# Linear Regression

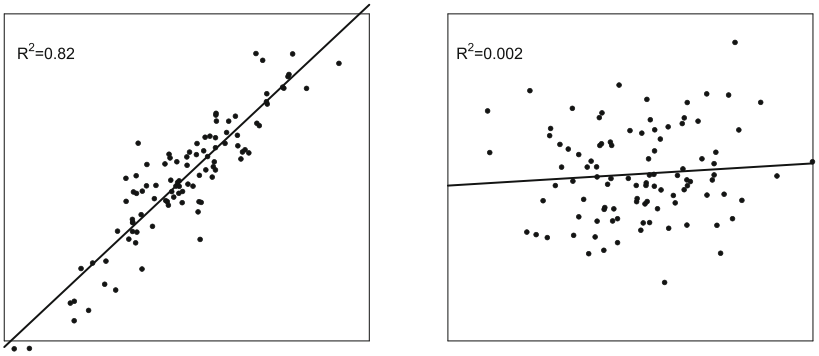
# 11

We learnt about various measures of association in Chap. 4. Such measures are used to understand the degree of *relationship* or *association* between two variables. The correlation coefficient of Bravais–Pearson, for example, can help us to quantify the degree of a linear relationship, but it does not tell us about the functional form of a relationship. Moreover, any association of interest may depend on more than one variable, and we would therefore like to explore *multivariate* relationships. For example, consider a situation where we measured the body weights and the long jump results (distances) of school children taking part in a competition. The correlation coefficient of Bravais–Pearson between the body weight and distance jumped can tell us how strong or weak the association between the two variables is. It can also tell us whether the distance jumped increases or decreases with the body weight in the sense that a negative correlation coefficient indicates shorter jumps for higher body weights, whereas a positive coefficient implies longer jumps for higher body weights. Suppose we want to know how far a child with known body weight, say 50 kg, will jump. Such questions cannot be answered from the value and sign of the correlation coefficient. Another question of interest may be to explore whether the relationship between the body weight and distance jumped is linear or not. One may also be interested in the joint effect of age and body weight on the distance jumped. Could it be that older children, who have on average a higher weight than younger children, perform better? What would be the association of weight and the long jump results of children of the same age? Such questions are addressed by linear regression analysis which we introduce in this chapter.

In many situations, the outcome does not depend on one variable but on several variables. For example, the recovery time of a patient after an operation depends on several factors such as weight, haemoglobin level, blood pressure, body temperature, diet control, rehabilitation, and others. Similarly, the weather depends on many factors such as temperature, pressure, humidity, speed of winds, and others. Multiple



C. Heumann et al., *Introduction to Statistics and Data Analysis*,
DOI 10.1007/978-3-319-46162-5_11




(a) Positive linear relationship (b) No clear relationship

**Fig. 11.1** Scatter plots

linear regression, introduced in Sect. 11.6, addresses the issue where the outcome depends on more than one variable.

To begin with, we consider only two quantitative variables $X$ and $Y$ in which the outcome $Y$ depends on $X$ and we explore the quantification of their relationship.

*Examples* Examples of associations in which we might be interested in are:

- body height ($X$) and body weight ($Y$) of persons,
- speed ($X$) and braking distance ($Y$) measured on cars,
- money invested (in €) in the marketing of a product ($X$) and sales figures for this product (in €) ($Y$) measured in various branches,
- amount of fertilizer used ($X$) and yield of rice ($Y$) measured on different acres, and
- temperature ($X$) and hotel occupation ($Y$) measured in cities.

## 11.1 The Linear Model

Consider the scatter plots from Fig. 4.2 on p. 80. Plotting $X$-values against $Y$-values enables us to visualize the relationship between two variables. Figure 11.1a reviews what a positive (linear) association between $X$ and $Y$ looks like: the higher the $X$ values, the higher the $Y$-values (and vice versa). The plot indicates that there may be a linear relationship between $X$ and $Y$. On the other hand, Fig. 11.1b displays a scatter plot which shows no clear relationship between $X$ and $Y$. The $R^2$ measure, shown in the two figures and explained in more detail later, equates to the squared correlation coefficient of Bravais–Pearson.

To summarize the observed association between $X$ and $Y$, we can postulate the following linear relationship between them:

$$Y = \alpha + \beta X. \tag{11.1}$$

This Eq. (11.1) represents a straight line where $\alpha$ is the **intercept** and $\beta$ represents the **slope** of the line. The slope indicates the change in the $Y$-value when the $X$-value changes by one unit. If the sign of $\beta$ is positive, it indicates that the value of $Y$ increases as the value of $X$ increases. If the sign of $\beta$ is negative, it indicates that the value of $Y$ decreases as the value of $X$ increases. When $X = 0$, then $Y = \alpha$. If $\alpha = 0$, then $Y = \beta X$ represents a line passing through the origin. Suppose the height and body weights in the example of school children are represented in Fig. 11.1a. This has the following interpretation: when the height of a child increases by 1 cm, then the body weight increases by $\beta$ kilograms. The slope $\beta$ in Fig. 11.1a would certainly be positive because we have a positive linear relationship between $X$ and $Y$. Note that in this example, the intercept term has no particular meaning because when the height $X = 0$, the body weight $Y = \alpha = 0$. The scatter diagram in Fig. 11.1b does not exhibit any clear relationship between $X$ and $Y$. Still, the slope of a possible line would likely be somewhere around 0.

It is obvious from Fig. 11.1a that by assuming a linear relationship between $X$ and $Y$, any straight line will not exactly match the data points in the sense that it cannot pass through all the observations: the observations will lie above and below the line. The line represents a *model* to describe the process generating the data.

In our context, a model is a mathematical representation of the relationship between two or more variables. A model has two components—variables (e.g. $X$, $Y$) and parameters (e.g. $\alpha$, $\beta$). A model is said to be *linear* if it is linear in its parameters. A model is said to be *nonlinear* if it is nonlinear in its parameters (e.g. $\beta^2$ instead of $\beta$). Now assume that each observation potentially deviates by $e_i$ from the line in Fig. 11.1a. The **linear model** in Eq. (11.1) can then be written as follows to take this into account:

$$Y = \alpha + \beta X + e. \tag{11.2}$$

Suppose we have $n$ observations $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, then each observation satisfies

$$y_i = \alpha + \beta x_i + e_i. \tag{11.3}$$

Each deviation $e_i$ is called an *error*. It represents the deviation of the data points $(x_i, y_i)$ from the **regression line**. The line in Fig. 11.1a is the fitted (regression) line which we will discuss in detail later. We assume that the errors $e_i$ are identically and independently distributed with mean 0 and constant variance $\sigma^2$, i.e. $\mathrm{E}(e_i) = 0$, $\mathrm{Var}(e_i) = \sigma^2$ for all $i = 1, 2, \ldots, n$. We will discuss these assumptions in more detail in Sect. 11.7.

In the model (11.2), $Y$ is called the **response**, **response variable**, **dependent variable** or **outcome**; $X$ is called the **covariate**, **regressor** or **independent variable**. The scalars $\alpha$ and $\beta$ are the parameters of the model and are described as **regression coefficients** or **parameters** of the linear model. In particular, $\alpha$ is called the **intercept term** and $\beta$ is called the **slope parameter**.

It may be noted that if the regression parameters $\alpha$ and $\beta$ are known, then the linear model is completely known. An important objective in identifying the model is to determine the values of the regression parameters using the available observations for $X$ and $Y$. It can be understood from Fig. 11.1a that an ideal situation would be when all the data points lie exactly on the line or, in other words, the error $e_i$ is zero for each observation. It seems meaningful to determine the values of the parameters in such a way that the errors are minimized.

There are several methods to estimate the values of the regression parameters. In the remainder of this chapter, we describe the methods of least squares and maximum likelihood estimation.

## 11.2 Method of Least Squares

Suppose $n$ sets of observations $P_i = (x_i, y_i)$, $i = 1, 2, \ldots, n$, are obtained on two variables $P = (X, Y)$ and are plotted in a scatter plot. An example of four observations, $(x_1, y_1)$, $(x_2, y_2)$, $(x_3, y_3)$, and $(x_4, y_4)$, is given in Fig. 11.2.

The **method of least squares** says that a line can be fitted to the given data set such that the errors are minimized. This implies that one can determine $\alpha$ and $\beta$ such that the sum of the squared distances between the data points and the line $Y = \alpha + \beta X$ is minimized. For example, in Fig. 11.2, the first data point $(x_1, y_1)$ does not lie on the plotted line and the deviation is $e_1 = y_1 - (\alpha + \beta x_1)$. Similarly, we obtain the

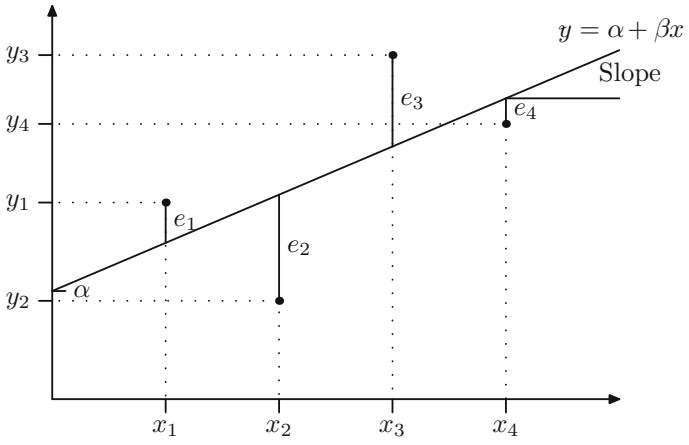



**Fig. 11.2** Regression line, observations, and errors $e_i$*

difference of the other three data points from the line: $e_2, e_3, e_4$. The error is zero if the point lies exactly on the line. The problem we would like to solve in this example is to minimize the sum of squares of $e_1, e_2, e_3$, and $e_4$, i.e.

$$\min_{\alpha,\beta} \sum_{i=1}^{4} (y_i - \alpha - \beta x_i)^2. \tag{11.4}$$

We want the line to fit the data well. This can generally be achieved by choosing $\alpha$ and $\beta$ such that the squared errors of all the $n$ observations are minimized:

$$\min_{\alpha,\beta} \sum_{i=1}^{n} e_i^2 = \min_{\alpha,\beta} \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2. \tag{11.5}$$

If we solve this optimization problem by the principle of maxima and minima, we obtain estimates of $\alpha$ and $\beta$ as

$$\left.\begin{aligned} \hat{\beta} &= \frac{S_{xy}}{S_{xx}} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{\sum_{i=1}^{n} x_i\, y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} \\ \hat{\alpha} &= \bar{y} - \hat{b}\bar{x} \end{aligned}\right\}, \tag{11.6}$$

see Appendix C.6 for a detailed derivation. Here, $\hat{\alpha}$ and $\hat{\beta}$ represent the estimates of the parameters $\alpha$ and $\beta$, respectively, and are called the **least squares estimator** of $\alpha$ and $\beta$, respectively. This gives us the model $y = \hat{\alpha} + \hat{\beta}x$ which is called the *fitted model* or the *fitted regression line*. The literal meaning of "regression" is to move back. Since we are acquiring the data and then moving back to find the parameters of the model using the data, it is called a *regression model*. The fitted regression line $y = \hat{\alpha} + \hat{\beta}x$ describes the postulated relationship between $Y$ and $X$. The sign of $\beta$ determines whether the relationship between $X$ and $Y$ is positive or negative. If the sign of $\beta$ is positive, it indicates that if $X$ increases, then $Y$ increases too. On the other hand, if the sign of $\beta$ is negative, it indicates that if $X$ increases, then $Y$ decreases. For any given value of $X$, say $x_i$, the *predicted* value $\hat{y}_i$ is calculated by

$$\hat{y}_i = \hat{\alpha} + \hat{\beta}x_i$$

and is called the $i$th *fitted value*.

If we compare the observed data point $(x_i, y_i)$ with the point suggested (predicted, fitted) by the regression line, $(x_i, \hat{y}_i)$, the difference between $y_i$ and $\hat{y}_i$ is called the **residual** and is given as

$$\hat{e}_i = y_i - \hat{y}_i = y_i - (\hat{\alpha} + \hat{\beta}x_i). \tag{11.7}$$

This can be viewed as an estimator of the error $e_i$. Note that it is not really an estimator in the true statistical sense because $e_i$ is random and so it cannot be estimated. However, since the $e_i$ are unknown and $\hat{e}_i$ measures the same difference between the estimated and true values of the $y$'s, see for example Fig. 11.2, it can be treated as estimating the error $e_i$.

