% from city-run housing societies, which are known to provide comparably cheap accommodation, and 50% from private landlords, whose apartments are relatively expensive.

However, in reality, only 25% of the apartments are rented out by housing societies, while 75% are offered by private landlords. Clearly, the data are biased as the 50% of the data represent only 25% of the rental market, while the other 50% represent 75%. Under these circumstances, weighting can again play a useful role. The weights in this case are commonly called survey weights and would be $w_i = 75/50 = 3/2$ for observations from private and commercial landlords and would be set to $w_i = 25/50 = 1/2$ for city-run apartments. A correct use of survey weights in regression is deceptively complex and one must first assume a model where rents of the one group of landlords behave differently than the other. Hence, we need to in fact assume two separate regression models. This idea extends to stratified sampling. In our example, this refers to the fact that the apartments in the city fall into two different groups, or "strata". See DuMouchel and Duncan (1983) and Gelman (2007) for a deeper discussion and also Hu and Zidek (2002), who generalise the idea towards weighted likelihoods. In this example, we will keep things simple and just include weights in the least squares. This leads to (7.2.1), which is however a biased estimate.

To understand this bias, let us look again at the rental example. We assume that apartments from city-run housing organisations have a different (average) monthly rent than those of private and commercial landlords. Let z_i be an indicator variable for the i-th apartment expressing whether the apartment is city run ($z_i = 1$) or commercially run ($z_i = 2$). Then, assuming the same variance in the two groups, we have

$$Y_i | x_i, z_i \sim N(x_i \beta_{(z_i)}, \sigma^2),$$

where $\beta_{(1)}$ and $\beta_{(2)}$ are the slope parameters in the two groups. The population parameter, that is the slope when the groups are omitted, is

$$\beta = 0.25\beta_{(1)} + 0.75\beta_{(2)} = P(Z_i = 1)\beta_{(1)} + P(Z_i = 2)\beta_{(2)}.$$

With these prerequisites, we can now calculate the expectation of $\hat{\beta}$, which leads us to

$$E(\hat{\beta}) = (\boldsymbol{X}^T \boldsymbol{W} \boldsymbol{X}) \sum_{i=1}^{n} w_i x_i^T x_i \beta_{(z_i)},$$

which clearly is not equal to β. Hence, the bias is difficult to calculate. If, however, the distribution of the covariate x_i is independent of z_i, then the estimate is asymptotically unbiased. Instead of looking at the bias of the estimate, let us focus on its variance, which, assuming the same variance in both groups, is given by

$$\begin{aligned} Var(\hat{\beta}) &= (\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X})^{-1}(\boldsymbol{X}^T\boldsymbol{W}Var(\varepsilon)\boldsymbol{W}\boldsymbol{X})(\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X})^{-1} \\ &= (\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X})^{-1}(\sum_{i=1}^{n} w_i^2 x_i^T Var(Y_i - x_i\beta)x_i)(\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X}) \qquad (7.2.4) \\ &= \sigma^2(\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X})^{-1}(\boldsymbol{X}^T\boldsymbol{W}^2\boldsymbol{X})(\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X})^{-1}. \qquad (7.2.5) \end{aligned}$$

The variance has a sandwich-type structure and clearly differs from (7.2.3) unless all weights are equal. Note also that the weights do not need to sum up to 1, as any normalisation of the weights cancels out in (7.2.5). A direct estimate of (7.2.4) was proposed by Huber (1967), which is better known as the Eicker–White estimator (see White, 1980, or Mage, 1998). The idea is to replace $Var(Y_i - x_i\beta)$ with its empirical counterpart (assuming unbiasedness), i.e.

$$\widehat{Var}(\hat{\beta}) = (\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X})^{-1}(\sum_{i=1}^{n} w_i^2 x_i x_i^T (y - x_i\hat{\beta})^2)(\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X})^{-1}.$$

By default, standard software packages that allow for weighted regression do not provide the variance estimate (7.2.5) but instead (7.2.3). Hence, weighting is considered to account for variance heterogeneity but not for survey weights, which should be kept in mind when applying survey weighting in regression. This approach using weights can also be extended to the case of dependent error terms. Then the weighting matrix is not a diagonal matrix and reflects the dependence structure within the error terms. For details, see Fahrmeir et al. (2015), Chap. 4.

7.3 Quantile Regression

The above regression model plays a central role in statistics and a number of extensions and generalisations have been proposed over the last few decades. One of these is quantile regression, which has proven itself to be both practical and powerful in many situations. The most essential piece of literature in the field is “Quantile Regression” by Koenker (1996). To motivate the idea behind quantile regression, let us first repeat the definition of a quantile. A quantile can be comprehended as the inverse of the distribution function. The τ-quantile of a distribution function $F(y)$ is defined as

$$Q(\tau) = \inf\{y : F(y) \geq \tau\} \qquad (7.3.1)$$

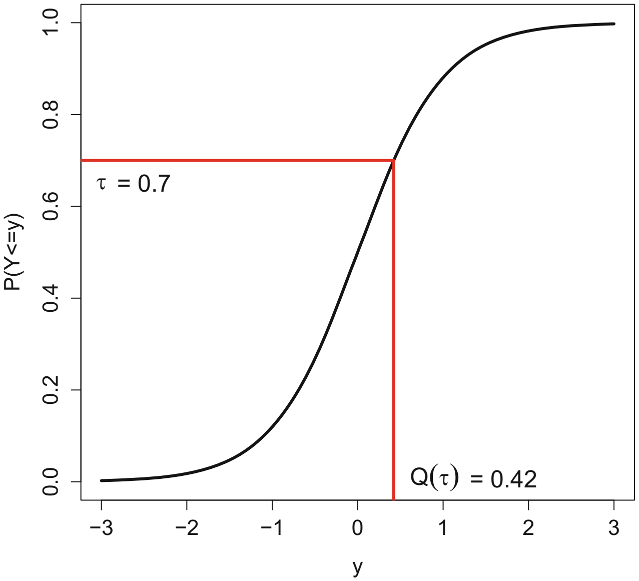

Fig. 7.4 Sketch of the definition of a quantile

for $\tau \in (0, 1)$. This can be seen in Fig. 7.4. Note that if $F(.)$ is invertable, even if only on a subset of $\mathbb{R}$, then $Q(\tau)$ is the inverse of $F(.)$ such that $F(Q(\tau)) = \tau$. A commonly used quantile is the median, defined as $Q(0.5)$, but the quartiles $Q(0.25)$ and $Q(0.75)$ are also used in practice. The idea of quantile regression is now to model the relationship between x and the quantile of the variable Y, in contrast to linear regression which models the *mean* of Y. To do so, we condition everything in (7.3.1) on x, that is

$$Q(\tau|x) = \inf\{y : F(y|x) \geq \tau\},$$

where $F(y|x)$ denotes the conditional distribution of Y given x. We make the dependence on x explicit by modelling $Q(\tau|x)$ in a regression framework of the type

$$Q(\tau|x) = \beta_{0,\tau} + x\beta_{x,\tau}. \tag{7.3.2}$$

This gives a linear quantile regression model, which clearly can be generalised in the same way as linear regression models, i.e. the covariate x can be multivariate or discrete.

Assume now that we have observed the data $(y_i, x_i), i = 1, \ldots, n$ with x_i as the covariates. In order to demonstrate the estimation of the parameters in (7.3.2), let us first look at median regression, i.e. $\tau = 0.5$. Remember that the least squares estimates in linear models were derived by minimising the sum of squared residuals

$\sum_{i=1}^{n}(y_i - x_i\beta)^2$ for the given data $(y_i; x_i)$, $i = 1, \dots, n$. If we replace the squared distance with the absolute distance, we get the framework for median regression. Setting $Q(0.5|x) = \beta_{0,0.5} + x\beta_{x,0.5}$, the parameters are found with

$$(\hat{\beta}_{0,0.5}, \hat{\beta}_{x,0.5}) = \arg\min \sum_{i=1}^{n} |y_i - \beta_{0,0.5} - x_i\beta_{x,0.5}|. \tag{7.3.3}$$

We will now demonstrate how these estimates can be computed with linear programming. To do so, we will ignore the regression framework for the moment and give a general definition of quantiles. For $\tau \in (0, 1)$, we define the **check function** $\delta_\tau(y)$ with

$$\delta_\tau(y) = \begin{cases} \tau y & \text{if } y \geq 0 \\ (\tau - 1)y & \text{if } y \leq 0, \end{cases}$$

which can also be written as

$$\delta_\tau(y) = y(\tau - \mathbb{1}_{\{y<0\}}).$$

The typical shape of a check function is visualised in Fig. 7.5 for different values of τ. The check function allows the definition of the quantile function $Q(\tau)$ with

$$\begin{aligned} Q(\tau) &= \arg\min_q E\Big\{\delta_\tau(Y - q)\Big\} \\ &= \arg\min_q \Big\{(\tau - 1)\int_{-\infty}^{q}(y - q)f(y)dy + \tau\int_{q}^{\infty}(y - q)f(y)dy\Big\}. \end{aligned} \tag{7.3.4}$$

If we now differentiate $E\Big\{\delta_\tau(Y - q)\Big\}$ with respect to q, we obtain

$$(1 - \tau)\int_{-\infty}^{q} f(y)dy - \tau\int_{q}^{\infty} f(y)dy = (1 - \tau)F(q) - \tau(1 - F(q)) = F(q) - \tau.$$

Clearly, this defines q as the τ-quantile.