*Example 11.2.1* A physiotherapist advises 12 of his patients, all of whom had the same knee surgery done, to regularly perform a set of exercises. He asks them to record how long they practise. He then summarizes the average time they practised ($X$, time in minutes) and how long it takes them to regain their full range of motion again ($Y$, time in days). The results are as follows:

| $i$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $x_i$ | 24 | 35 | 64 | 20 | 33 | 27 | 42 | 41 | 22 | 50 | 36 | 31 |
| $y_i$ | 90 | 65 | 30 | 60 | 60 | 80 | 45 | 45 | 80 | 35 | 50 | 45 |

To estimate the linear regression line $y = \hat{\alpha} + \hat{\beta}x$, we first calculate $\bar{x} = 35.41$ and $\bar{y} = 57.08$. To obtain $S_{xy}$ and $S_{xx}$ we need the following table:

| $i$ | $x_i$ | $y_i$ | $(x_i - \bar{x})$ | $(y_i - \bar{y})$ | $(x_i - \bar{x})(y_i - \bar{y})$ | $(x_i - \bar{x})^2$ |
|---|---|---|---|---|---|---|
| 1 | 24 | 90 | −11.41 | 32.92 | −375.61 | 130.19 |
| 2 | 35 | 65 | −0.41 | 7.92 | −3.25 | 0.17 |
| 3 | 64 | 30 | 28.59 | −27.08 | −774.22 | 817.39 |
| 4 | 20 | 60 | −15.41 | 2.92 | −45.00 | 237.47 |
| 5 | 33 | 60 | −2.41 | 2.92 | −7.27 | 5.81 |
| 6 | 27 | 80 | −8.41 | 22.92 | −192.75 | 70.73 |
| 7 | 42 | 45 | 6.59 | −12.08 | −79.61 | 43.43 |
| 8 | 41 | 45 | 5.59 | −12.08 | −67.53 | 31.25 |
| 9 | 22 | 80 | −13.41 | 22.92 | −307.36 | 179.83 |
| 10 | 50 | 35 | 14.59 | −22.08 | −322.14 | 212.87 |
| 11 | 36 | 50 | 0.59 | −7.08 | −4.18 | 0.35 |
| 12 | 31 | 45 | −4.41 | −12.08 | 53.27 | 19.45 |
| Total | | | | | −2125.65 | 1748.94 |

Using (11.6), we can now easily find the least squares estimates $\hat{\alpha}$ and $\hat{\beta}$ as

$$\hat{\beta} = \frac{S_{xy}}{S_{xx}} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} = \frac{-2125.65}{1748.94} \approx -1.22,$$

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} = 57.08 - (-1.215) \cdot 35.41 = 100.28.$$

The fitted regression line is therefore

$$y = 100.28 - 1.22 \cdot x.$$

We can interpret the results as follows:

- For an increase of 1 min in exercising, the recovery time decreases by 1.22 days because $\hat{\beta} = -1.22$. The negative sign of $\hat{\beta}$ indicates that the relationship between exercising time and recovery time is negative; i.e. as exercise time increases, the recovery time decreases.
- When comparing two patients with a difference in exercising time of 10 min, the linear model estimates a difference in recovery time of 12.2 days because $10 \cdot 1.22 = 12.2$.

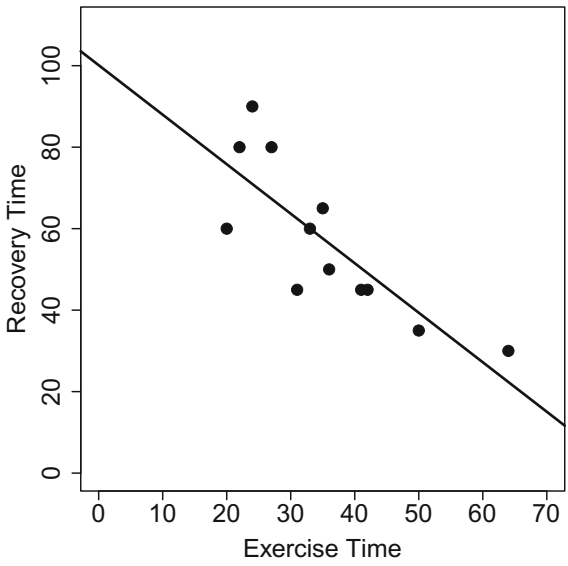



**Fig. 11.3** Scatter plot and regression line for Example 11.2.1

- The model predicts an average recovery time of $100.28 - 1.22 \cdot 38 = 53.92$ days for a patient who practises the set of exercises for 38 min.
- If a patient did not exercise at all, the recovery time would be predicted as $\hat{a} = 100.28$ days by the model. However, see item (i) in Sect. 11.2.1 below about interpreting a regression line outside of the observed data range.

We can also obtain these results by using *R*. The command `lm(Y~X)` fits a linear model and provides the estimates of $\hat{\alpha}$ and $\hat{\beta}$.

```
lm(Y~X)
```

We can draw the regression line onto a scatter plot using the command `abline`, see also Fig. 11.3.

```
plot(X,Y)
abline(a=100.28,b=-1.22)
```

### 11.2.1 Properties of the Linear Regression Line

There are a couple of interesting results related to the regression line and the least square estimates.

(i) As a rule of thumb, one should interpret the regression line $\hat{y}_i = \hat{\alpha} + \hat{\beta} x_i$ only in the interval $[x_{(1)}, x_{(n)}]$. For example, if $X$ denotes "Year", with observed values

from 1999 to 2015, and $Y$ denotes the "annual volume of sales of a particular company", then a prediction for the year 2030 may not be meaningful or valid because a linear relationship discovered in the past may not continue to hold in the future.

(ii) For the points $\hat{P}_i = (x_i, \hat{y}_i)$, forming the regression line, we can write

$$\hat{y}_i = \hat{\alpha} + \hat{\beta} x_i = \bar{y} + \hat{\beta}(x_i - \bar{x}). \tag{11.8}$$

(iii) It follows for $x_i = \bar{x}$ that $\hat{y}_i = \bar{y}$, i.e. the point $(\bar{x}, \bar{y})$ always lies on the regression line. The linear regression line therefore always passes through $(\bar{x}, \bar{y})$.

(iv) The sum of the residuals is zero. The $i$th residual is defined as

$$\hat{e}_i = y_i - \hat{y}_i = y_i - (\hat{\alpha} + \hat{\beta} x_i) = y_i - [\bar{y} + \hat{\beta}(x_i - \bar{x})].$$

The sum is therefore

$$\begin{aligned} \sum_{i=1}^{n} \hat{e}_i &= \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \bar{y} - \hat{\beta} \sum_{i=1}^{n} (x_i - \bar{x}) \\ &= n\bar{y} - n\bar{y} - \hat{\beta}(n\bar{x} - n\bar{x}) = 0. \end{aligned} \tag{11.9}$$

(v) The arithmetic mean of $\hat{y}$ is the same as the arithmetic mean of $y$:

$$\bar{\hat{y}} = \frac{1}{n} \sum_{i=1}^{n} \hat{y}_i = \frac{1}{n}(n\bar{y} + \hat{\beta}(n\bar{x} - n\bar{x})) = \bar{y}. \tag{11.10}$$

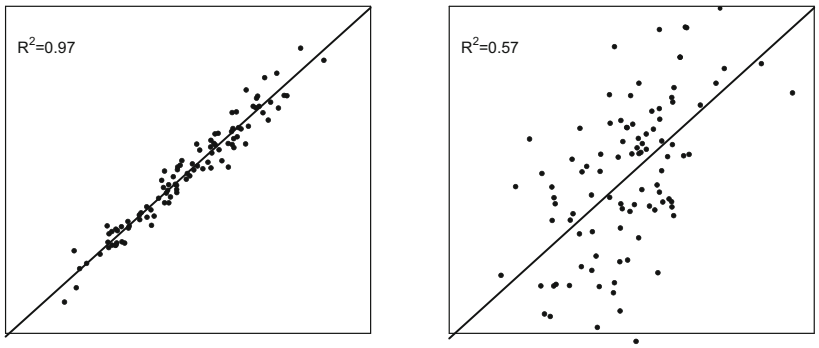
(vi) The least squares estimate $\hat{\beta}$ has a direct relationship with the correlation coefficient of Bravais–Pearson:

$$\hat{\beta} = \frac{S_{xy}}{S_{xx}} = \frac{S_{xy}}{\sqrt{S_{xx}}\sqrt{S_{yy}}} \cdot \sqrt{\frac{S_{yy}}{S_{xx}}} = r\sqrt{\frac{S_{yy}}{S_{xx}}}. \tag{11.11}$$

The slope of the regression line is therefore proportional to the correlation coefficient $r$: a positive correlation coefficient implies a positive estimate of $\beta$ and vice versa. However, a stronger correlation does not necessarily imply a steeper slope in the regression analysis because the slope depends upon $\sqrt{S_{yy}/S_{xx}}$ as well.

## 11.3 Goodness of Fit

While one can easily fit a linear regression model, this does not mean that the model is necessarily good. Consider again Fig. 11.1: In Fig. 11.1a, the regression line is clearly capturing the linear trend seen in the scatter plot. The fit of the model to the data seems good. Figure 11.1b shows however that the data points vary considerably around the line. The quality of the model does not seem to be very good. If we would use the regression line to predict the data, we would likely obtain poor results. It is obvious from Fig. 11.1 that the model provides a good fit to the data when the observations are close to the line and capture a linear trend.




(a) Strong linear relationship (b) Weak linear relationship

**Fig. 11.4** Different goodness of fit for different data

A look at the scatter plot provides a visual and qualitative approach to judging the quality of the fitted model. Consider Fig. 11.4a, b which both show a linear trend in the data, but the observations in Fig. 11.4a are closer to the regression line than those in Fig. 11.4b. Therefore, the goodness of fit is worse in the latter figure and any quantitative measure should capture this.

A quantitative measure for the goodness of fit can be derived by means of **variance decomposition** of the data. The total variation of $y$ is partitioned into two components—sum of squares due to the fitted model and sum of squares due to random errors in the data:

$$\underbrace{\sum_{i=1}^{n}(y_i - \bar{y})^2}_{SQ_{\text{Total}}} = \underbrace{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}_{SQ_{\text{Regression}}} + \underbrace{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}_{SQ_{\text{Error}}}. \tag{11.12}$$

The proof of the above equation is given in Appendix C.6.

The left-hand side of (11.12) represents the total variability in $y$ with respect to $\bar{y}$. It is proportional to the sample variance (3.21) and is called $SQ_{\text{Total}}$ (total sum of squares). The first term on the right-hand side of the equation measures the variability which is explained by the fitted linear regression model ($SQ_{\text{Regression}}$, sum of squares due to regression); the second term relates to the error sum of squares ($SQ_{\text{Error}}$) and reflects the variation due to random error involved in the fitted model. Larger values of $SQ_{\text{Error}}$ indicate that the deviations of the observations from the regression line are large. This implies that the model provides a bad fit to the data and that the goodness of fit is low.

Obviously, one would like to have a fit in which the error sum of squares is as small as possible. To judge the goodness of fit, one can therefore look at the error sum of squares in relation to the total sum of squares: in an ideal situation, if the error sum of squares equals zero, the total sum of squares is equal to the sum of squares due to regression and the goodness of fit is optimal. On the other hand, if the sum of squares due to error is large, it will make the sum of squares due to regression smaller

and the goodness of fit should be bad. If the sum of squares due to regression is zero, it is evident that the model fit is the worst possible. These thoughts are reflected in the criterion for the goodness of fit, also known as $R^2$:

$$R^2 = \frac{SQ_{\text{Regression}}}{SQ_{\text{Total}}} = 1 - \frac{SQ_{\text{Error}}}{SQ_{\text{Total}}}. \tag{11.13}$$

It follows from the above definition that $0 \leq R^2 \leq 1$. The closer $R^2$ is to 1, the better the fit because $SQ_{\text{Error}}$ will then be small. The closer $R^2$ is to 0, the worse the fit, because $SQ_{\text{Error}}$ will then be large. If $R^2$ takes any other value, say $R^2 = 0.7$, it means that only 70 % of the variation in data is explained by the fitted model, and hence, in simple language, the model is 70 % good. An important point to remember is that $R^2$ is defined only when there is an intercept term in the model (an assumption we make throughout this chapter). So it is not used to measure the goodness of fit in models without an intercept term.

*Example 11.3.1* Consider again Fig. 11.1: In Fig. 11.1a, the line and data points fit well together. As a consequence $R^2$ is high, $R^2 = 0.82$. Figure 11.1b shows data points with large deviations from the regression line; therefore, $R^2$ is small, here $R^2 = 0.002$. Similarly, in Fig. 11.4a, an $R^2$ of 0.97 relates to an almost perfect model fit, whereas in Fig. 11.4b, the model describes the data only moderately well ($R^2 = 0.57$).