The quantiles can be derived using the method of moments estimation. To do so, one replaces the expectation in (7.3.4) with its empirical counterpart. This gives

$$\hat{Q}(\tau) = \arg\min_{q_\tau}\left(\sum_{i=1}^{n}\delta_\tau(y_i - q_\tau)\right). \tag{7.3.5}$$

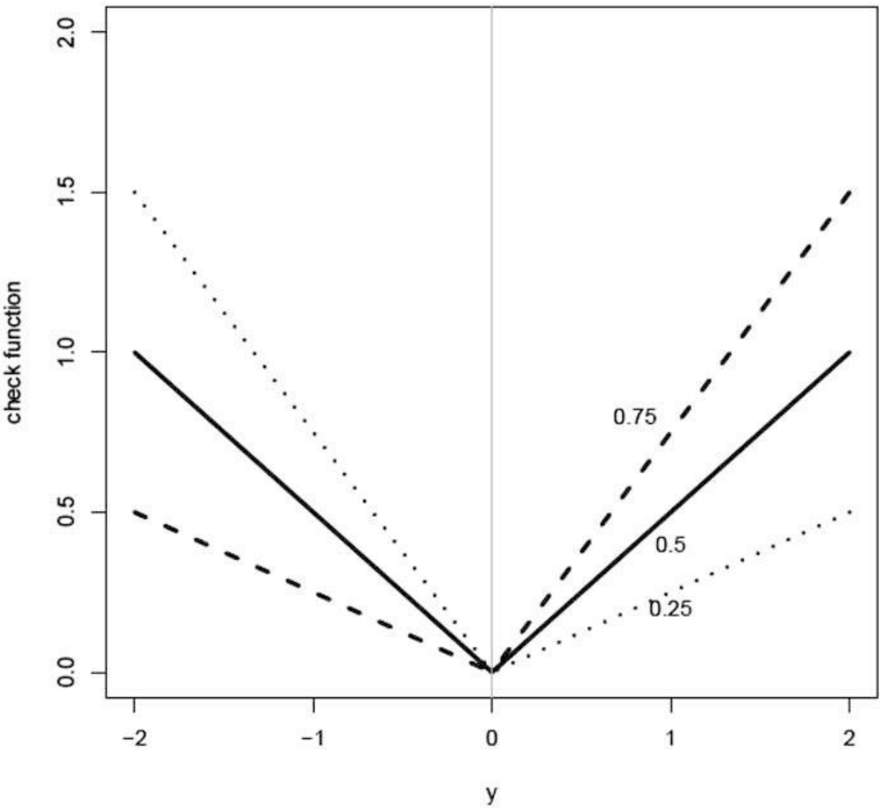

Fig. 7.5 Check function $\delta_\tau(y)$ for different values of τ

Replacing q_τ with a linear regression, for example,

$$q_\tau = \beta_{0,\tau} + x\beta_{x,\tau}$$

gives the general formula for quantile regression. To derive the estimates numerically, we can rewrite the minimisation problem as

$$\delta_\tau(y_i - q_\tau) = \tau u_i + (1-\tau)v_i,$$

where

$$u_i = \max\{y_i - q_\tau, 0\} \text{ and } v_i = \max\{-(y_i - q_\tau), 0\}$$

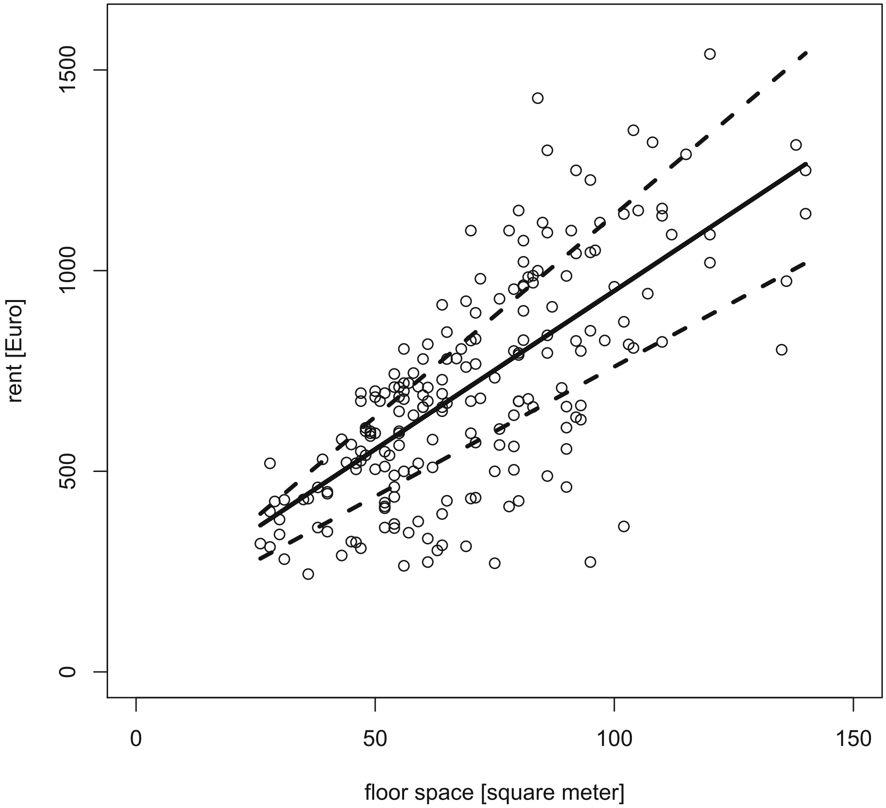

Fig. 7.6 Quantile regression functions for $\tau = 0.25$ (bottom dashed line), $\tau = 0.5$ (centre line) and $\tau = 0.75$ (upper dashed line)

such that $x_i - q_\tau = u_i - v_i$. Hence, minimising (7.3.5) is equivalent to minimising the linear function

$$\min \sum_{i=1}^{n} (\tau u_i + (1-\tau) v_i) \tag{7.3.6}$$

subject to the linear constraints $u_i \geq 0$ and $v_i \geq 0$ for $i = 1, \dots, n$. Clearly, (7.3.6) is a linear programming problem. For an overview, see e.g. Danzig and Thapa (1997)

Figure 7.6 shows the resulting quantile regression functions for the rental data with $\tau_1 = 0.25$, $\tau_2 = 0.5$ and $\tau_3 = 0.75$. The estimated parameters are given in Table 7.4. The plot shows the advantages of quantile regression, in that it allows the modelling and visualisation of variance heteroskedasticity. The range of the observed rents gets larger with increasing floor space. Hence, it seems that not only the median and mean but also the variance of the rents depends upon x.

Table 7.4 Parameter estimates for quantile regression

	Value	Std. error	t value	Pr(>|t|)
$\tau = 0.25$				
(Intercept)	114.49	65.36	1.75	0.08
fspc	6.46	0.88	7.32	< 0.001
$\tau = 0.5$				
(Intercept)	159.96	43.86	3.64	< 0.0001
fspc	7.90	0.59	13.34	< 0.001
$\tau = 0.75$				
(Intercept)	132.33	38.07	3.47	< 0.01
fspc	10.07	0.51	19.59	< 0.01

To draw proper statistical inference, we need to take the estimation variability into account. Hence, we consider the distributional properties of our estimates. It is shown in Koenker (1996) that

$$\hat{\beta}_\tau - \beta_\tau \sim N\left(0, n^{-1}H^{-1}(\tau)J(\tau)H^{-1}(\tau)\right),$$

where $J(\tau)$ and $H(\tau)$ can be approximated with

$$J(\tau) = \frac{\tau(1-\tau)}{n}\sum_{i=1}^{n}(1, x_i)^T(1, x_i)$$

$$H(\tau) = \frac{1}{n}\sum_{i=1}^{n}(1, x_i)^T(1, x_i)f(\beta_{0\tau} + x_i\beta_{1\tau})$$

with $f(.)$ as density of Y. The asymptotic normality is slightly more complicated to prove, because the check function $\delta_\tau(.)$ is not differentiable. We therefore do not go into more detail here.

7.4 Nonparametric Smooth Models

So far, we have only considered parametric models, which are simple in structure and are often merely defined by a linear relationship $\beta_0 + x\beta_x$. Let us now extend these models towards **nonparametric models**, sometimes also labelled as **semiparametric models**. The term nonparametric here refers to functional, nonparametric components in the model. Returning to the original linear regression model (7.1.1), we assume a metric covariate x that can take arbitrary values on the real axis. The idea is now to replace the linear relationship between covariate and response with a more flexible model of the type

$$Y = m(x) + \varepsilon,$$

where the influence of x on y is mediated by $m(.)$, a smooth, i.e. differentiable, but otherwise unspecified function. Like above we assume that ε is a vector of normally distributed random variables with mean zero and variance σ^2. Given the data (y_i, x_i), the intention is to estimate $m(x)$ in a smooth form, meaning a sufficiently differentiable function. The literature on smooth estimation of $m(x)$ is vast and we refer to Hastie and Tibshirani (1990), Fan and Gijbels (1996) or Ruppert et al. (2003) for a comprehensive discussion or see Fahrmeir et al. (2015) for a more recent work.

Various methods have been proposed for the estimation of $m(.)$, including kernel regression and spline smoothing. Here, we demonstrate the use of penalised splines, originally proposed in Eilers and Marx (1996). The idea is closely related to regression splines, where one replaces the unknown function $m(x)$ with a linear combination of known basis functions. To begin, we start with

$$m(x) = \boldsymbol{B}(x)\boldsymbol{\theta} = \sum_{k=1}^{K} B_k(x)\theta_k.$$

There are many possibilities for choosing the basis function. A convenient and practical choice is to work with B-splines as proposed in de Boor (1972). A B-spline basis is constructed by first locating knots on the x-axis. Between the knots, the B-spline is a piecewise polynomial function. A linear B-spline is thereby linear between the knots and a quadratic B-spline is built from polynomials of order 2. In Fig. 7.7, we show a linear B-spline basis (top row) and quadratic B-splines (bottom row). Linear B-splines lead to a continuous piecewise linear function $m(x)$ and a piecewise differentiable quadratic function is given by the quadratic B-spline. In the left-hand plots, we show the B-spline for equidistant knots. Often, however, the covariate measurements x_i are not uniformly distributed and it is advisable to instead allocate the knots based on the quantiles of x_i. This is sensible, because it gives the spline basis more structure where the x_i are dense and there is more information content, while giving less structure in areas with less information. Given the rent data, the resulting bases for non-equidistant knots are shown on the right of Fig. 7.7. The B-spline basis now replaces the original linear relationship and therefore provides more structure and flexibility. Note that we can now in principle follow the arguments from Sect. 7.1 and write the design matrix $\boldsymbol{X}$ as

$$\boldsymbol{X} = \begin{pmatrix} B_1(x_1) & \dots & B_K(x_1) \\ \vdots & & \vdots \\ B_1(x_n) & \dots & B_K(x_n) \end{pmatrix}.$$

The intercept column is incorporated into the B-splines, and hence an extra column of ones is not necessary here. This becomes more clear in Fig. 7.7. The overall level, which is usually included in the intercept, can be expressed by the overall level of the B-splines.