*Example 11.3.2* Consider again Example 11.2.1 where we analysed the relationship between exercise intensity and recovery time for patients convalescing from knee surgery. To calculate $R^2$, we need the following table:

| $i$ | $y_i$ | $\hat{y}_i$ | $(y_i - \bar{y})$ | $(y_i - \bar{y})^2$ | $(\hat{y}_i - \bar{y})$ | $(\hat{y}_i - \bar{y})^2$ |
|---|---|---|---|---|---|---|
| 1 | 90 | 70.84 | 32.92 | 1083.73 | 13.76 | 189.34 |
| 2 | 65 | 57.42 | 7.92 | 62.73 | 0.34 | 0.12 |
| 3 | 30 | 22.04 | −27.08 | 733.33 | −35.04 | 1227.80 |
| 4 | 60 | 75.72 | 2.92 | 8.53 | 18.64 | 347.45 |
| 5 | 60 | 59.86 | 2.92 | 8.53 | 2.78 | 7.73 |
| 6 | 80 | 67.18 | 22.92 | 525.33 | 10.10 | 102.01 |
| 7 | 45 | 48.88 | −12.08 | 145.93 | −8.2 | 67.24 |
| 8 | 45 | 50.10 | −12.08 | 145.93 | −6.83 | 48.72 |
| 9 | 80 | 73.28 | 22.92 | 525.33 | 16.20 | 262.44 |
| 10 | 35 | 39.12 | −22.08 | 487.53 | −17.96 | 322.56 |
| 11 | 50 | 56.20 | −7.08 | 50.13 | −0.88 | 0.72 |
| 12 | 45 | 62.30 | −12.08 | 145.93 | 5.22 | 27.25 |
| Total | | | | 3922.96 | | 2603.43 |

We calculate $R^2$ with these results as

$$R^2 = \frac{SQ_{\text{Regression}}}{SQ_{\text{Total}}} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = \frac{2603.43}{3922.96} = 0.66.$$

We conclude that the regression model provides a reasonable but not perfect fit to the data because 0.66 is not close to 0, but also not very close to 1. About 66 % of the variability in the data can be explained by the fitted linear regression model. The rest is random error: for example, individual variation in the recovery time of patients due to genetic and environmental factors, surgeon performance, and others.

We can also obtain this result in *R* by looking at the summary of the linear model:

```
summary(lm(Y~X))
```

Please note that we give a detailed explanation of the model summary in Sect. 11.7.

There is a direct relationship between $R^2$ and the correlation coefficient of Bravais–Pearson $r$:

$$R^2 = r^2 = \left( \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}} \right)^2 . \tag{11.14}$$

The proof is given in Appendix C.6.

*Example 11.3.3* Consider again Examples 11.3 and 11.5 where we analysed the association of exercising time and time to regain full range of motion after knee surgery. We calculated $R^2 = 0.66$. We therefore know that the correlation coefficient is $r = \sqrt{R^2} = \sqrt{0.66} \approx 0.81$.

## 11.4 Linear Regression with a Binary Covariate

Until now, we have assumed that the covariate $X$ is continuous. It is however also straightforward to fit a linear regression model when $X$ is binary, i.e. if $X$ has two categories. In this case, the values of $X$ in the first category are usually coded as 0 and the values of $X$ in the second category are coded as 1. For example, if the binary variable is "gender", we replace the word "male" with the number 0 and the word "female" with 1. We can then fit a linear regression model using these numbers, but the interpretation differs from the interpretations in case of a continuous variable. Recall the definition of the linear model, $Y = \alpha + \beta X + e$ with $\mathrm{E}(e) = 0$; if $X = 0$ (male) then $\mathrm{E}(Y|X = 0) = \alpha$ and if $X = 1$ (female), then $\mathrm{E}(Y|X = 1) = \alpha + \beta \cdot 1 = \alpha + \beta$. Thus, $\alpha$ is the average value of $Y$ for males, i.e. $E(Y|X = 0)$, and $\beta = E(Y|X = 1) - E(Y|X = 0)$. It follows that those subjects with $X = 1$ (e.g. females) have on average $Y$-values which are $\beta$ units higher than subjects with $X = 0$ (e.g. males).

*Example 11.4.1* Recall Examples 11.2.1, 11.3.2, and 11.3.3 where we analysed the association of exercising time and recovery time after knee surgery. We keep the values of $Y$ (recovery time, in days) and replace values of $X$ (exercising time, in minutes) with 0 for patients exercising for less than 40 min and with 1 for patients exercising for 40 min or more. We have therefore a new variable $X$ indicating whether a patient is exercising a lot ($X = 1$) or not ($X = 0$). To estimate the linear regression line $\hat{y} = \hat{\alpha} + \hat{\beta}x$, we first calculate $\bar{x} = 4/12$ and $\bar{y} = 57.08$. To obtain $S_{xy}$ and $S_{xx}$, we need the following table:

| $i$ | $x_i$ | $y_i$ | $(x_i - \bar{x})$ | $(y_i - \bar{y})$ | $(x_i - \bar{x})(y_i - \bar{y})$ | $(x_i - \bar{x})^2$ |
|---|---|---|---|---|---|---|
| 1 | 0 | 90 | $-\frac{4}{12}$ | 32.92 | −10.97 | 0.11 |
| 2 | 0 | 65 | $-\frac{4}{12}$ | 7.92 | −2.64 | 0.11 |
| 3 | 1 | 30 | $\frac{8}{12}$ | −27.08 | −18.05 | 0.44 |
| 4 | 0 | 60 | $-\frac{4}{12}$ | 2.92 | −0.97 | 0.11 |
| 5 | 0 | 60 | $-\frac{4}{12}$ | 2.92 | −0.97 | 0.11 |
| 6 | 0 | 80 | $-\frac{4}{12}$ | 22.92 | −7.64 | 0.11 |
| 7 | 1 | 45 | $\frac{8}{12}$ | −12.08 | −8.05 | 0.44 |
| 8 | 1 | 45 | $\frac{8}{12}$ | −12.08 | −8.05 | 0.44 |
| 9 | 0 | 80 | $-\frac{4}{12}$ | 22.92 | −7.64 | 0.11 |
| 10 | 1 | 35 | $\frac{8}{12}$ | −22.08 | −14.72 | 0.44 |
| 11 | 0 | 50 | $-\frac{4}{12}$ | −7.08 | 2.36 | 0.11 |
| 12 | 0 | 45 | $-\frac{4}{12}$ | −12.08 | 4.03 | 0.11 |
| Total | Total | | | | −72.34 | 2.64 |

We can now calculate the least squares estimates of $\alpha$ and $\beta$ using (11.6) as

$$\hat{\beta} = \frac{S_{xy}}{S_{xx}} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} = \frac{-72.34}{2.64} \approx -27.4,$$

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} = 57.08 - (-27.4) \cdot \frac{4}{12} = 66.2.$$

The fitted regression line is:

$$y = 66.2 - 27.4 \cdot x.$$

The interpretation is:

- The average recovery time for patients doing little exercise is $66.2 - 27.4 \cdot 0 = 66.2$ days.
- Patients exercising heavily ($x = 1$) have, on average, a 27.4 days shorter recovery period ($\beta = -27.4$) than patients with a short exercise time ($x = 0$).
- The average recovery time for patients exercising heavily is 38.8 days ($66.2 - 27.4 \cdot 1$ for $x = 1$).
- These interpretations are visualized in Fig. 11.5. Because "exercise time" ($X$) is considered to be a nominal variable (two categories representing the intervals $[0; 40)$ and $(40; \infty)$), we can plot the regression line on a scatter plot as two parallel lines.

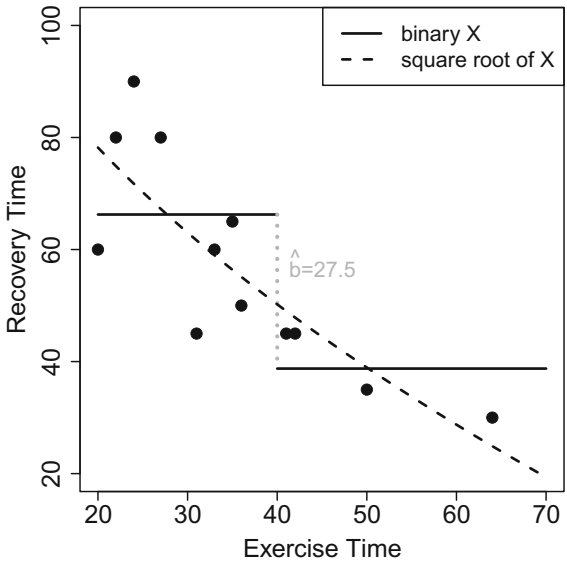



**Fig. 11.5** Scatter plot and regression lines for Examples 11.4.1 and 11.5.1

## 11.5 Linear Regression with a Transformed Covariate

Recall that a model is said to be linear when it is linear in its parameters. The definition of a linear model is not at all related to the linearity in the covariate. For example,

$$Y = \alpha + \beta^2 X + e$$

is *not* a linear model because the right-hand side of the equation is not a linear function in $\beta$. However,

$$Y = \alpha + \beta X^2 + e$$

is a linear model. This model can be fitted as for any other linear model: we obtain $\hat{\alpha}$ and $\hat{\beta}$ as usual and simply use the squared values of $X$ instead of $X$. This can be justified by considering $X^* = X^2$, and then, the model is written as $Y = \alpha + \beta X^* + e$ which is again a linear model. Only the interpretation changes: for each unit increase in the squared value of $X$, i.e. $X^*$, $Y$ increases by $\beta$ units. Such an interpretation is often not even needed. One could simply plot the regression line of $Y$ on $X^*$ to visualize the functional form of the effect. This is also highlighted in the following example.

*Example 11.5.1* Recall Examples 11.2.1, 11.3.2, 11.3.3, and 11.4.1 where we analysed the association of exercising time and recovery time after knee surgery. We estimated $\hat{\beta}$ as $-1.22$ by using $X$ as it is, as a continuous variable, see also Fig. 11.3. When using a binary $X$, based on a cut-off of 40 min, we obtained $\hat{\beta} = -27.4$, see also Fig. 11.5. If we now use $\sqrt{X}$ rather than $X$, we obtain $\hat{\beta} = -15.1$. This means for

an increase of 1 unit of the square root of exercising time, the recovery time decreases by 15.1 days. Such an interpretation will be difficult to understand for many people. It is better to plot the linear regression line $y = 145.8 - 15.1 \cdot \sqrt{x}$, see Fig. 11.5. We can see that the new nonlinear line (obtained from a *linear* model) fits the data nicely and it may even be preferable to a straight regression line. Moreover, the value of $\alpha$ substantially increased with this modelling approach, highlighting that no exercising at all may severely delay recovery from the surgery (which is biologically more meaningful). In *R*, we obtain these results by either creating a new variable $\sqrt{X}$ or by using the `I()` command which allows specifying transformations in regression models.

```
newX <- sqrt(X)
lm(Y~newX)          #option 1
lm(Y~I(sqrt(X)))    #option 2
```

Generally the covariate $X$ can be replaced by any transformation of it, such as using $\log(X)$, $\sqrt{X}$, $\sin(X)$, or $X^2$. The details are discussed in Sect. 11.6.3. It is also possible to compare the quality and goodness of fit of models with different transformations (see Sect. 11.8).

## 11.6 Linear Regression with Multiple Covariates

Up to now, we have only considered two variables—$Y$ and $X$. We therefore assume that $Y$ is affected by only one covariate $X$. However, in many situations, there might be more than one covariate which affects $Y$ and consequently all of them are relevant to the analysis (see also Sect. 11.10 on the difference between association and causation). For example, the yield of a crop depends on several factors such as quantity of fertilizer, irrigation, and temperature, among others. In another example, in the pizza data set (described in Appendix A.4), many variables could potentially be associated with delivery time—driver, amount of food ordered, operator handling the order, whether it is a weekend, and so on. A reasonable question to ask is: Do different drivers have (on average) a different delivery time—given they deal with the same operator, the amount of food ordered is the same, the day is the same, etc. These kind of questions can be answered with **multiple linear regression**. The model contains more than one, say $p$, covariates $X_1, X_2, \ldots, X_p$. The linear model (11.2) can be extended as

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + e. \tag{11.15}$$

Note that the intercept term is denoted here by $\beta_0$. In comparison with (11.2), $\alpha = \beta_0$ and $\beta = \beta_1$.