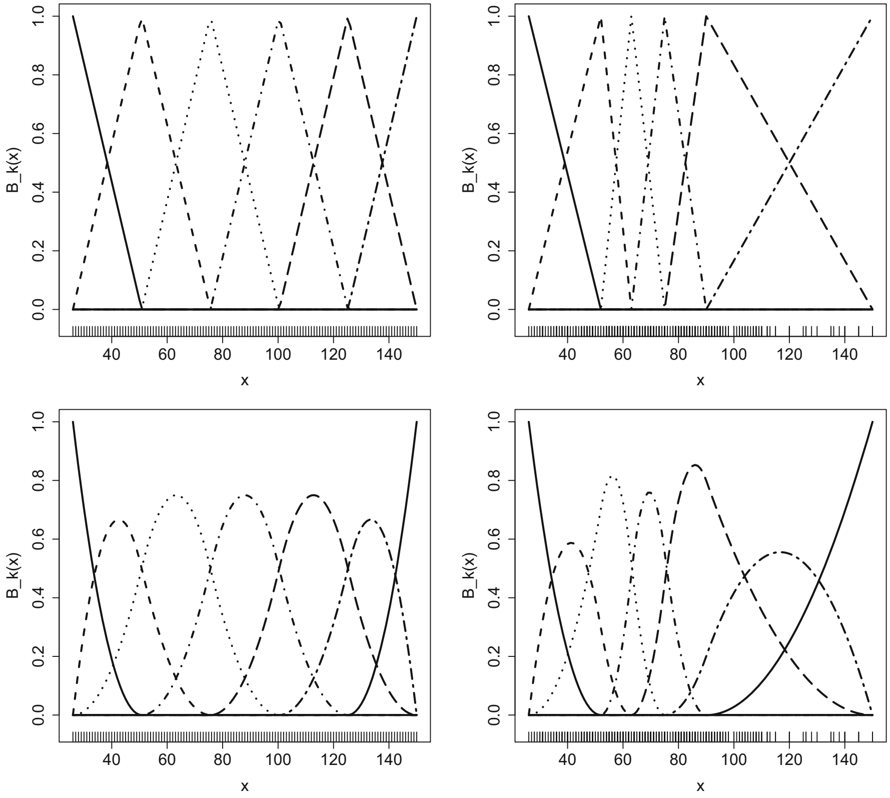

Fig. 7.7 B-spline basis for equidistant (left-hand side) and non-equidistant knots (right-hand side). The top row shows linear B-splines, and the bottom row shows quadratic B-splines

As previously, the estimate of the regression parameter θ is given by

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

We have plotted the resulting fit for different values of the spline dimension K in Fig. 7.8. As before, this is based on the Munich rent data, but this time with Y as the rent divided by the number of square metres of the apartment. The top row shows the fit with linear B-splines and the bottom with quadratic B-splines. The corresponding basis $\boldsymbol{B}(x) = (B_1(x), \ldots, B_K(x))$ is shown in the bottom of the plot, where the basis functions $B_k(x)$ are weighted by their corresponding estimated coefficient $\hat{\theta}_k$. The resulting fit is included in the plot as a line. For $K = 5$, shown in the left column, we obtain a reasonably smooth fit, but for $K = 15$, shown in the right-hand column, the fit gets too "wiggly" and appears unrealistic. This suggests that the basis for $K = 15$ is too complex and the estimation variability too large. To overcome this problem, we impose a penalty on the spline coefficients $\boldsymbol{\theta}$. Note

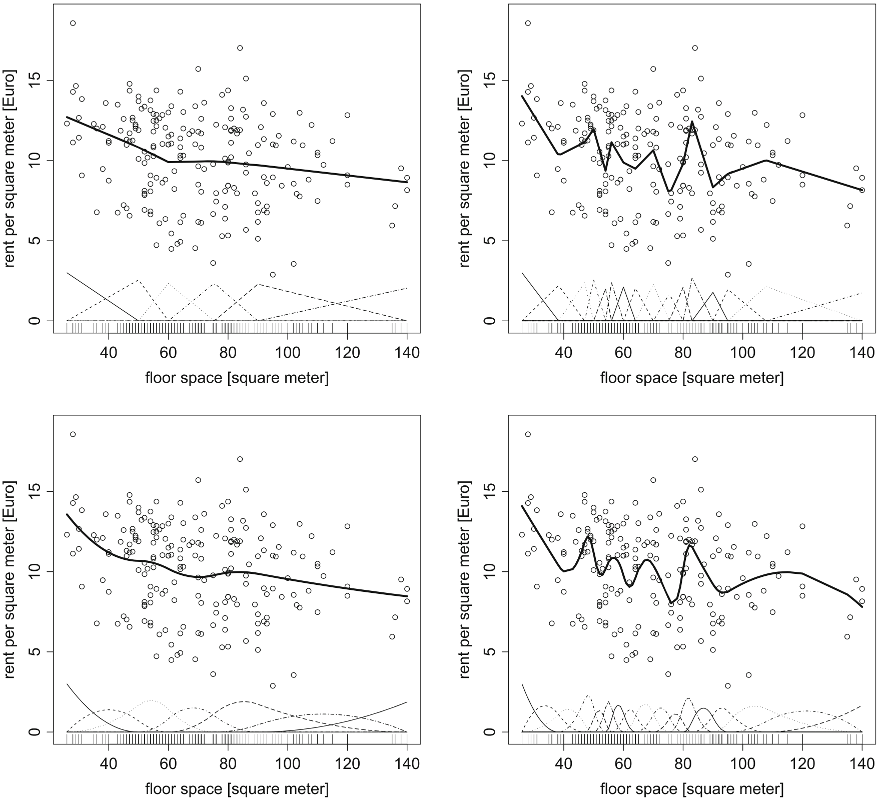

Fig. 7.8 Fitted functional shape for different spline dimensions. The top row shows linear B-splines, and the bottom row shows quadratic B-splines. $K = 5$ is given in left-hand plots and $K = 15$ in right-hand plots

that the wiggliness occurs when two neighbouring spline coefficients have highly differing values, which seems implausible given our assumption of smoothness. We therefore favour solutions where neighbouring coefficients have similar values, that is $|\theta_k - \theta_{k-1}|$ should be small. We enforce this by imposing a penalty on neighbouring spline coefficients, i.e. $\sum_{j=2}^{K}(\theta_j - \theta_{j-1})^2$. An alternative is to use second order differences, i.e.

$$\sum_{j=2}^{K}\big((\theta_j - \theta_{j-1}) - (\theta_{j-1} - \theta_{j-2})\big)^2 = \sum_{j=2}^{K}(\theta_j - 2\theta_{j-1} + \theta_{j-2})^2 \to \text{ small.}$$

To formulate this in matrix notation, we define the difference matrix

$$\boldsymbol{L} = \begin{pmatrix} 1 & -2 & 1 & 0 & \dots & 0 \\ 0 & 1 & -2 & 1 & & \\ \vdots & \ddots & \ddots & \ddots & & \vdots \\ 0 & & \dots & 1 & -2 & 1 \end{pmatrix}$$

and replace the least squares with a penalised version. To do so, we add a penalty to the log-likelihood (7.1.4) and write

$$\begin{aligned} l_p(\boldsymbol{\theta}, \sigma^2, \lambda) &= -\frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{X\theta})^T(\boldsymbol{y} - \boldsymbol{X\theta}) - \frac{1}{2}\frac{\lambda}{\sigma^2}\boldsymbol{\theta}^T\boldsymbol{L}\boldsymbol{L}^T\boldsymbol{\theta} \qquad (7.4.1) \\ &= l(\boldsymbol{\theta}, \sigma^2) - \frac{1}{2}\frac{\lambda}{\sigma^2}P(\boldsymbol{\theta}, \lambda), \end{aligned}$$

where $P(\boldsymbol{\theta}, \lambda)$ defines the penalty term. The scalar term λ defines the desired level of smoothness and will be discussed later. Simple differentiation gives the estimate as

$$\hat{\boldsymbol{\theta}} = (\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{L}\boldsymbol{L}^T)^{-1}(\boldsymbol{X}^T\boldsymbol{y}). \qquad (7.4.2)$$

The effect of penalisation can be seen in Fig. 7.9, which has the same format as Fig. 7.8. The left-hand column shows the unpenalised and the right the penalised fit. The penalisation makes the fit smooth, even if the basis is complex. In this respect, penalisation compensates for a basis that is too complex and forces the resulting fit to be smooth.

The **smoothing parameter** λ plays a central role in the above estimation process. If we set $\lambda = 0$, we obtain an unpenalised, wiggly estimate. On the other hand, if we set $\lambda \to \infty$, given our current difference matrix L, the coefficients must satisfy $\theta_j - 2\theta_{j-1} + \theta_{j-2} \equiv 0$ for $3 \leq j \leq K$. In practice, λ needs to be chosen based on the data and there are well-established routines that allow its numerical calculation. These are, for instance, cross validation or the AIC and BIC criteria, which we will discuss more generally in Chaps. 8 and 9 and therefore only briefly sketch out here. The key idea behind the **Akaike Information Criterion (AIC)** (see Akaike, 1973) is to balance the complexity of a model and the goodness of fit. The more complex a model, the better the fit, but if we allow the model to become too complex, then parsimony is violated. Thus, we want a model that is as simple as possible, but no simpler. We measure the fit using the log squared error

$$\text{fit}(\lambda) := n\log\left\{\sum_{i=1}^{n}(y_i - \hat{m}(x_i))^2\right\},$$

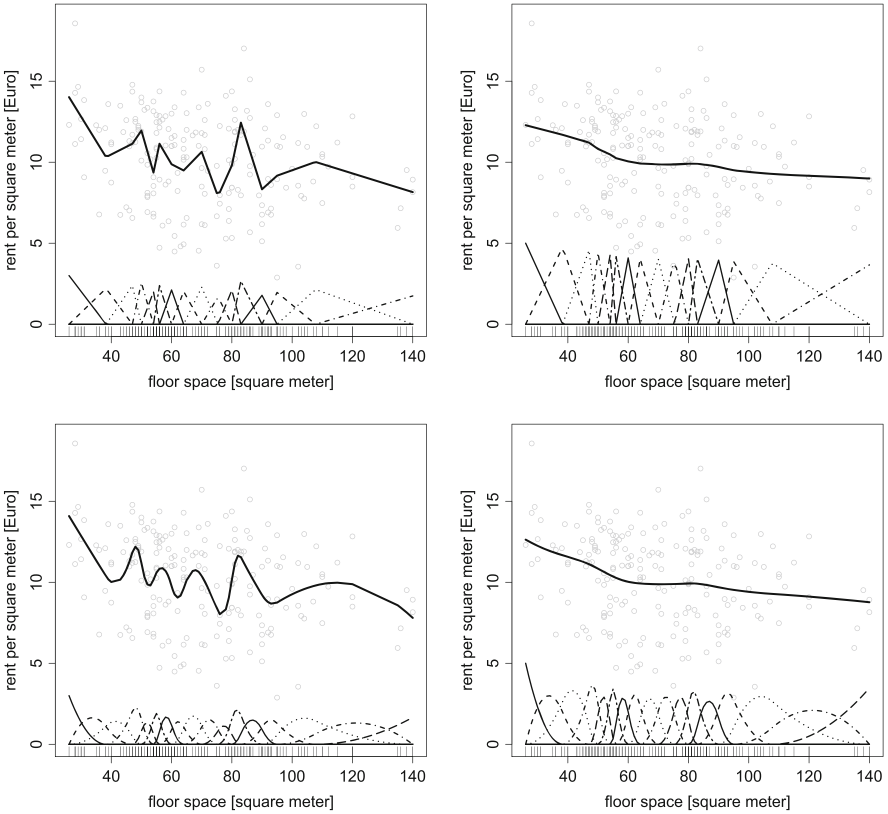

Fig. 7.9 Unpenalised (left) and penalised fit (right) for rental data. Top row is for linear B-splines and the bottom row for quadratic B-splines

where $\hat{m}(x_i) = \boldsymbol{B}(x_i)\hat{\boldsymbol{\theta}}$ with $\hat{\boldsymbol{\theta}}$ as in (7.4.2). Note that the fit $\hat{m}(x) = \boldsymbol{B}(x)\hat{\boldsymbol{\theta}}$ still depends on the smoothing parameter λ, because our estimated $\hat{\boldsymbol{\theta}}$ in (7.4.2) is calculated for a given λ. This is suppressed in the notation for simplicity but should be kept in mind. Clearly, the more complex the model, i.e. the smaller the smoothing parameter λ, the better the fit and the closer $\hat{m}(x_i)$ is to y_i. However, the resulting function is wiggly and complex. As a measure of the complexity of a model, we can define its dimension with

$$\dim(\lambda) = tr\left\{\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{L}\boldsymbol{L}^T)^{-1}\boldsymbol{X}\right\} = tr\left\{(\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{L}\boldsymbol{L}^T)^{-1}(\boldsymbol{X}^T\boldsymbol{X})\right\}. \tag{7.4.3}$$

Note that if $\lambda = 0$, then $\dim(\lambda) = K$, the number of splines in the basis, and the matrix in the first component of (7.4.3) becomes the hat matrix. If $\lambda \to \infty$, the dimension decreases, and hence for general λ, the matrix in the first component of

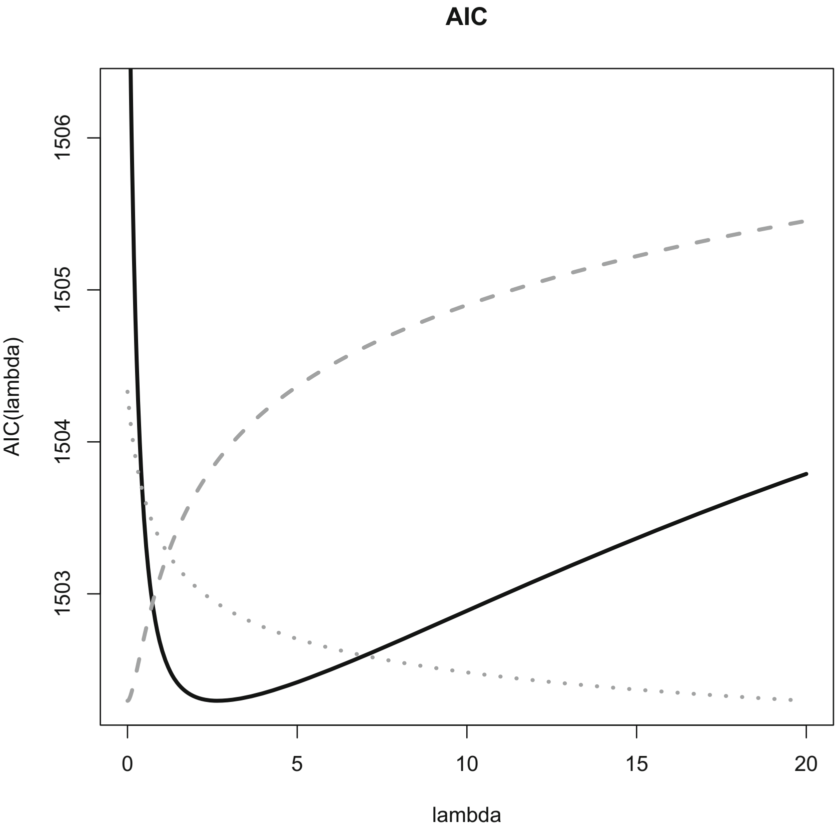

Fig. 7.10 AIC(λ) (solid line) and corresponding functions fit(λ) (dashed) and dim(λ) (dotted)

(7.4.3) can be seen as a generalisation of the hat matrix, and in fact for $\lambda = 0$ it is the hat matrix. These two components are now combined in the Akaike Information Criterion and give

$$\text{AIC}(\lambda) = \text{fit}(\lambda) + 2\,\text{dim}(\lambda).$$

A deeper motivation for this formula is given in Chap. 8. For now let us look at the criterion in the rent example from before, which is shown in Fig. 7.10. We show the curve AIC(λ) for different values of λ, which has a minimum at approximately $\lambda = 4$. We also include the curves for the fit(λ) (dashed grey line) and dim(λ) (dotted grey line).

Minimising the AIC(λ) to obtain an optimal λ requires its calculation for a dense grid of possible values. This can be a computational burden or even infeasible if multiple functions are fitted and thus several smoothing parameters need to be optimised simultaneously. In recent years, a numerically more feasible routine to select λ has been developed, which relies on a Bayesian approach. To understand this idea, we must first look at the penalty term in (7.4.1), which has a quadratic form in the log-likelihood. Taking the exponential of the penalty gives

$$\exp(-\frac{1}{2\sigma_\theta^2}\theta^T D\theta),$$

where $D = LL^T$ and $\sigma_\theta^2 = 1/\lambda$. This term mirrors the structure of a multivariate normal distribution. In fact, taking a Bayesian viewpoint, we can impose a prior on the parameter vector $\boldsymbol{\theta}$ in the form

$$\boldsymbol{\theta} \sim N(0, \sigma_\theta^2 \boldsymbol{D}^-), \tag{7.4.4}$$

where $\boldsymbol{D}^-$ is the generalised inverse of $\boldsymbol{LL}^T$, that is $(D^-)^- = \boldsymbol{LL}^T$. Because $\boldsymbol{LL}^T$ is not invertable, the prior (7.4.4) is degenerate and not proper. We have seen, however, that improper priors may still lead to proper posterior distributions. Note that the log posterior is now, up to an additive constant, equal to

$$\log f(\boldsymbol{\theta}, \sigma^2; \sigma_\theta^2 | y) = const + l(\boldsymbol{\theta}, \sigma^2) - \frac{\tilde{K}}{2}\log(\sigma_\theta^2) - \frac{1}{2\sigma_\theta^2}\boldsymbol{\theta}^T \boldsymbol{D}^- \boldsymbol{\theta},$$

where $\tilde{K}$ is the rank of $\boldsymbol{D}^-$. We can see now that the penalty parameter λ is equal to the inverse of the prior variance, i.e. $\lambda = \sigma_\theta^{-2}$. This is, in fact, a (hyper)parameter in the Bayesian model, which allows the derivation of its posterior distribution. Instead of using the posterior, it is more common to simply set σ_θ^2 to the posterior mode estimate or, equivalently, estimate the hyperparameters using empirical Bayes. In fact, one can comprehend the whole model as a linear mixed model (see e.g. Fahrmeir et al. (2015)) and estimate σ_θ^2 with Maximum Likelihood estimation to derive a penalised fit for $\boldsymbol{\theta}$. The resulting fit is shown in Fig. 7.9 in the right-hand column. Clearly, the above results have only been sketched and need to be explained in much more depth to be fully understandable. We consider this to be beyond the scope of this chapter but refer to Ruppert et al. (2003) or Wood (2017) for a more comprehensive introduction to nonparametric regression.