*Example 11.6.1* For the pizza delivery data, a particular linear model with multiple covariates could be specified as follows:

$$\begin{aligned}\text{Delivery Time} = \beta_0 &+ \beta_1 \text{Number pizzas ordered} + \beta_2 \text{Weekend(0/1)} \\ &+ \beta_3 \text{Operator (0/1)} + e.\end{aligned}$$

### 11.6.1 Matrix Notation

If we have a data set of $n$ observations, then every set of observations satisfies model (11.15) and we can write

$$\begin{aligned} y_1 &= \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + \cdots + \beta_p x_{1p} + e_1 \\ y_2 &= \beta_0 + \beta_1 x_{21} + \beta_2 x_{22} + \cdots + \beta_p x_{2p} + e_2 \\ &\vdots \qquad\qquad \vdots \qquad\qquad \vdots \\ y_n &= \beta_0 + \beta_1 x_{n1} + \beta_2 x_{n2} + \cdots + \beta_p x_{np} + e_n \end{aligned} \tag{11.16}$$

It is possible to write the $n$ equations in (11.16) in matrix notation as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e} \tag{11.17}$$

where

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & & & & \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}, \quad \mathbf{e} = \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix}.$$

The letters $\mathbf{y}$, $\mathbf{X}$, $\boldsymbol{\beta}$, $\mathbf{e}$ are written in bold because they refer to vectors and matrices rather than scalars. The capital letter $\mathbf{X}$ makes it clear that $\mathbf{X}$ is a matrix of order $n \times p$ representing the $n$ observations on each of the covariates $X_1, X_2, \ldots, X_p$. Similarly, $\mathbf{y}$ is a $n \times 1$ vector of $n$ observations on $Y$, $\boldsymbol{\beta}$ is a $p \times 1$ vector of regression coefficients associated with $X_1, X_2, \ldots, X_p$, and $\mathbf{e}$ is a $n \times 1$ vector of $n$ errors. The lower case letter $\mathbf{x}$ relates to a vector representing a variable which means we can denote the multiple linear model from now on also as

$$\mathbf{y} = \beta_0 \mathbf{1} + \beta_1 \mathbf{x_1} + \cdots + \beta_p \mathbf{x_p} + \mathbf{e} \tag{11.18}$$

where $\mathbf{1}$ is the $n \times 1$ vector of 1's. We assume that $\mathrm{E}(\mathbf{e}) = 0$ and $\mathrm{Cov}(\mathbf{e}) = \sigma^2 I$ (see Sect. 11.9 for more details).

We would like to highlight that $\mathbf{X}$ is not the data matrix. The matrix $\mathbf{X}$ is called the **design matrix** and contains both a column of 1's denoting the presence of the intercept term and all explanatory variables which are relevant to the multiple linear model (including possible transformations and interactions, see Sects. 11.6.3 and 11.7.3). The errors $\mathbf{e}$ reflect the deviations of the observations from the regression line and therefore the difference between the observed and fitted relationships. Such deviations may occur for various reasons. For example, the measured values can be

affected by variables which are not included in the model, the variables may not be accurately measured, there is unmeasured genetic variability in the study population, and all of which are covered in the error **e**. The estimate of $\beta$ is obtained by using the least squares principle by minimizing $\sum_{i=1}^{n} e_i^2 = e'e$. The **least squares estimate** of $\beta$ is given by

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}. \tag{11.19}$$

The vector $\hat{\boldsymbol{\beta}}$ contains the estimates of $(\beta_0, \beta_1, \ldots, \beta_p)'$. We can interpret it as earlier: $\hat{\beta}_0$ is the estimate of the intercept which we obtain because we have added the column of 1's (and is identical to $\alpha$ in (11.1)). The estimates $\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_p$ refer to the regression coefficients associated with the variables $\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_p}$, respectively. The interpretation of $\hat{\beta}_j$ is that it represents the partial change in $y_i$ when the value of $x_i$ changes by one unit keeping all other covariates fixed.

A possible interpretation of the intercept term is that when all covariates equal zero then

$$\mathrm{E}(\mathbf{y}) = \beta_0 + \beta_1 \cdot 0 + \cdots + \beta_p \cdot 0 = \beta_0. \tag{11.20}$$

There are many situations in real life for which there is no meaningful interpretation of the intercept term because one or many covariates cannot take the value zero. For instance, in the pizza data set, the bill can never be €0, and thus, there is no need to describe the average delivery time for a given bill of €0. The intercept term here serves the purpose of improvement of the model fit, but it will not be interpreted.

In some situations, it may happen that the average value of $y$ is zero when all covariates are zero. Then, the intercept term is zero as well and does not improve the model. For example, suppose the salary of a person depends on two factors—education level and type of qualification. If any person is completely illiterate, even then we observe in practice that his salary is never zero. So in this case, there is a benefit of the intercept term. In another example, consider that the velocity of a car depends on two variables—acceleration and quantity of petrol. If these two variables take values of zero, the velocity is zero on a plane surface. The intercept term will therefore be zero as well and yields no model improvement.

*Example 11.6.2* Consider the pizza data described in Appendix A.4. The data matrix $\mathcal{X}$ is as follows:

$$\mathcal{X} = \begin{array}{c} \begin{array}{cccccc} \text{Day} & \text{Date} & \text{Time} & \cdots & \text{Discount} \end{array} \\ \begin{pmatrix} \text{Thursday} & \text{1-May-14} & 35.1 & \cdots & 1 \\ \text{Thursday} & \text{1-May-14} & 25.2 & \cdots & 0 \\ \vdots & & & & \vdots \\ \text{Saturday} & \text{31-May-14} & 35.7 & \cdots & 0 \end{pmatrix} \end{array}$$

Suppose the manager has the hypothesis that the operator and the overall bill (as a proxy for the amount ordered from the customer) influence the delivery time. We can postulate a linear model to describe this relationship as

$$\text{Delivery Time} = \beta_0 + \beta_1 \text{Bill} + \beta_2 \text{Operator} + e.$$

The model in matrix notation is as follows:

$$\underbrace{\begin{pmatrix} 35.1 \\ 25.2 \\ \vdots \\ 35.7 \end{pmatrix}}_{\mathbf{y}} = \underbrace{\begin{pmatrix} 1 & 58.4 & 1 \\ 1 & 26.4 & 0 \\ \vdots & \vdots & \vdots \\ 1 & 42.7 & 0 \end{pmatrix}}_{\mathbf{X}} \underbrace{\begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix}}_{\boldsymbol{\beta}} + \underbrace{\begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_{1266} \end{pmatrix}}_{\mathbf{e}}$$

To understand the associations in the data, we need to estimate $\boldsymbol{\beta}$ which we obtain by the least squares estimator $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$. Rather than doing this tiresome task manually, we use *R*:

```
lm(time~bill+operator)
```

If we have more than one covariate in a linear regression model, we simply add all of them separated by the + sign when specifying the model formula in the `lm()` function. We obtain $\hat{\beta}_0 = 23.1$, $\hat{\beta}_1 = 0.26$, $\hat{\beta}_2 = 0.16$. The interpretation of these parameters is as follows:

- For each extra € that is spent, the delivery time increases by 0.26 min. Or, for each extra €10 spent, the delivery time increases on average by 2.6 min.
- The delivery time is on average 0.16 min longer for operator 1 compared with operator 0.
- For operator 0 and a bill of €0, the expected delivery time is $\beta_0 = 23.1$ min. However, there is no bill of €0, and therefore, the intercept $\beta_0$ cannot be interpreted meaningfully here and is included only for improvement of the model fit.

### 11.6.2 Categorical Covariates

We now consider the case when covariates are categorical rather than continuous.

*Examples*

- Region: East, West, South, North,
- Marital status: single, married, divorced, widowed,
- Day: Monday, Tuesday, . . ., Sunday.

We have already described how to treat a categorical variable which consists of two categories, i.e. a binary variable: one category is replaced by 1's and the other category is replaced by 0's subjects who belong to the category $x = 1$ have on average $y$-values which are $\hat{\beta}$ units higher than those subjects with $x = 0$.

Consider a variable $x$ which has more than two categories, say $k > 2$ categories. To include such a variable in a regression model, we create $k - 1$ new variables $x_i$, $i = 1, 2, \ldots, k - 1$. Similar to how we treat binary variables, each of these variables is a **dummy variable** and consists of 1's for units which belong to the category of interest and 0's for the other category

$$x_i = \begin{cases} 1 \text{ for category } i, \\ 0 \text{ otherwise.} \end{cases} \tag{11.21}$$

The category for which we do not create a dummy variable is called the **reference category**, and the interpretation of the parameters of the dummy variables is with respect to this category. For example, for category $i$, the $\mathbf{y}$-values are on average $\beta_i$ higher than for the reference category. The concept of creating dummy variables and interpreting them is explained in the following example.

*Example 11.6.3* Consider again the pizza data set described in Appendix A.4. The manager may hypothesize that the delivery times vary with respect to the branch. There are $k = 3$ branches: East, West, Centre. Instead of using the variable $\mathbf{x} =$ branch, we create $(k - 1)$, i.e. $(3 - 1) = 2$ new variables denoting $\mathbf{x_1} =$ East and $\mathbf{x_2} =$ West. We set $\mathbf{x_1} = 1$ for those deliveries which come from the branch in the East and set $\mathbf{x_1} = 0$ for other deliveries. Similarly, we set $\mathbf{x_2} = 1$ for those deliveries which come from the West and $\mathbf{x_2} = 0$ for other deliveries. The data then is as follows:

| Delivery | $\mathbf{y}$ (Delivery Time) | $\mathbf{x}$ (Branch) | $\mathbf{x_1}$ (East) | $\mathbf{x_2}$ (West) |
|---|---|---|---|---|
| 1 | 35.1 | East | 1 | 0 |
| 2 | 25.2 | East | 1 | 0 |
| 3 | 45.6 | West | 0 | 1 |
| 4 | 29.4 | East | 1 | 0 |
| 5 | 30.0 | West | 0 | 1 |
| 6 | 40.3 | Centre | 0 | 0 |
| ⋮ | ⋮ | ⋮ | ⋮ | |
| 1266 | 35.7 | West | 0 | 1 |

Deliveries which come from the East have $\mathbf{x_1} = 1$ and $\mathbf{x_2} = 0$, deliveries which come from the West have $\mathbf{x_1} = 0$ and $\mathbf{x_2} = 1$, and deliveries from the Centre have $\mathbf{x_1} = 0$ and $\mathbf{x_2} = 0$. The regression model of interest is thus

$$\mathbf{y} = \beta_0 + \beta_1 \text{East} + \beta_2 \text{West} + e.$$

We can calculate the least squares estimate $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2)' = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ via *R*: either (i) we create the dummy variables ourselves or (ii) we ask *R* to create it for us. This requires that "branch" is a factor variable (which it is in our data, but if it was not, then we would have to define it in the model formula via `as.factor()`).

```
East <- as.numeric(branch=='East')
West <- as.numeric(branch=='West')
lm(time~ East+West)              # option 1
lm(time~ branch)                 # option 2a
lm(time~ as.factor(branch))      # option 2b
```

We obtain the following results:

$$Y = 36.3 - 5.2\text{East} - 1.1\text{West}.$$

The interpretations are as follows:

- The average delivery time for the branch in the Centre is 36.3 min, the delivery time for the branch in the East is $36.3 - 5.2 = 31.1$ min, and the predicted delivery time for the West is $36.3 - 1.1 = 35.2$ min.
- Therefore, deliveries arrive on average 5.2 min earlier in the East compared with the centre ($\hat{\beta}_1 = -5.2$), and deliveries in the West arrive on average 1.1 min earlier than in the centre ($\hat{\beta}_2 = -1.1$). In other words, it takes 5.2 min less to deliver a pizza in the East than in the Centre. The deliveries in the West take on average 1.1 min less than in the Centre.

Consider now a covariate with $k = 3$ categories for which we create two new dummy variables, $\mathbf{x_1}$ and $\mathbf{x_2}$. The linear model is $\mathbf{y} = \beta_0 + \beta_1\mathbf{x_1} + \beta_2\mathbf{x_2} + e$.

- For $\mathbf{x_1} = 1$, we obtain $\mathrm{E}(\mathbf{y}) = \beta_0 + \beta_1 \cdot 1 + \beta_2 \cdot 0 = \beta_0 + \beta_1 \equiv E(\mathbf{y}|\mathbf{x_1} = 1, \mathbf{x_2} = 0)$;
- For $\mathbf{x_2} = 1$, we obtain $\mathrm{E}(\mathbf{y}) = \beta_0 + \beta_1 \cdot 0 + \beta_2 \cdot 1 = \beta_0 + \beta_2 \equiv E(\mathbf{y}|\mathbf{x_1} = 0, \mathbf{x_2} = 1)$; and
- For the reference category ($\mathbf{x_1} = \mathbf{x_2} = \mathbf{0}$), we obtain $\mathbf{y} = \beta_0 + \beta_1 \cdot 0 + \beta_2 \cdot 0 = \beta_0 \equiv \mathrm{E}(\mathbf{y}|\mathbf{x_1} = \mathbf{x_2} = 0)$.