7.5 Generalised Linear Models

Thus far, we have assumed that Y is metric with normal residuals. Let us now relax this assumption to include any exponential family distribution for Y given the covariates x, see Sect. 2.1.5. Moreover, we assume that the parameter θ in the exponential family distribution depends on some covariates x. To be specific, let

$$Y|x \sim \exp\{t(y)\theta(x) - \kappa(\theta(x))\} h(y),$$

where, for simplicity, we limit ourselves to a univariate model. Parameter θ is assumed to depend on covariates x in the following way: first, we define what is called the linear predictor η, which, in the univariate case, is given by $\eta = \beta_0 + x\beta_x$. This is the linear part of the model. Second, let

$$\mu = \frac{\partial \kappa(\theta)}{\partial \theta} = E\left(t(Y); \theta\right)$$

be the expectation of the statistic $t(Y)$ of the distribution. Note that in the regression framework, parameter θ depends on x, which is however suppressed in the notation. We now link μ with the linear predictor η through a link function $g(.)$, such that $\mu = g^{-1}(\eta)$. This results in the relationship

$$g(E(Y_i|x_i)) = \beta_0 + x_i\beta_x.$$

The link function $g(.)$ needs to be suitably chosen, which becomes more clear in concrete examples given below. In fact, given that μ and θ have a one-to-one relation, a mathematically convenient setting is to link η and μ by taking

$$\theta = \eta. \tag{7.5.1}$$

This is also called the **canonical link function** and makes for simpler derivation of the log-likelihood. Let $(y_i, x_i), i = 1, \ldots, n$, be the observed data points and assume a canonical link (7.5.1). The log-likelihood is then given by

$$l(\beta) = \sum_{i=1}^{n} \{t(y_i)\eta_i - \kappa(\eta_i)\} + \log(h(y)),$$

where $\eta_i = \beta_0 + x_i\beta_x$. Note that $\log(h(y))$ can be dropped, as it has no influence on the maximum of $l(\beta)$. Taking derivatives gives the score function

$$\frac{\partial l(\beta)}{\partial \beta} = \sum_{i=1}^{n} \begin{pmatrix} 1 \\ x_i \end{pmatrix} \{t(y_i) - \mu_i\}$$

with $\mu_i = E(t(y_i)|x_i)$. We can reformulate this in matrix notation by setting

$$t(\boldsymbol{y}) = (t(y_y), \ldots, t(y_n))^T$$

and

$$t(\boldsymbol{Y}) = (t(Y_1), \ldots, t(Y_n))^T$$

and

$$\boldsymbol{X} = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}.$$

The score function can then be written as

$$\boldsymbol{X}^T\{t(\boldsymbol{y}) - E(t(\boldsymbol{Y}); \boldsymbol{\eta})\},$$

where $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_n)$. The Maximum Likelihood estimate is given by the fix-point equation

$$\boldsymbol{X}^T E(t(\boldsymbol{Y}); \hat{\boldsymbol{\eta}}) = \boldsymbol{X}^T t(\boldsymbol{y}).$$

Note that the Fisher matrix is given by

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

where $\boldsymbol{W}$ is a diagonal matrix with

$$\frac{\partial^2 \kappa(\eta_i)}{\partial\eta\partial\eta} = \mathrm{Var}(t(Y_i), \eta_i), i = 1, \ldots, n$$

on the diagonal. The above introduction to generalised linear models is certainly not directly illuminative. It shows, however, the principles behind the construction of the model class. Looking at a few examples should help to gather some intuition.

Example 27 The linear model discussed in Sect. 7.1 is a special case of a generalised linear model. We take the results from Sect. 2.1.5 and write the normal distribution as exponential family model, where we consider the variance σ^2 as given. For simplicity of notation, we set $\sigma^2 = 1$. We can then write the normal distribution model as

$$f(y; \mu) = \exp\left\{y\mu - \frac{\mu^2}{2}\right\} h(y^2)h(y^2; \sigma^2).$$

We get $E(t(y)) = \mu = \partial\kappa(\theta)/\partial\theta$. Setting $\theta = \beta_0 + x\beta_x$ leads to the linear model as previously discussed.

▷

Example 28 We now move on to logistic regression. Assume that $Y_i \in \{0, 1\}$ and consider the model $P(Y_i = 1|x_i)$. Using the results from Sect. 2.1.5, we obtain as canonical link the logit function and get the following regression model:

$$\mathrm{logit}\, \mathrm{P}(Y_i = 1|x_i) = \log\left(\frac{P(Y_i = 1|x_i)}{1 - P(Y_i = 1|x_i)}\right) = \beta_0 + x_i\beta_x.$$

Note that $\mathrm{Var}(Y_i|x_i) = P(Y_i = 1|x_i)(1 - P(Y_i = 1|x_i))$, which defines the weights in the weight matrix $\boldsymbol{W}$ from above.

▷

Example 29 If Y_i are count data, it is suitable to apply a Poisson model. In this case, the log is the canonical link, such that we model

$$\log E(Y_i|x_i) = \beta_0 + x_i\beta_x.$$

Note that $\mathrm{Var}(Y_i|x_i) = E(Y_i|x_i) = \exp(\beta_0 + x\beta_x)$, which defines the weights in the weight matrix $\boldsymbol{W}$.

▷

Generalised linear models are a powerful tool and have seen multiple extensions since their introduction by Wedderburn and Nelder (1972). The canonical reference is Nelder and McCullagh (1989), see also Myers et al. (2010).

7.6 Case Study in Generalised Additive Models

As generalised linear models are a very central tool in statistical modelling, let us demonstrate their power and flexibility with a case study from the e-commerce sector. An internet retailer ran a marketing campaign for 30 days, with TV advertisements broadcast on various channels. The number of visits to their website was tracked for this period. The data are visualised in Fig. 7.11, where we can see striking peaks and clearly some diurnal variation, both of which need to be addressed in our statistical model. The daily pattern is also visualised in Fig. 7.12, where we plot the number of clicks (i.e. visits) as boxplots for each individual hour of the day. This plot shows that the data is highly skewed, with a large number of upper outliers. In fact, the boxes, which cover 50% of the data points, cover only a negligible portion of the data's range.

Before exploring and modelling the data, we zoom in and look arbitrarily at day 15, shown in Fig. 7.13. In this plot, we are better able to observe the nature of the peaks which will be explained shortly. First of all, however, we need to select a suitable probability model for the data. As we are recording counts, namely counts of clicks, the Poisson model appears suitable. In Sect. 2.1.4, we saw that the Poisson distribution has equal mean and variance. This feature is common for count data and thankfully our data also satisfy this assumption. To visualise this, we zoom once again into two separate hours, 5–6 am and 6–7 pm. These are indicated in grey in Fig. 7.13 and were chosen because they exhibit no extreme outliers. In Fig. 7.14, we plot the number of clicks in these two windows and report the (arithmetic) mean and the (empirical) variance of the data. Between 5 am and 6 am, we observe only low traffic with an average of 4.11 clicks per minute and a similar variance of 4.3. In the evening window, the traffic is ten times as large, with about 44.39 visits per

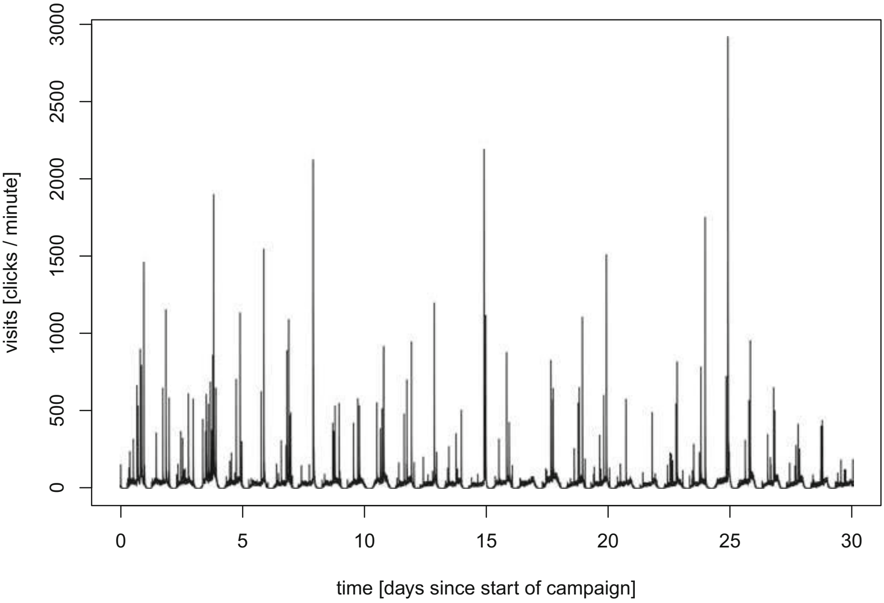

Fig. 7.11 Number of visits per minute over 30 days

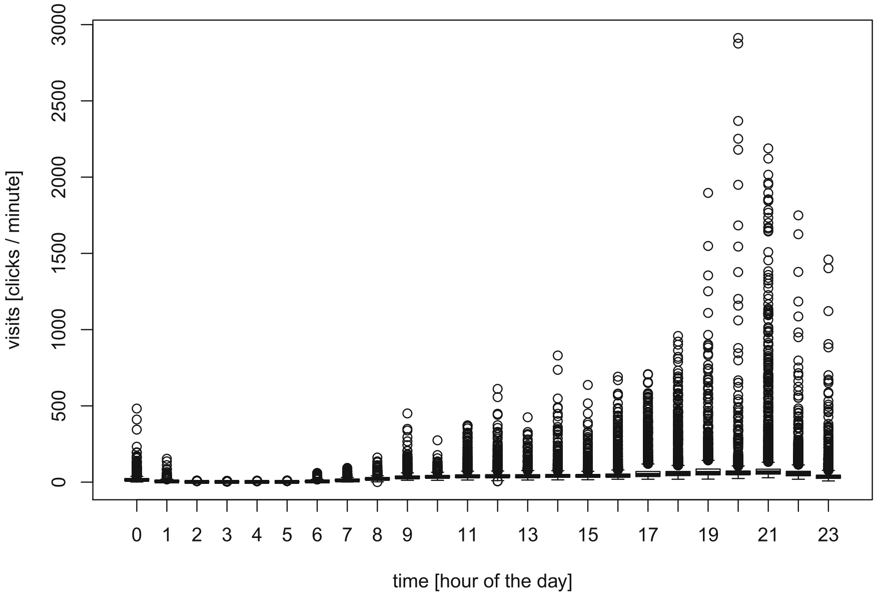

Fig. 7.12 Number of visits per minute plotted against hour of the day

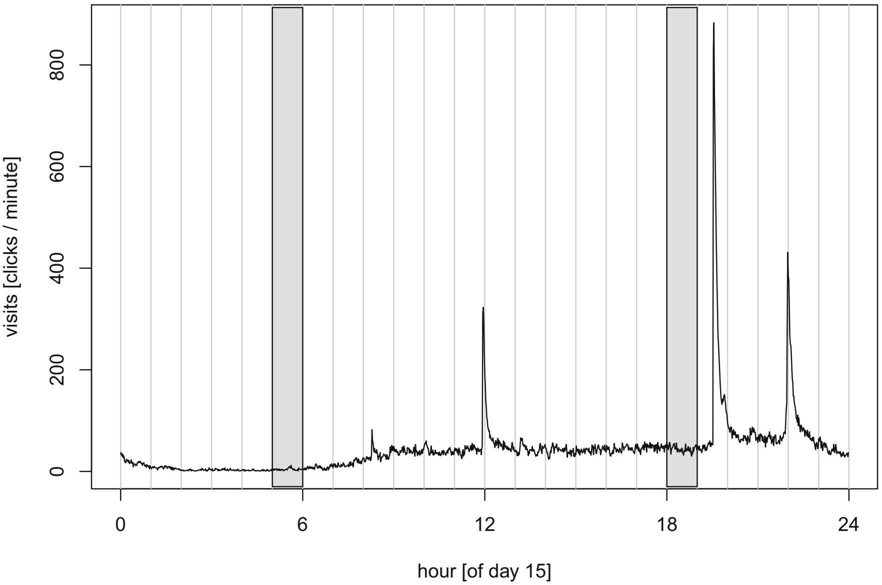

Fig. 7.13 Number of visits for day 15 with two hours being highlighted

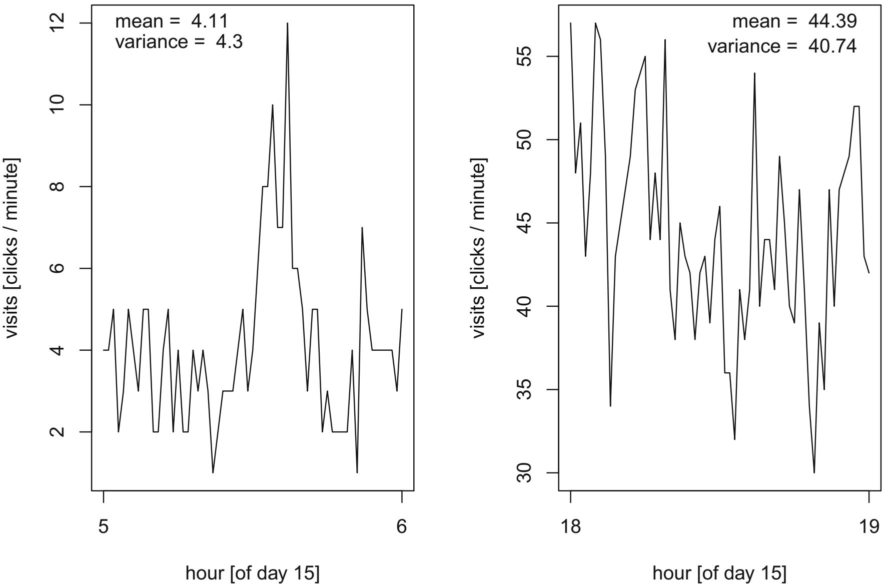

Fig. 7.14 Number of clicks for two hours of day 15 with mean and variance

minute and, again, a similar variance of 40.74. Thus, we can conclude that a Poisson distribution appears suitable for analysing the data. To start with, we define Y_i as the number of visits per minute and assume it is Poisson distributed

$$Y_i \sim Po(\lambda_i). \tag{7.6.1}$$

The intensity λ_i now needs to be modelled appropriately and, to that end, we define with t_i the timepoint at which Y_i was observed. t_i ranges from 0 to 30 in step size 1/(24*60), covering the time range of the campaign in minute intervals. With h_i we denote the hour of the day corresponding to t_i, where $h_i = (t_i \mod 60)$. We follow the framework of a generalised additive model (GAM) as discussed in Wood (2017), which combines smoothing models with generalised linear models, which we covered in Sects. 7.4 and 7.5, respectively. To be specific, we make use of the distribution in (7.6.1) and set

$$\lambda_i = \exp\{\beta_0 + t_i\beta_t + \beta_{\text{weekday}_i} + m(d_i)\}. \tag{7.6.2}$$

The exp guarantees that the intensity of the Poisson distribution is positive and β_t expresses the overall trend in the data, i.e. the long-term trend of the campaign. $m(d)$ carries the diurnal variation, and we postulate that it is cyclic, that is $\lim_{d \to 0}(m(d)) = \lim_{d \to 24}(m(d))$. The coefficient vector β_{weekday_i} captures differences between the different days of the week. The model can be easily fitted with the GAM procedure provided in the mgcv package in R. Again, the technically inclined reader can refer to Wood (2017) for more detail as the aim here is to demonstrate the flexibility of the model class without going into too much technical detail.

The fit of $m(D)$ is shown in Fig. 7.15, where the function is shown on a log scale and centred around zero, which can be interpreted as the average, and quantifies the deviation of the visits per minute throughout the day. We include horizontal lines, for instance, log(2) and $-\log(2)$, which delineate traffic at double or half the average rate, respectively. We can see that at around 3 am we observe only about 5% of the average traffic, while around 10 pm the traffic is 4 times greater than average.

The other parameter estimates are shown in Table 7.5. On Mondays, the reference category, we average about 31 (= exp(3.439)) visits per minute. We see a slight variation over the weekdays and the negative time coefficient shows that there is a small downward trend. We do not want to interpret these effects at this stage and instead focus on the spikes shown in Figs. 7.11 and 7.13. The spikes occur, as expected, at the same time as TV advertisements. To model this effect, we make use of a second dataset that specifies the exact minute that each advertisement was broadcast. This is visualised for day 15 in Fig. 7.16. The plot shows the same data as in Fig. 7.13, but now the vertical lines indicate the exact time of each advertisement. In particular, for the timepoints after midday, we see that the peaks occur exactly at the same time as the advertisements. Consumers view the advertisement on TV and immediately visit the advertised website. This explains the spikes and the

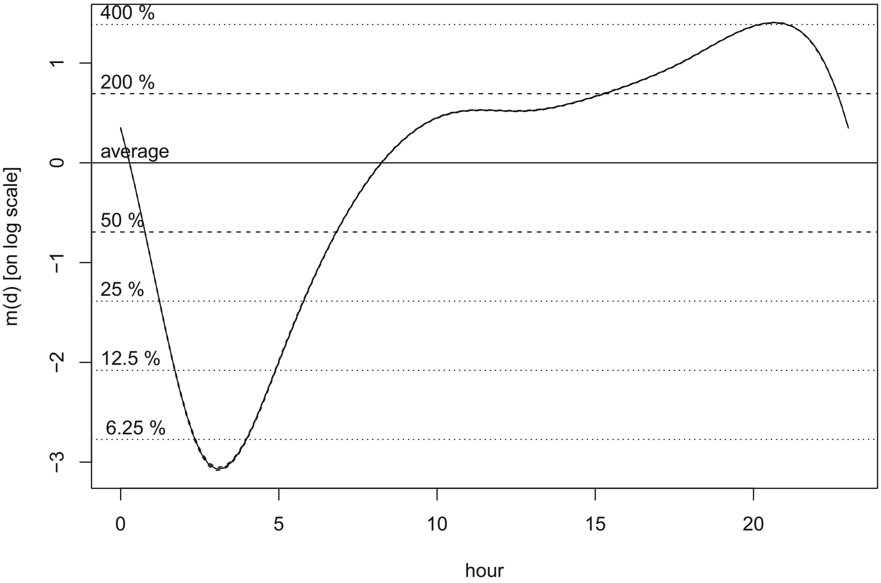

Fig. 7.15 Diurnal effect of the number of clicks

Table 7.5 Parameter estimates for quantile regression

	Value	Std. error	t value	Pr(>|t|)
(Intercept)	3.439e+00	2.650e−03	1297.68	<2e−16 ***
t	−1.057e−02	8.574e−05	−123.22	<2e−16 ***
wdTuesday	−7.028e−02	2.915e−03	−24.11	<2e−16 ***
wdWednesday	−1.489e−01	2.988e−03	−49.82	<2e−16 ***
wdThursday	6.398e−02	2.667e−03	23.99	<2e−16 ***
wdFriday	−1.154e−01	2.782e−03	−41.47	<2e−16 ***
wdSaturday	−1.012e−01	2.919e−03	−34.68	<2e−16 ***
wdSunday	3.852e−01	2.613e−03	147.45	<2e−16 ***