We can thus conclude that the intercept $\beta_0 = \mathrm{E}(\mathbf{y}|\mathbf{x_1} = 0, \mathbf{x_2} = 0)$ describes the average $\mathbf{y}$ for the reference category, that $\beta_1 = \mathrm{E}(\mathbf{y}|\mathbf{x_1} = 1, \mathbf{x_2} = 0) - E(\mathbf{y}|\mathbf{x_1} = 0, \mathbf{x_2} = 0)$ describes the difference between the first and the reference category, and that $\beta_2 = \mathrm{E}(\mathbf{y}|\mathbf{x_1} = 0, \mathbf{x_2} = 1) - \mathrm{E}(\mathbf{y}|\mathbf{x_1} = 0, \mathbf{x_2} = 0)$ describes the difference between the second and the reference category.

*Remark 11.6.1* There are other ways of recoding categorical variables, such as effect coding. However, we do not describe them in this book.

### 11.6.3 Transformations

As we have already outlined in Sect. 11.5, if $\mathbf{x}$ is transformed by a function $T(\mathbf{x}) = (T(x_1), T(x_2), \ldots, T(x_n))$ then

$$\mathbf{y} = f(\mathbf{x}) = \beta_0 + \beta_1 T(\mathbf{x}) + e \tag{11.22}$$

is still a linear model because the model is linear in its parameters $\beta$. Popular transformations $T(\mathbf{x})$ are $\log(\mathbf{x})$, $\sqrt{\mathbf{x}}$, $\exp(\mathbf{x})$, $\mathbf{x}^p$, among others. The choice of such a function is typically motivated by the application and data at hand. Alternatively, a relationship between $\mathbf{y}$ and $\mathbf{x}$ can be modelled via a **polynomial** as follows:

$$\mathbf{y} = f(\mathbf{x}) = \beta_0 + \beta_1\mathbf{x} + \beta_2\mathbf{x}^2 + \cdots + \beta_p\mathbf{x}^p + e. \tag{11.23}$$

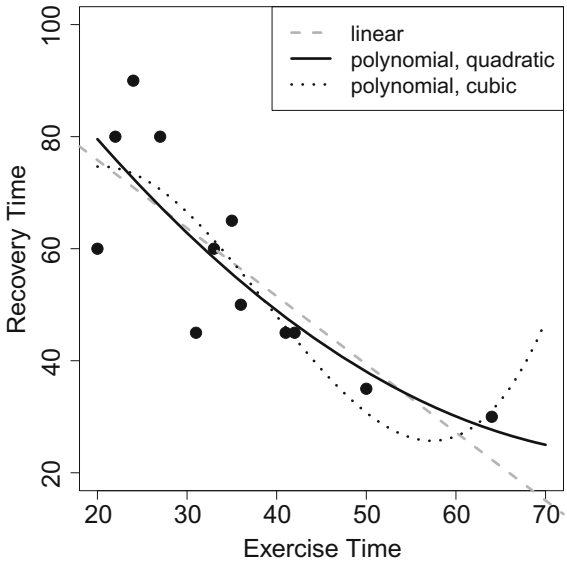



**Fig. 11.6** Scatter plot and regression lines for Example 11.6.4

*Example 11.6.4* Consider again Examples 11.2.1, 11.3.2, 11.3.3, 11.4.1, and 11.5.1 where we analysed the association of intensity of exercising and recovery time after knee surgery. The linear regression line was estimated as

$$\text{Recovery Time} = 100.28 - 1.22\,\text{Exercising Time}.$$

One could question whether the association is indeed linear and fit the second- and third-order polynomials:

$$\text{Recovery Time} = \beta_0 - \beta_1 \text{Exercise} + \beta_2 \text{Exercise}^2,$$
$$\text{Recovery Time} = \beta_0 - \beta_1 \text{Exercise} + \beta_2 \text{Exercise}^2 + \beta_3 \text{Exercise}^3.$$

To obtain the estimates we can use *R*:

```
lm(Y~X+I(X^2))
lm(Y~X+I(X^2)+I(X^3))
```

The results are $\beta_0 = 121.8$, $\beta_1 = -2.4$, $\beta_2 = 0.014$ and $\beta_0 = 12.9$, $\beta_1 = 6.90$, $\beta_2 = -0.23$, $\beta_3 = 0.002$ for the two models respectively. While it is difficult to interpret these coefficients directly, we can simply plot the regression lines on a scatter plot (see Fig. 11.6).

We see that the regression based on the quadratic polynomial visually fits the data slightly better than the linear polynomial. It seems as if the relation between recovery time and exercise time is not exactly linear and the association is better modelled through a second-order polynomial. The regression line based on the cubic polynomial seems to be even closer to the measured points; however, the functional form of the association looks questionable. Driven by a single data point, we obtain a

regression line which suggests a heavily increased recovery time for exercising times greater than 65 min. While it is possible that too much exercising causes damage to the knee and delays recovery, the data does not seem to be clear enough to support the shape suggested by the model. This example illustrates the trade-off between fit (i.e. how good the regression line fits the data) and parsimony (i.e. how well a model can be understood if it becomes more complex). Section 11.8 explores this non-trivial issue in more detail.

**Transformations of the Outcome Variable.** It is also possible to apply transformations to the outcome variable. While in many situations, this makes interpretations quite difficult, a log transformation is quite common and easy to interpret. Consider the **log-linear model**

$$\log \mathbf{y} = \beta_0 \mathbf{1} + \beta_1 \mathbf{x_1} + \cdots + \beta_p \mathbf{x_p} + \mathbf{e}$$

where $y$ is required to be greater than 0. Exponentiating on both sides leads to

$$\mathbf{y} = e^{\beta_0 \mathbf{1}} \cdot e^{\beta_1 \mathbf{x_1}} \cdot \ldots \cdot e^{\beta_p \mathbf{x_p}} \cdot e^{\mathbf{e}}.$$

It can be easily seen that a one unit increase in $\mathbf{x_j}$ multiplies $\mathbf{y}$ by $e^{\beta_j}$. Therefore, if $\mathbf{y}$ is log-transformed, one simply interprets $\exp(\boldsymbol{\beta})$ instead of $\boldsymbol{\beta}$. For example, if $\mathbf{y}$ is the yearly income, $\mathbf{x}$ denotes gender (1 = male, 0 = female), and $\beta_1 = 0.2$ then one can say that a man's income is $\exp(0.2) = 1.22$ times higher (i.e. 22 %) than a woman's.

## 11.7 The Inductive View of Linear Regression

So far we have introduced linear regression as a method which *describes* the relationship between dependent and independent variables through the best fit to the data. However, as we have highlighted in earlier chapters, often we are interested not only in describing the data but also in drawing conclusions from a sample about the population of interest. For instance, in an earlier example, we have estimated the association of branch and delivery time for the pizza delivery data. The linear model was estimated as delivery time $= 36.3 - 5.2\,\text{East} - 1.1\,\text{West}$; this indicates that the delivery time is on average 5.2 min shorter for the branch in the East of the town compared with the central branch and 1.1 min shorter for the branch in the West compared with the central branch. When considering all pizza deliveries (the population), not only those collected by us over the period of a month (the sample), we might ask ourselves whether 1.1 min is a real difference valid for the entire population or just caused by random error involved in the sample we chose. In particular, we would also like to know what the 95 % confidence interval for our parameter estimate is. Does the interval cover "zero"? If yes, we might conclude that we cannot be very sure that there is an association and do not reject the null hypothesis that $\beta_{\text{West}} = 0$. In reference to Chaps. 9 and 10, we now apply the concepts of statistical inference and testing in the context of regression analysis.

**Point and interval estimation in the linear model.** We now rewrite the model for the purpose of introducing statistical inference to linear regression as

$$\begin{aligned} \mathbf{y} &= \beta_0 \mathbf{1} + \beta_1 \mathbf{x_1} + \cdots + \beta_p \mathbf{x_p} + \mathbf{e} \\ &= \mathbf{X}\boldsymbol{\beta} + \mathbf{e} \quad \text{with} \quad \mathbf{e} \sim N(\mathbf{0}, \sigma^2 \mathbf{I}) \,. \end{aligned} \tag{11.24}$$

where $\mathbf{y}$ is the $n \times 1$ vector of outcomes and $\mathbf{X}$ is the $n \times (p+1)$ design matrix (including a column of 1's for the intercept). The identity matrix $\mathbf{I}$ consists of 1's on the diagonal and 0's elsewhere, and the parameter vector is $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)'$. We would like to estimate $\boldsymbol{\beta}$ to make conclusions about a relationship in the population. Similarly, $\mathbf{e}$ reflects the random errors in the population. Most importantly, the linear model now contains assumptions about the errors. They are assumed to be normally distributed, $N(0, \sigma^2 I)$, which means that the expectation of the errors is 0, $\mathrm{E}(e_i) = 0$, the variance is $\mathrm{Var}(e_i) = \sigma^2$ (and therefore the same for *all* $e_i$), and it follows from $\mathrm{Var}(\mathbf{e}) = \sigma^2 \mathbf{I}$ that $\mathrm{Cov}(e_i, e_{i'}) = 0$ for all $i \neq i'$. The assumption of a normal distribution is required to construct confidence intervals and test of hypotheses statistics. More details about these assumptions and the procedures to test their validity on the basis of a given sample of data are explained in Sect. 11.9.

The least squares estimate of $\boldsymbol{\beta}$ is obtained by

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}. \tag{11.25}$$

It can be shown that $\hat{\boldsymbol{\beta}}$ follow a normal distribution with mean $\mathrm{E}(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$ and covariance matrix $\mathrm{Var}(\hat{\boldsymbol{\beta}}) = \sigma^2 \mathbf{I}$ as

$$\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \sigma^2 (\mathbf{X}'\mathbf{X})^{-1}). \tag{11.26}$$

Note that $\hat{\boldsymbol{\beta}}$ is unbiased (since $\mathrm{E}(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$); more details about (11.26) can be found in Appendix C.6. An unbiased estimator of $\sigma^2$ is

$$\hat{\sigma}^2 = \frac{(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})}{n - (p+1)} = \frac{\hat{\mathbf{e}}'\hat{\mathbf{e}}}{n - (p+1)} = \frac{1}{n - (p+1)} \sum_{i=1}^{n} \hat{e}_i^2. \tag{11.27}$$

The errors are estimated from the data as $\hat{\mathbf{e}} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}$ and are called **residuals**.

Before giving a detailed data example, we would like to outline how to draw conclusions about the population of interest from the linear model. As we have seen, both $\boldsymbol{\beta}$ and $\sigma^2$ are unknown in the model and are estimated from the data using (11.25) and (11.27). These are our point estimates. We note that if $\boldsymbol{\beta}_j = 0$, then $\boldsymbol{\beta}_j \mathbf{x_j} = 0$, and then, the model will not contain the term $\boldsymbol{\beta}_j \mathbf{x_j}$. This means that the covariate $\mathbf{x_j}$ does not contribute to explaining the variation in $\mathbf{y}$. Testing the hypothesis $\beta_j = 0$ is therefore equivalent to testing whether $\mathbf{x}_j$ is associated with $\mathbf{y}$ or not in the sense that it helps to explain the variations in $\mathbf{y}$ or not. To test whether the point estimate is different from zero, or whether the deviations of estimates from zero can be explained by random variation, we have the following options:

1. The $100(1-\alpha)\%$ confidence interval for each $\hat{\beta}_j$ is
$$\hat{\beta}_j \pm t_{n-p-1;1-\alpha/2} \cdot \hat{\sigma}_{\hat{\beta}_j} \tag{11.28}$$
with $\hat{\sigma}_{\hat{\beta}} = \sqrt{s_{jj}\hat{\sigma}^2}$ where $s_{jj}$ is the $j$th diagonal element of the matrix $(\mathbf{X}'\mathbf{X})^{-1}$. If the confidence interval does not overlap 0, then we can conclude that $\beta$ is different from 0 and therefore $\mathbf{x_i}$ is associated with $\mathbf{y}$ and it is a relevant variable. If the confidence interval includes 0, we cannot conclude that there is an association between $\mathbf{x_i}$ and $\mathbf{y}$.
2. We can formulate a formal test which tests if $\hat{\beta}_j$ is different from 0. The hypotheses are:
$$H_0 : \beta_j = 0 \quad \text{versus} \quad H_1 : \beta_j \neq 0.$$
The test statistic
$$T = \frac{\hat{\beta}_j}{\sqrt{\hat{\sigma}_{\hat{\beta}_j^2}}}$$
follows a $t_{n-p-1}$ distribution under $H_0$. If $|T| > t_{n-p-1;1-\alpha/2}$, we can reject the null hypothesis (at $\alpha$ level of significance); otherwise, we accept it.
3. The decisions we get from the confidence interval and the $T$-statistic from points 1 and 2 are identical to those obtained by checking whether the $p$-value (see also Sect. 10.2.6) is smaller than $\alpha$, in which case we also reject the null hypothesis.