All coefficients are significantly different from zero, which is indicated with *** in the table

next modelling step is to include this phenomenon appropriately in the model. We propose a simple yet effective approach here. To motivate this we look at a single peak, namely the peak occurring at approximately half past 7 on the evening of the 15th day. The advertisement was sent out at 7:31 pm, leading to a sharp increase and subsequent exponential decay, which is clearly visible in Fig. 7.17. We model (or approximate) this behaviour by introducing a stimulus function.For the k-th advertisement broadcast, this is defined as

$$z_k(t) = \begin{cases} 0 & \text{for } t \leq t_{(k)} \\ \exp(-(t - t_{(k)})\delta) & \text{for } t > t_{(k)}, \end{cases}$$

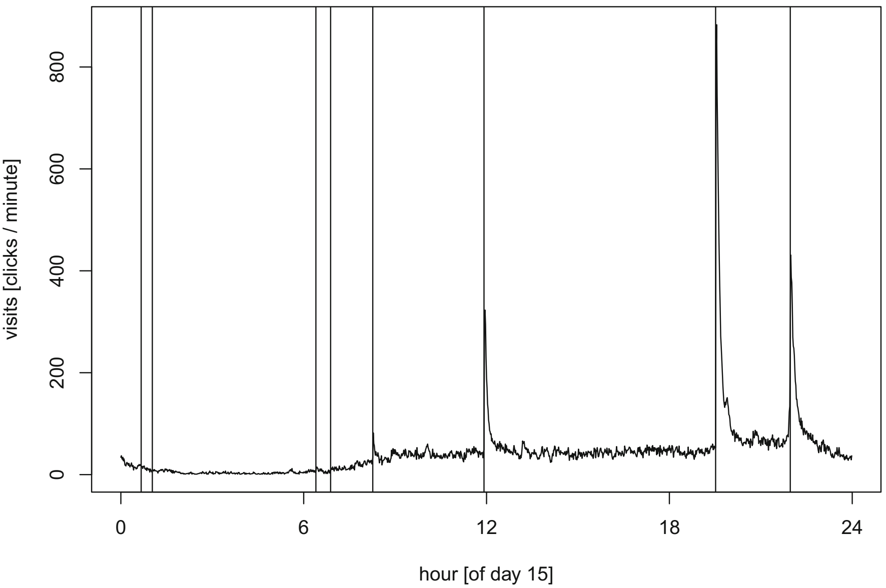

Fig. 7.16 Visits for day 15 with timepoints of advertisements indicated as vertical lines

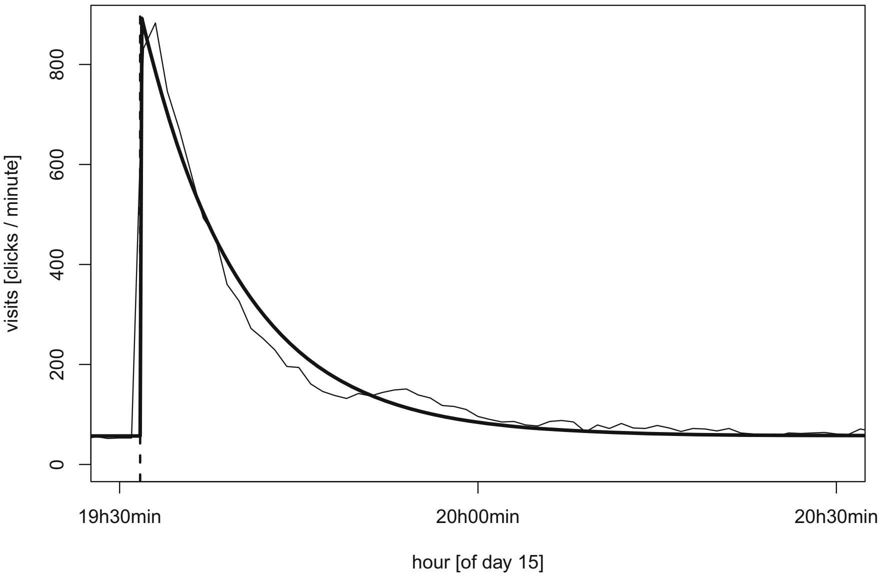

Fig. 7.17 Visualisation of function $z_k(t)$ (thick line) and the data (thin line) for one broadcasting timepoint

where $t_{(k)}$ is the timepoint of broadcast of the k-th advertisement. The decay parameter δ could in principle be estimated from the data, but, for simplicity, we use a calibrated value, which gives the function shown in Fig. 7.17. In actuality, $z_k(t)$ decays from a peak value of 1 but was scaled in the diagram to better demonstrate its fit. By defining for each of the $K = 254$ advertisements a corresponding stimulus function, we can extend Model (7.6.2) to

$$\lambda_i = \exp\{\beta_0 + t_i\beta_t + \beta_{\text{weekday}_i} + \sum_{k=1}^{K} z_k(t_i)\beta_k + m(d_i)\}. \tag{7.6.3}$$

We fit Model (7.6.3) and obtain parameter estimates and a fitted curve $\hat{m}(d)$. The curve looks comparable to the one shown in Fig. 7.15 and is therefore not shown again. The weekday effects change slightly, but our focus is generally more on the fitted effects $\hat{\beta}_k$. Before drawing conclusions from the model, it is however necessary to look at goodness of the model fit. We take a rudimentary approach here and compare the observed values Y_i with the fitted values $\hat{\lambda}_i$. If the model is appropriate, then the difference between Y_i and $\hat{\lambda}_i$ should be random. In Fig. 7.18, we plot the original data in the upper plot (this is a repetition of Fig. 7.11) and the fitted values $\hat{\lambda}_i$ in the bottom row. Without going into depth here, we see sufficient concordance. Now that we are more confident in our model, we can look at the fitted advertisement effects $\hat{\beta}_k$. The coefficients β_k can be interpreted as the efficacy of the advertisement. Given that the maximum value of $z_k(t)$ is 1, we can interpret β_k on a log scale., i.e. the value of $\beta_k \geq \log(2) = 0.693$ means that the traffic was temporarily more than doubled. This also means that $\beta_k \leq 0$ indicates timepoints where the advertisement was not at all effective. We fit Model (7.6.3) and plot the corresponding fitted advertisement effects in Fig. 7.19. We also include the confidence intervals calculated by adding and subtracting twice the standard deviation of the estimates. We can see that most fitted values of $\hat{\beta}_k$ are positive and, in fact, the largest fitted values are in the order of 4.5, indicating a 90-fold temporary increase to the number of website visits ($90 \approx \exp(4.5)$). The next step in our analysis is to investigate and explore the coefficients $\hat{\beta}_k$, to determine the timepoints when advertisements are most effective. To do so, we first look again at day 15 and plot both the data and the corresponding advertisement effects $\hat{\beta}_k$ for this day. There were 8 advertisements broadcast, which can be seen in Fig. 7.20. In the top plot, we again show the data for day 15, and in the bottom plot we show the fitted advertisement effects $\hat{\beta}_k$. We see that advertisements shown in the early morning had hardly any effect, and in fact one fitted effect is even negative, while there is a clear benefit to advertising in the evening. This is a general pattern, which can be seen from Fig. 7.21, where we plot the fitted advertisement effect $\hat{\beta}_k$ against the hour of the day. It becomes obvious that evening advertisements, broadcast between 8 pm and 11 pm, are the most effective.

While there was plenty of room for improvement in each of the individual steps of the analysis, we hope to have demonstrated the flexibility of generalised regression models with nonparametric and parametric effects.

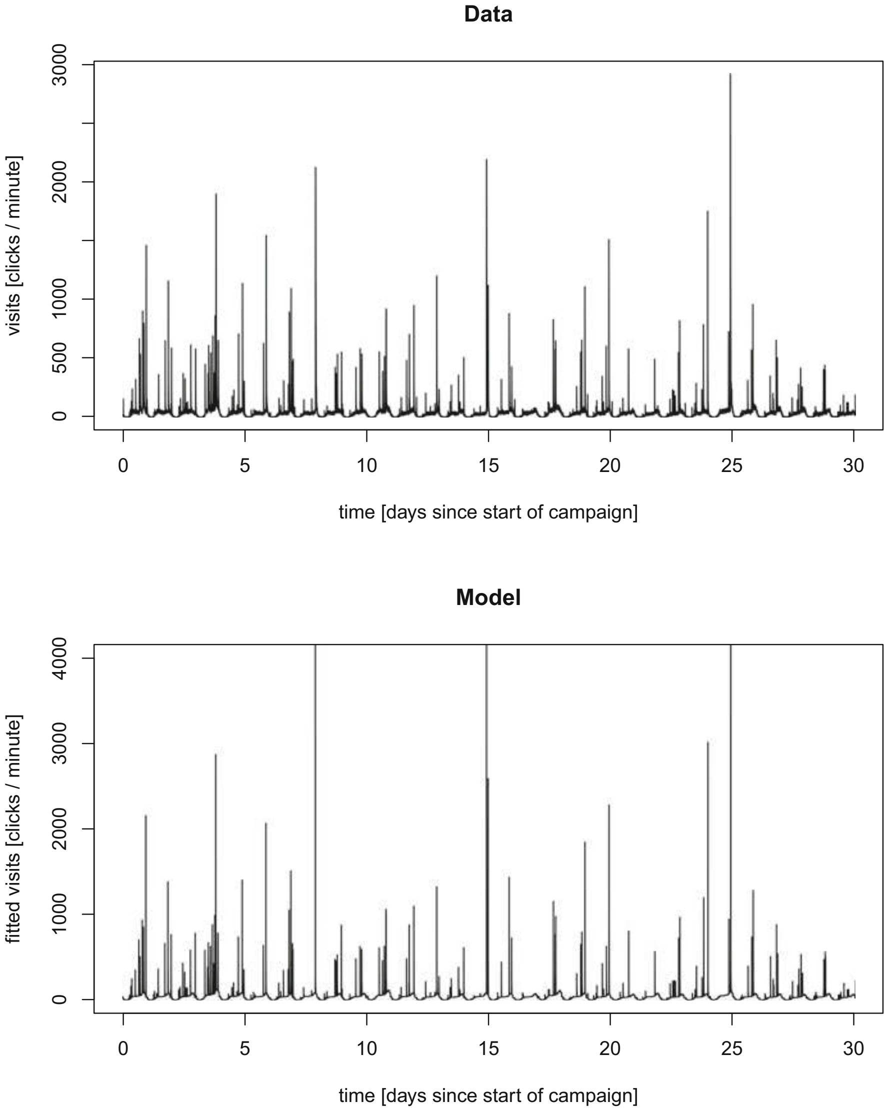

Fig. 7.18 Number of visits (top plot) and fitted model (bottom plot)