*Example 11.7.1* Recall Examples 11.2.1, 11.3.2, 11.3.3, 11.4.1, 11.5.1, and 11.6.4 where we analysed the association of exercising time and time of recovery from knee surgery. We can use *R* to obtain a full inductive summary of the linear model:

```
summary(lm(Y~X))
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 100.1244    10.3571   9.667 2.17e-06 ***
X            -1.2153     0.2768  -4.391  0.00135 **
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 11.58 on 10 degrees of freedom
```

Now, we explain the meaning and interpretation of different terms involved in the output.

- Under "Estimate", the parameter estimates are listed and we read that the linear model is fitted as $\mathbf{y}$ (recovery time) $= 100.1244 - 1.2153 \cdot \mathbf{x}$ (exercise time). The subtle differences to Example 11.2.1 are due to rounding of numerical values.
- The variance is estimated as $\hat{\sigma}^2 = 11.58^2$ ("Residual standard error"). This is easily calculated manually using the residual sum of squares: $\sum \hat{e}_i^2/(n-p-1) = 1/10 \cdot \{(90-70.84)^2 + \cdots + (45-62.30)^2\} = 11.58^2$.

- The standard errors $\hat{\sigma}_{\hat{\beta}}$ are listed under "Std. Error". Given that $n - p - 1 = 12 - 1 - 1 = 10$, and therefore $t_{10;0.975} = 2.28$, we can construct a confidence interval for age:
$$-1.22 \pm 2.28 \cdot 0.28 = [-1.86; -0.58]$$
The interval does not include 0, and we therefore conclude that there is an association between exercising time and recovery time. The random error involved in the data is not sufficient to explain the deviation of $\hat{\beta}_i = -1.22$ from 0 (given $\alpha = 0.05$).
- We therefore reject the null hypothesis that $\beta_j = 0$. This can also be seen by comparing the test statistic (listed under "t value" and obtained by $(\hat{\beta}_j - 0)/\sqrt{\hat{\sigma}^2_{\hat{\beta}}} = -1.22/0.277$) with $t_{10,0.975}, |-4.39| > 2.28$. Moreover, $p = 0.001355 < \alpha = 0.05$. We can say that there is a significant association between exercising and recovery time.

Sometimes, one is interested in whether a regression model is useful in the sense that all $\beta_i$'s are different from zero, and we therefore can conclude that there is an association between any $\mathbf{x}_i$ and $\mathbf{y}$. The null hypothesis is

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0$$

and can be tested by the **overall *F*-test**

$$F = \frac{(\hat{\mathbf{y}} - \bar{\mathbf{y}})'(\hat{\mathbf{y}} - \bar{\mathbf{y}})/(p)}{(\mathbf{y} - \hat{\mathbf{y}})'(\mathbf{y} - \hat{\mathbf{y}})/(n - p - 1)} = \frac{n - p - 1}{p} \frac{\sum_i (\hat{y}_i - \bar{y})^2}{\sum_i e_i^2}.$$

The null hypothesis is rejected if $F > F_{1-\alpha;p,n-p-1}$. Note that the null hypothesis in this case tests only the equality of slope parameters and does not include the intercept term.

*Example 11.7.2* In this chapter, we have already explored the associations between branch, operator, and bill with delivery time for the pizza data (Appendix A.4). If we fit a multiple linear model including all of the three variables, we obtain the following results:

```
summary(lm(time~branch+bill+operator))
```

```
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     26.19138    0.78752  33.258  < 2e-16 ***
branchEast      -3.03606    0.42330  -7.172 1.25e-12 ***
branchWest      -0.50339    0.38907  -1.294    0.196
bill             0.21319    0.01535  13.885  < 2e-16 ***
operatorMelissa  0.15973    0.31784   0.503    0.615
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 5.653 on 1261 degrees of freedom
Multiple R-squared:  0.2369,    Adjusted R-squared:  0.2345
F-statistic: 97.87 on 4 and 1261 DF,  p-value: < 2.2e-16
```

By looking at the $p$-values in the last column, one can easily see (without calculating the confidence intervals or evaluating the $t$-statistic) that there is a significant association between the bill and delivery time; also, it is evident that the average delivery time in the branch in the East is significantly different ($\approx 3$ min less) from the central branch (which is the reference category here). However, the estimated difference in delivery times for both the branches in the West and the operator was not found to be significantly different from zero. We conclude that some variables in the model are associated with delivery time, while for others, we could not show the existence of such an association. The last line of the output confirms this: the overall $F$-test has a test statistic of 97.87 which is larger than $F_{1-\alpha;p,n-p-1} = F_{0.95;4,1261} = 2.37$; therefore, the $p$-value is smaller 0.05 ($2.2 \times 10^{-16}$) and the null hypothesis is rejected. The test suggests that there is at least one variable which is associated with delivery time.

### 11.7.1 Properties of Least Squares and Maximum Likelihood Estimators

The least squares estimator of $\boldsymbol{\beta}$ has several properties:

1. The least squares estimator of $\boldsymbol{\beta}$ is *unbiased*, $\mathrm{E}(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$ (see Appendix C.6 for the proof).
2. The estimator $\hat{\sigma}^2$ as introduced in equation (11.27) is also *unbiased*, i.e. $E(\hat{\sigma}^2) = \sigma^2$.
3. The least squares estimator of $\boldsymbol{\beta}$ is *consistent*, i.e. $\hat{\boldsymbol{\beta}}$ converges to $\boldsymbol{\beta}$ as $n$ approaches infinity.
4. The least squares estimator of $\boldsymbol{\beta}$ is *asymptotically normally distributed*.
5. The least squares estimator of $\boldsymbol{\beta}$ has the smallest variance among all linear and unbiased estimators (*best linear unbiased estimator*) of $\boldsymbol{\beta}$.

We do not discuss the technical details of these properties in detail. It is more important to know that the least squares estimator has good features and that is why we choose it for fitting the model. Since we use a "good" estimator, we expect that the model will also be "good". One may ask whether it is possible to use a different estimator. We have already made distributional assumptions in the model: we require the errors to be normally distributed, given that it is indeed possible to apply the maximum likelihood principle to obtain estimates for $\boldsymbol{\beta}$ and $\sigma^2$ in our set-up.

**Theorem 11.7.1** *For the linear model* (11.24)*, the least squares estimator and the maximum likelihood estimator for* $\beta$ *are identical. However, the maximum likelihood estimator of* $\sigma^2$ *is* $\hat{\sigma}^2_{ML} = 1/n(\hat{\mathbf{e}}'\hat{\mathbf{e}})$ *of* $\sigma^2$ *which is a biased estimator of* $\sigma^2$*, but it is asymptotically unbiased.*

How to obtain the maximum likelihood estimator for the linear model is presented in Appendix C.6. The important message of Theorem 11.7.1 is that no matter whether we apply the least squares principle or the maximum likelihood principle, we always

obtain $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$; this does not apply when estimating the variance, but it is an obvious choice to go for the unbiased estimator (11.27) in any given analysis.

### 11.7.2 The ANOVA Table

A table that is frequently shown by software packages when performing regression analysis is the analysis of variance (ANOVA) table. This table can have several meanings and interpretations and may look slightly different depending on the context in which it is used. We focus here on its interpretation i) as a way to summarize the effect of categorical variables and ii) as a hypothesis test to compare $k$ means. This is illustrated in the following example.

*Example 11.7.3* Recall Example 11.6.3 where we established the relationship between branch and delivery time as

$$\text{Delivery Time} = 36.3 - 5.2\text{East} - 1.1\text{West}.$$

The *R* output for the respective linear model is as follows:

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   36.3127     0.2957 122.819  < 2e-16 ***
branchEast    -5.2461     0.4209 -12.463  < 2e-16 ***
branchWest    -1.1182     0.4148  -2.696  0.00711 **
```

We see that the average delivery time of the branches in the East and the Centre (reference category) is different and that the average delivery time of the branches in the West and the Centre is different (because $p < 0.05$). This is useful information, but it does not answer the question if branch as a whole influences the delivery time. It seems that this is the case, but the hypothesis we may have in mind may be

$$H_0 : \mu_{\text{East}} = \mu_{\text{West}} = \mu_{\text{Centre}}$$

which corresponds to

$$H_0 : \beta_{\text{East}} = \beta_{\text{West}} = \beta_{\text{Centre}}$$

in the context of the regression model. These are two identical hypotheses because in the regression set-up, we are essentially comparing three conditional means $\mathrm{E}(Y|X = x_1) = \mathrm{E}(Y|X = x_2) = \mathrm{E}(Y|X = x_3)$. The ANOVA table summarizes the corresponding $F$-Test which tests this hypothesis:

```
m1 <- lm(time~branch)
anova(m1)
```

```
Response: time
           Df Sum Sq Mean Sq F value    Pr(>F)
branch      2   6334  3166.8   86.05 < 2.2e-16 ***
Residuals 1263  46481    36.8
```

We see that the null hypothesis of 3 equal means is rejected because $p$ is close to zero.

What does this table mean more generally? If we deal with linear regression with one (possibly categorical) covariate, the table will be as follows:

| | df | Sum of squares | Mean squares | $F$-statistic |
|---|---|---|---|---|
| Var | $p$ | $SQ_{Reg.}$ | $MSR=SQ_{Reg.}/p$ | MSR/MSE |
| Res | $n-p-1$ | $SQ_{Error}$ | $MSE= SQ_{Error}/(n-p-1)$ | |

The table summarizes the sum of squares regression and residuals (see Sect. 11.3 for the definition), standardizes them by using the appropriate degrees of freedom (df), and uses the corresponding ratio as the $F$-statistic. Note that in the above example, this corresponds to the overall $F$-test introduced earlier. The overall $F$-test tests the hypothesis that any $\beta_j$ is different from zero which is identical to the hypothesis above. Thus, if we fit a linear regression model with one variable, the ANOVA table will yield the same conclusions as the overall $F$-test which we obtain through the main summary. However, if we consider a multiple linear regression model, the ANOVA table may give us more information.

*Example 11.7.4* Suppose we are not only interested in the branch, but also in how the pizza delivery times are affected by operator and driver. We may for example hypothesize that the delivery time is identical for all drivers given they deliver for the same branch and speak to the same operator. In a standard regression output, we would get 4 coefficients for 5 drivers which would help us to compare the average delivery time of each driver to the reference driver; it would however not tell us if on an average, they all have the same delivery time. The overall $F$-test would not help us either because it would test if any $\beta_j$ is different from zero which includes coefficients from branch and operator, not only driver. Using the `anova` command yields the results of the $F$-test of interest:

```
m2 <- lm(time~branch+operator+driver)
anova(m2)
```

```
Response: time
            Df Sum Sq Mean Sq F value    Pr(>F)
branch       2   6334  3166.8 88.6374 < 2.2e-16 ***
operator     1     16    16.3  0.4566    0.4994
driver       4   1519   379.7 10.6282 1.798e-08 ***
Residuals 1258  44946    35.7
```

We see that the null hypothesis of equal delivery times of the drivers is rejected. We can also test other hypotheses with this output: for instance, the null hypothesis of equal delivery times for each operator is not rejected because $p \approx 0.5$.

### 11.7.3 Interactions

It may be possible that the joint effect of some covariates affects the response. For example, drug concentrations may have a different effect on men, woman, and children; a fertilizer could work differently in different geographical areas; or a new education programme may show benefit only with certain teachers. It is fairly simple to target such questions in linear regression analysis by using **interactions**. Interactions are measured as the product of two (or more) variables. If either one or both variables are categorical, then one simply builds products for each dummy variable, thus creating $(k-1)\times(l-1)$ new variables when dealing with two categorical variables (with $k$ and $l$ categories respectively). These product terms are called interactions and estimate how an association of one variable differs with respect to the values of the other variable. Now, we give examples for continuous–categorical, categorical–categorical, and continuous–continuous interactions for two variables $\mathbf{x}_1$ and $\mathbf{x}_2$.

**Categorical–Continuous Interaction.** Suppose one variable $\mathbf{x}_1$ is categorical with $k$ categories, and the other variable $\mathbf{x}_2$ is continuous. Then, $k-1$ new variables have to be created, each consisting of the product of the continuous variable and a dummy variable, $\mathbf{x}_2 \times \mathbf{x}_{1_i}, i \in 1,\ldots,k-1$. These variables are added to the regression model in addition to the *main effects* relating to $\mathbf{x}_1$ and $\mathbf{x}_2$ as follows:

$$\begin{aligned}\mathbf{y} = \beta_0 + \beta_1\mathbf{x_{1_1}} + \cdots + \beta_{k-1}\mathbf{x_{1_{k-1}}} + \beta_k\mathbf{x_2}\\ +\beta_{k+1}\mathbf{x_{1_1}x_2} + \cdots + \beta_p\mathbf{x_{1_{k-1}}x_2} + \mathbf{e}.\end{aligned}$$

It follows that for the reference category of $\mathbf{x}_1$, the effect of $\mathbf{x}_2$ is just $\beta_k$ (because each $\mathbf{x}_2\mathbf{x}_{1_i}$ is zero since each $\mathbf{x}_{1_j}$ is zero). However, the effect for all other categories is $\beta_2+\beta_j$ where $\beta_j$ refers to $\mathbf{x_{1_j}x_2}$. Therefore, the association between $\mathbf{x}_2$ and the outcome $\mathbf{y}$ differs by $\beta_j$ between category $j$ and the reference category. Testing $H_0 : \beta_j = 0$ thus helps to identify whether there is an interaction effect with respect to category $j$ or not.

*Example 11.7.5* Consider again the pizza data described in Appendix A.4. We may be interested in whether the association of delivery time and temperature varies with respect to branch. In addition to time and branch, we therefore need additional interaction variables. Since there are 3 branches, we need $3-1=2$ interaction variables which are essentially the product of (i) time and branch "East" and (ii) time and branch "West". This can be achieved in *R* by using either the "$\star$" operator (which will create both the main and interaction effects) or the ":" operator (which only creates the interaction term).

```
int.model.1 <- lm(temperature~time*branch)
int.model.2 <- lm(temperature~time+branch+time:branch)
summary(int.model.1)
summary(int.model.2)
```

```
Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     70.718327   1.850918  38.207  < 2e-16 ***
time            -0.288011   0.050342  -5.721 1.32e-08 ***
branchEast      10.941411   2.320682   4.715 2.69e-06 ***
branchWest       1.102597   2.566087   0.430  0.66750
time:branchEast -0.195885   0.066897  -2.928  0.00347 **
time:branchWest  0.004352   0.070844   0.061  0.95103
```

The main effects of the model tell us that the temperature is almost 11 degrees higher for the eastern branch compared to the central branch (reference) and about 1 degree higher for the western branch. However, only the former difference is significantly different from 0 (since the $p$-value is smaller than $\alpha = 0.05$). Moreover, the longer the delivery time, the lower the temperature (0.29 degrees for each minute). The parameter estimates related to the interaction imply that this association differs by branch: the estimate is indeed $\beta_{\text{time}} = -0.29$ for the reference branch in the Centre, but the estimate for the branch in the East is $-0.29 - 0.196 = -0.486$ and the estimate for the branch in the West is $-0.29 + 0.004 = -0.294$. However, the latter difference in the association of time and temperature is not significantly different from zero. We therefore conclude that the delivery time and pizza temperature are negatively associated and this is more strongly pronounced in the eastern branch compared to the other two branches. It is also possible to visualize this by means of a separate regression line for each branch, see Fig. 11.7.

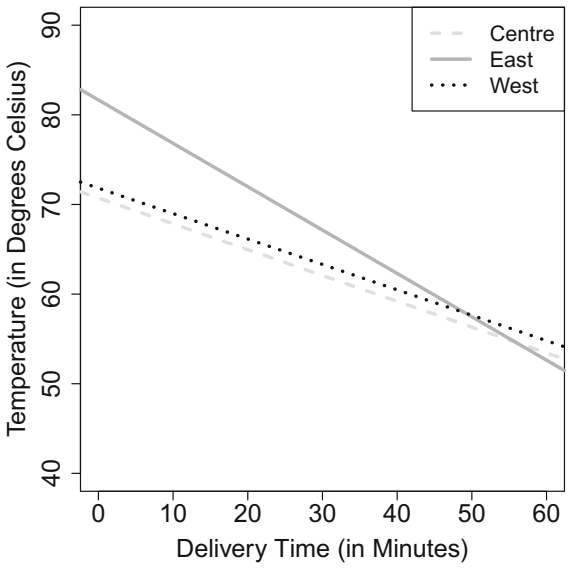



**Fig. 11.7** Interaction of delivery time and branch in Example 11.7.5

- Centre: Temperature $= 70.7 - 0.29 \times$ time,
- East: Temperature $= 70.7 + 10.9 - (0.29 + 0.195) \times$ time,
- West: Temperature $= 70.7 + 1.1 - (0.29 - 0.004) \times$ time.

One can see that the pizzas delivered by the branch in the East are overall hotter but longer delivery times level that benefit off. One might speculate that the delivery boxes from the eastern branch are not properly closed and therefore—despite the overall good performance—the temperature falls more rapidly over time for this branch.

**Categorical–Categorical Interaction.** For two categorical variables $\mathbf{x}_1$ and $\mathbf{x}_2$, with $k$ and $l$ categories, respectively, $(k-1) \times (l-1)$ new dummy variables $\mathbf{x}_{1i} \times \mathbf{x}_{2j}$ need to be created as follows:

$$\begin{aligned}\mathbf{y} = \beta_0 + \beta_1 \mathbf{x_{1_1}} + \cdots + \beta_{k-1} \mathbf{x_{1_{k-1}}} + \beta_k \mathbf{x_{2_1}} + \cdots + \beta_{k+l-2} \mathbf{x_{2_{l-1}}} \\ + \beta_{k+l-1} \mathbf{x_{1_1} x_{2_1}} + \cdots + \beta_p \mathbf{x_{1_{k-1}} x_{2_{l-1}}} + \mathbf{e}.\end{aligned}$$

The interpretation of the regression parameters of interest is less complicated than it looks at first. If the interest is in $\mathbf{x}_1$, then the estimate referring to the category of interest (i.e. $\mathbf{x}_{1_i}$) is interpreted as usual—with the difference that it relates only to the reference category of $\mathbf{x}_2$. The effect of $\mathbf{x}_{1_i}$ may vary with respect to $\mathbf{x}_2$, and the sum of the respective main and interaction effects for each category of $\mathbf{x}_2$ yields the respective estimates. These considerations are explained in more detail in the following example.

*Example 11.7.6* Consider again the pizza data. If we have the hypothesis that the delivery time depends on the operator (who receives the phone calls), but the effect is different for different branches, then a regression model with branch (3 categories, 2 dummy variables), operator (2 categories, one dummy variable), and their interaction (2 new variables) can be used.

```
lm(time~operator*branch)
```

```
Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                36.4203     0.4159  87.567   <2e-16 ***
operatorMelissa            -0.2178     0.5917  -0.368   0.7129
branchEast                 -5.6685     0.5910  -9.591   <2e-16 ***
branchWest                 -1.3599     0.5861  -2.320   0.0205 *
operatorMelissa:branchEast  0.8599     0.8425   1.021   0.3076
operatorMelissa:branchWest  0.4842     0.8300   0.583   0.5598
```

The interaction terms can be interpreted as follows:

- If we are interested in the operator, we see that the delivery time is on average 0.21 min shorter for operator "Melissa". When this operator deals with a branch other than the reference (Centre), the estimate changes to $-0.21 + 0.86 = 0.64$ min longer delivery in the case of branch "East" and $-0.21 + 0.48 = 0.26$ min for branch "West".
- If we are interested in the branches, we observe that the delivery time is shortest for the eastern branch which has on average a 5.66 min shorter delivery time than the central branch. However, this is the estimate for the reference operator only; if operator "Melissa" is in charge, then the difference in delivery times for the two branches is only $-5.66 + 0.86 = -4.8$ min. The same applies when comparing the western branch with the central branch: depending on the operator, the difference between these two branches is estimated as either $-1.36$ or $-1.36 + 0.48 = 0.88$ min, respectively.
- The interaction terms are not significantly different from zero. We therefore conclude that the hypothesis of different delivery times for the two operators, possibly varying by branch, could not be confirmed.

**Continuous–Continuous Interaction.** It is also possible to add an interaction of two continuous variables. This is done by adding the product of these two variables to the model. If $\mathbf{x}_1$ and $\mathbf{x}_2$ are two continuous variables, then $\mathbf{x}_1\mathbf{x}_2$ is the new variable added in the model as an interaction effect as

$$\mathbf{y} = \beta_1\mathbf{x}_1 + \beta_2\mathbf{x}_2 + \beta_3\mathbf{x}_1\mathbf{x}_2 + \mathbf{e}.$$

*Example 11.7.7* If we again consider the pizza data, with pizza temperature as an outcome, we may wish to test for an interaction of bill and time.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 92.555943   2.747414  33.688  < 2e-16 ***
bill        -0.454381   0.068322  -6.651 4.34e-11 ***
time        -0.679537   0.086081  -7.894 6.31e-15 ***
bill:time    0.008687   0.002023   4.294 1.89e-05 ***
```

The *R* output above reveals that there is a significant interaction effect. The interpretation is more difficult here. It is clear that a longer delivery time and a larger bill decrease the pizza's temperature. However, for a large product of bill and time (i.e., when both are large), these associations are less pronounced because the negative coefficients become more and more outweighed by the positive interaction term. On the contrary, for a small bill and short delivery time, the temperature can be expected to be quite high.

**Combining Estimates.** As we have seen in the examples in this section, it can make sense to combine regression estimates when interpreting interaction effects. While it is simple to obtain the point estimates, it is certainly also important to report

their 95 % confidence intervals. As we know from Theorem 7.7.1, the variance of the combination of two random variables can be obtained by $\mathrm{Var}(X \pm Y) = \sigma_{XY}^2 = \mathrm{Var}(X) + \mathrm{Var}(Y) \pm 2\,\mathrm{Cov}(X, Y)$. We can therefore estimate a confidence interval for the combined estimate $\hat{\beta}_i + \hat{\beta}_j$ as

$$(\hat{\beta}_i + \hat{\beta}_j) \pm t_{n-p-1;1-\alpha/2} \cdot \hat{\sigma}_{(\hat{\beta}_i+\hat{\beta}_j)} \tag{11.29}$$

where $\hat{\sigma}_{(\hat{\beta}_i+\hat{\beta}_j)}$ is obtained from the estimated covariance matrix $\widehat{\mathrm{Cov}}(\hat{\boldsymbol{\beta}})$ via

$$\hat{\sigma}_{(\hat{\beta}_i+\hat{\beta}_j)} = \sqrt{\widehat{\mathrm{Var}}(\beta_i) + \widehat{\mathrm{Var}}(\beta_j) + 2\widehat{\mathrm{Cov}}(\beta_i, \beta_j)}.$$

*Example 11.7.8* Recall Example 11.7.5 where we estimated the association between pizza temperature, delivery time, and branch. There was a significant interaction effect for time and branch. Using *R*, we obtain the covariance matrix as follows:

```
mymodel <- lm(temperature~time*branch)
vcov(mymodel)
```

```
            (Int.)       time    East     West time:East time:West
(Int.)      3.4258 -0.09202 -3.4258 -3.4258   0.09202   0.09202
time       -0.0920  0.00253  0.0920  0.0920  -0.00253  -0.00253
branchEast -3.4258  0.09202  5.3855  3.4258  -0.15232  -0.09202
branchWest -3.4258  0.09202  3.4258  6.5848  -0.09202  -0.17946
time:East   0.0920 -0.00253 -0.1523 -0.0920   0.00447   0.00253
time:West   0.0920 -0.00253 -0.0920 -0.1794   0.00253   0.00501
```

The point estimate for the association between time and temperature in the eastern branch is $-0.29 - 0.19$ (see Example 11.7.5). The standard error is calculated as $\sqrt{0.00253 + 0.00447 - 2 \cdot 0.00253} = 0.044$. The confidence interval is therefore $-0.48 \pm 1.96 \cdot 0.044 = [-0.56; -0.39]$.

*Remark 11.7.1* If there is more than one interaction term, then it is generally possible to test whether there is an overall interaction effect, such as $\beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$ in the case of four interaction variables. These tests belong to the general class of "linear hypotheses", and they are not explained in this book. It is also possible to create interactions between more than two variables. However, the interpretation then becomes difficult.

## 11.8 Comparing Different Models

There are many situations where different multiple linear models can be fitted to a given data set, but it is unclear which is the best one. This relates to the question of which variables should be included in a model and which ones can be removed.

For example, when we modelled the association of recovery time and exercising time by using different transformations, it remained unclear which of the proposed transformations was the best. In the pizza delivery example, we have about ten covariates: Does it make sense to include all of them in the model or do we understand the associations in the data better by restricting ourselves to the few most important variables? Is it necessary to include interactions or not?

There are various model selection criteria to compare the quality of different fitted models. They can be useful in many situations, but there are also a couple of limitations which make model selection a difficult and challenging task.

**Coefficient of Determination ($R^2$).** A possible criterion to check the goodness of fit of a model is the coefficient of determination ($R^2$). We discussed its development, interpretation, and philosophy in Sect. 11.3 (see p. 256 for more details).