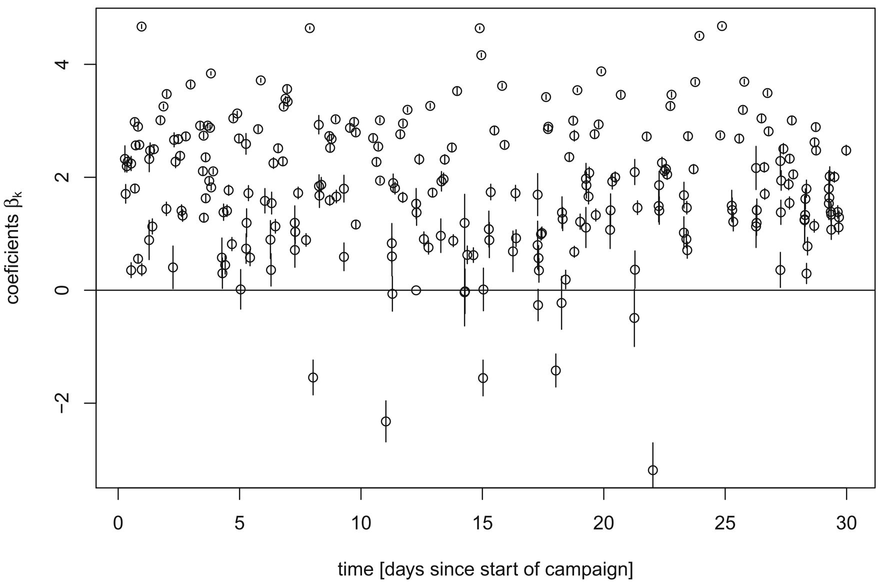

Fig. 7.19 Fitted coefficients $\hat{\beta}_k$ with confidence intervals plotted against the timepoint of broadcasting the advertisement

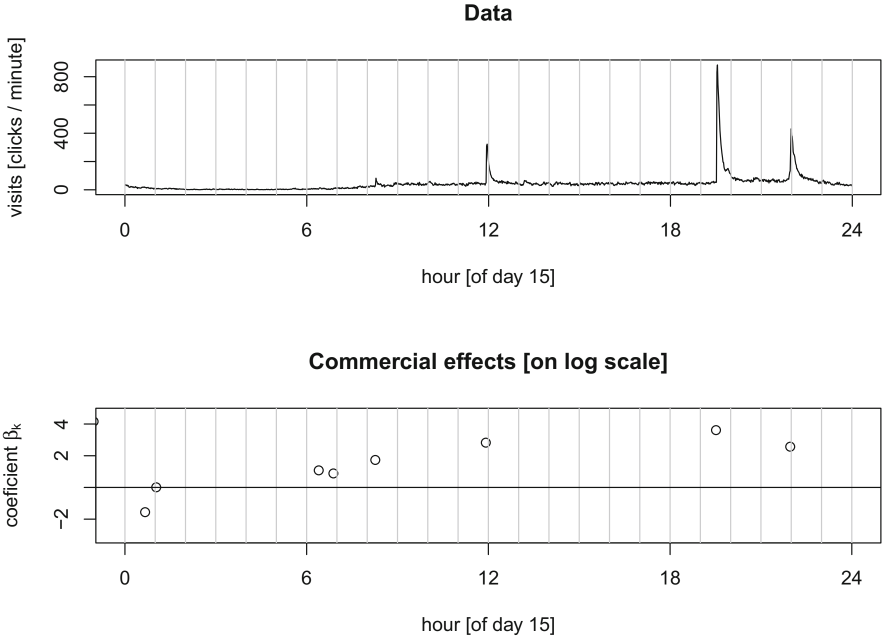

Fig. 7.20 Data for day 15 (top plot) and fitted advertisement effects $\hat{\beta}_k$ for that day

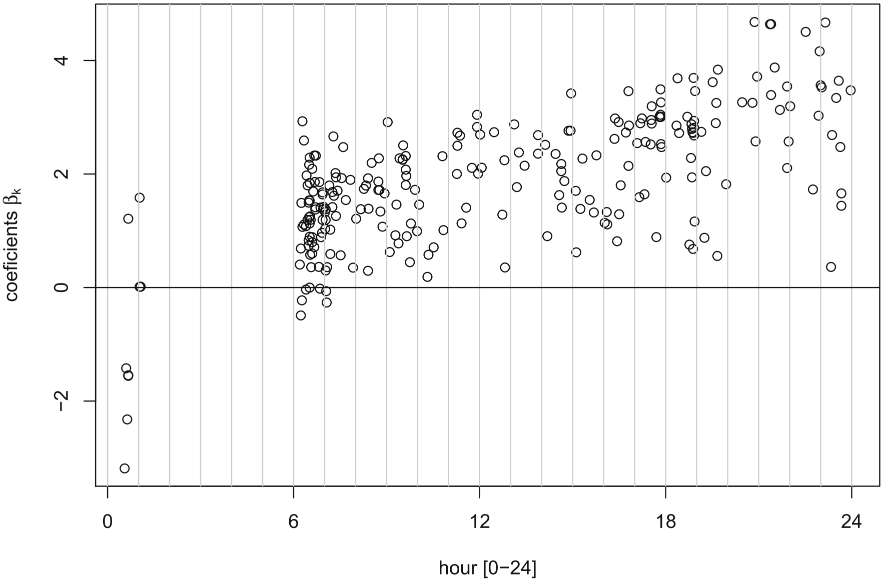

Fig. 7.21 Fitted advertisement effects $\hat{\beta}_k$ plotted against the hour of the day of the timepoint when the advertisement was broadcast

7.7 Exercises

Exercise 1 (Data Analysis and Linear Models in R)

The dataset `teengamb` in the **R** package `faraway` contains information on the gambling behaviour of teenagers in Great Britain:

Variable	Description
`gamble`	Money spent on gambling in pounds per year (response)
`status`	Socioeconomic status (based on parents' occupation)
`income`	Income (in pounds per week)
`verbal`	Test score based on the number of properly defined words (max. 12)
`sex`	Gender (0 = male and 1 = female)

1. Define and fit an ordinary linear regression model (including an intercept β_0) to explore the relationship between gambling and socioeconomic status and the income and verbal test score (`verbal`). Use the `lm` function in R.
2. Give the distribution of the estimated coefficient vector and compute the parameters if possible.
3. Perform an in-depth interpretation of the estimates of your coefficients.
4. Test if the model as a whole is useful, i.e. for the hypothesis $\beta_1 = \ldots = \beta_p = 0$. Interpret the result.

5. Make a linear model with all four covariates and also include the interactions between all variables and gender. Appropriately visualise the estimates of each effect. Finally, test whether we can reduce the set of covariates to `income` and `sex`.

Exercise 2 (Quantile Regression)
We revisit the dataset of the previous exercise. Instead of modelling the mean, we will now apply median regression.

1. Use the `quantreg` package in `R` to perform a quantile regression for the median, i.e. $\tau = 0.5$, to explore the relationship between gambling and socioeconomic status and the income and verbal test score. Try to interpret the results and compare them to the regression for the mean performed in Exercise 7.1 point 1.
2. What are possible advantages of quantile regression in comparison to ordinary linear regression?

Exercise 3 (Logistic Regression and Classification)
In credit scoring, banks want to check whether a client will pay back a credit in the specified time frame. We consider a binary outcome variable $Y_i \in \{0, 1\}$, which indicates whether client i pays back the credit ($Y_i = 0$) or not ($Y_i = 1$). In our example, further variables about the credit and the client are available: x_1 = duration of the credit, x_2 = amount of the credit, x_3 = previous payment behaviour (1 = good and 0 = bad), x_4 = intended use (1 = private and 0 = business) and x_5 = running account (0 = good running account and 1 = no or bad account running). The data can be downloaded from www.uni-goettingen.de/de/551625.html, see also the discussion in Fahrmeir et al. (2015).

1. Fit a logistic regression model $logit P(Y_i = 1|x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5$ resulting in estimations $\hat{\beta}_0, \ldots, \hat{\beta}_5$.
2. Using the fitted model, the probability of a failure of a credit can be estimated by

$$P(Y_i = 1|x) = G(\hat{\beta}_0 + \hat{\beta}_1 x_1 + \ldots + \hat{\beta}_5 x_5)$$

with $G(t) = logit^{-1}(t) = (1 + \exp(-t))^{-1}$. Use the result from 1 and give probabilities for different values of x: $x_1 = (0, 1)$, $x_2 = (0, 1)$, $x_3 = (0, 1)$, $x_4 = (12, 24, 36)$ and $x_5 = (4, 6, 8)$. The calculation can be used to decide whether the bank offers the credit or not by using a threshold for the failure probability.
3. Explain why logistic regression can be seen as a tool for binary classification based on information in the variables x.