**Adjusted $R^2$.** Although $R^2$ is a reasonable measure for the goodness of fit of a model, it also has some limitations. One limitation is that if the number of covariates increases, $R^2$ increases too (we omit the theorem and the proof for this statement). The added variables may not be relevant, but $R^2$ will still increase, wrongly indicating an improvement in the fitted model. This criterion is therefore not a good measure for model selection. An adjusted version of $R^2$ is defined as

$$R^2_{adj} = 1 - \frac{SQ_{\text{Error}}/(n-p-1)}{SQ_{\text{Total}}/(n-1)}. \tag{11.30}$$

It can be used to compare the goodness of fit of models with a different number of covariates. Please note that both $R^2$ and $R^2_{adj}$ are only defined when there is an intercept term in the model (which we assume throughout the chapter).

*Example 11.8.1* In Fig. 11.6, the association of exercising time and recovery time after knee surgery is modelled linearly, quadratically, and cubically; in Fig. 11.5, this association is modelled by means of a square-root transformation. The model summary in *R* returns both $R^2$ (under "Multiple R-squared") and the adjusted $R^2$ (under "adjusted R-squared"). The results are as follows:

| | $R^2$ | $R^2_{adj}$ |
|---|---|---|
| Linear | 0.6584 | 0.6243 |
| Quadratic | 0.6787 | 0.6074 |
| Cubic | 0.7151 | 0.6083 |
| Square root | 0.6694 | 0.6363 |

It can be seen that $R^2$ is larger for the models with more variables; i.e. the cubic model (which includes three variables) has the largest $R^2$. The adjusted $R^2$, which takes the different model sizes into account, favours the model with the square-root transformation. This model provides therefore the best fit to the data among the four models considered using $R^2$.

**Akaike's Information Criterion** (AIC) is another criterion for model selection. The AIC is based on likelihood methodology and has its origins in information theory. We do not give a detailed motivation of this criterion here. It can be written for the linear model as

$$\text{AIC} = n \log\left(\frac{SQ_{\text{Error}}}{n}\right) + 2(p+1). \tag{11.31}$$

The smaller the AIC, the better the model. The AIC takes not only the fit to the data via $SQ_{\text{Error}}$ into account but also the parsimony of the model via the term $2(p+1)$. It is in this sense a more mature criterion than $R^2_{adj}$ which considers only the fit to the data via $SQ_{\text{Error}}$. Akaike's Information Criterion can be calculated in *R* via the `extractAIC()` command. There are also other commands, but the results differ slightly because the different formulae use different constant terms. However, no matter what formula is used, the results regarding the model choice are always the same.

*Example 11.8.2* Consider again Example 11.8.1. $R^2_{adj}$ preferred the model which includes exercising time via a square-root transformation over the other options. The AIC value for the model where exercise time is modelled linearly is 60.59, when modelled quadratically 61.84, when modelled cubically 62.4, and 60.19 for a square-root transformation. Thus, in line with $R^2_{adj}$, the AIC also prefers the square-root transformation.

**Backward selection.** Two models, which differ only by one variable $\mathbf{x}_j$, can be compared by simply looking at the test result for $\beta_j = 0$: if the null hypothesis is rejected, the variable is kept in the model; otherwise, the other model is chosen. If there are more than two models, then it is better to consider a systematic approach to comparing them. For example, suppose we have 10 potentially relevant variables and we are not sure which of them to include in the final model. There are $2^{10} = 1024$ possible different combinations of variables and in turn so many choices of models! The inclusion or deletion of variables can be done systematically with various procedures, for example with **backward selection** (also known as **backward elimination**) as follows:

1. Start with the full model which contains all variables of interest, $\Upsilon = \{\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_p}\}$.
2. Remove the variable $\mathbf{x_i} \in \Upsilon$ which optimizes a criterion, i.e. leads to the smallest AIC (the highest $R^2_{adj}$, the highest test statistic, or the smallest significance) among all possible choices.
3. Repeat step 2. until a stop criterion is fulfilled, i.e. until no improvement regarding AIC, $R^2$, or the $p$-value can be achieved.

There are several other approaches. Instead of moving in the backward direction, we can also move in the forward direction. **Forward selection** starts with a model with no covariates and adds variables as long as the model continues to improve with respect to a particular criterion. The **stepwise selection** approach combines forward selection and backward selection: either one does backward selection and checks in between whether adding variables yields improvement, or one does forward selection and continuously checks whether removing the already added variables improves the model.

*Example 11.8.3* Consider the pizza data: if delivery time is the outcome, and branch, bill, operator, driver, temperature, and number of ordered pizzas are potential covariates, we may decide to include only the relevant variables in a final model. Using the `stepAIC` function of the `library(MASS)` allows us implementing backward selection with *R*.

```
library(MASS)
ms <- lm(time~branch+bill+operator+driver
+temperature+pizzas)
stepAIC(ms, direction='back')
```

At the first step, the *R* output states that the AIC of the full model is 4277.56. Then, the AIC values are displayed when variables are removed: 4275.9 if operator is removed, 4279.2 if driver is removed, and so on. One can see that the AIC is minimized if operator is excluded.

```
Start:  AIC=4277.56
time ~ branch+bill+operator+driver+temperature+pizzas

              Df Sum of Sq   RSS    AIC
- operator     1      9.16 36508 4275.9
<none>                     36499 4277.6
- driver       4    279.71 36779 4279.2
- branch       2    532.42 37032 4291.9
- pizzas       1    701.57 37201 4299.7
- temperature  1   1931.50 38431 4340.8
- bill         1   2244.28 38743 4351.1
```

Now *R* fits the model without operator. The AIC is 4275.88. Excluding further variables does not improve the model with respect to the AIC.

```
Step:  AIC=4275.88
time ~ branch + bill + driver + temperature + pizzas

               Df Sum of Sq   RSS    AIC
<none>                      36508 4275.9
- driver        4    288.45 36797 4277.8
- branch        2    534.67 37043 4290.3
- pizzas        1    705.53 37214 4298.1
- temperature   1   1923.92 38432 4338.9
- bill          1   2249.60 38758 4349.6
```

We therefore conclude that the final "best" model includes all variables considered, except operator. Using stepwise selection (with option `both` instead of `back`) yields the same results. We could now interpret the summary of the chosen model.

**Limitations and further practical considerations.** The above discussions of variable selection approaches makes it clear that there are a variety of options to compare fitted models, but there is no unique best way to do so. Different criteria are based on different philosophies and may lead to different outcomes. There are a few considerations that should however always be taken into account:

- If categorical variables are included in the model, then all respective dummy variables need to be considered as a whole: all dummy variables of this variable should either be removed or kept in the model. For example, a variable with 3 categories is represented by two dummy variables in a regression model. In the process of model selection, both of these variables should either be kept in the model or removed, but not just one of them.
- A similar consideration refers to interactions. If an interaction is added to the model, then the main effects should be kept in the model as well. This enables straightforward interpretations as outlined in Sect. 11.7.3.
- The same applies to polynomials. If adding a cubic transformation of $\mathbf{x_i}$, the squared transformation as well as the linear effect should be kept in the model.
- When applying backward, forward, or stepwise regression techniques, the results may vary considerably depending on the method chosen! This may give different models for the same data set. Depending on the strategy used to add and remove variables, the choice of the criterion (e.g. AIC versus $p$-values), and the detailed set-up (e.g. if the strategy is to add/remove a variable if the $p$-value is smaller than $\alpha$, then the choice of $\alpha$ is important), one may obtain different results.
- All the above considerations show that model selection is a non-trivial task which should always be complemented by substantial knowledge of the subject matter. If there are not too many variables to choose from, simply reporting the full model can be meaningful too.

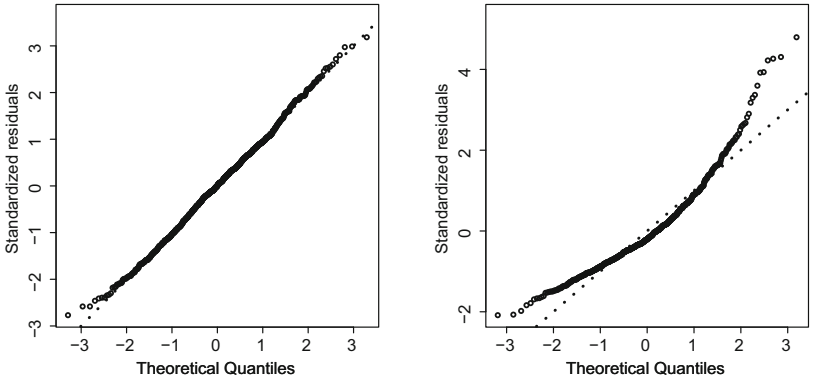
## 11.9 Checking Model Assumptions

The assumptions made for the linear model mostly relate to the error terms: $\mathbf{e} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$. This implies a couple of things: i) the difference between $\mathbf{y}$ and $\mathbf{X}\boldsymbol{\beta}$ (which is $\mathbf{e}$) needs to follow a normal distribution, ii) the variance for each error $e_i$ is constant as $\sigma^2$ and therefore does not depend on $i$, and iii) there is no correlation between the error terms, $\text{Cov}(e_i, e_{i'}) = 0$ for all $i, i'$. We also assume that $\mathbf{X}\boldsymbol{\beta}$ is adequate in the sense that we included all relevant variables and interactions in the model and that the estimator $\hat{\boldsymbol{\beta}}$ is stable and adequate due to influential observations or highly correlated variables. Now, we demonstrate how to check assumptions i) and ii).

**Normality assumption.** The errors $\mathbf{e}$ are assumed to follow a normal distribution. We never know the errors themselves. The only possibility is to estimate the errors by means of the residuals $\hat{\mathbf{e}} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}$. To ensure that the estimated residuals are comparable to the scale of a *standard* normal distribution, they can be standardized via

$$\hat{e}_i^* = \frac{\hat{e}_i}{\hat{\sigma}^2\sqrt{1 - P_{ii}}}$$

where $P_{ii}$ is the $i$th diagonal element of the hat matrix $P = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$. Generally, the estimated **standardized residuals** are used to check the normality assumption. To obtain a first impression about the error distribution, one can simply plot a histogram of the standardized residuals. The normality assumption is checked with a QQ-plot where the theoretical quantiles of a standard normal distribution are plotted against the quantiles from the standardized residuals. If the data approximately matches the bisecting line, then we have evidence for fulfilment of the normality assumption (Fig. 11.8a), otherwise we do not (Fig. 11.8b).




(a) Normality without problems (b) Normality with some problems

**Fig. 11.8** Checking the normality assumption with a QQ-plot

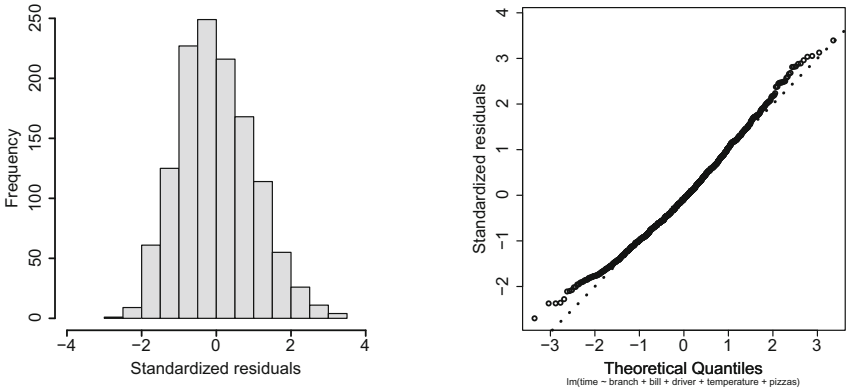
Small deviations from the normality assumption are not a major problem as the least squares estimator of $\boldsymbol{\beta}$ remains unbiased. However, confidence intervals and tests of hypothesis rely on this assumption, and particularly for small sample sizes and/or stronger violations of the normality assumptions, these intervals and conclusions from tests of hypothesis no longer remain valid. In some cases, these problems can be fixed by transforming $\mathbf{y}$.

**Heteroscedasticity.** If errors have a constant variance, we say that the errors are homoscedastic and we call this phenomenon **homoscedasticity**. When the variance of errors is not constant, we say that the errors are heteroscedastic. This phenomenon is also known as **heteroscedasticity**. If the variance $\sigma^2$ depends on $i$, then the variability of the $e_i$ will be different for different groups of observations. For example, the daily expenditure on food ($\mathbf{y}$) may vary more among persons with a high income ($\mathbf{x_1}$), so fitting a linear model yields stronger variability of the $e_i$ among higher income groups. Plotting the fitted values $\hat{y}_i$ (or alternatively $x_i$) against the standardized residuals (or a transformation thereof) can help to detect whether or not problems exist. If there is no pattern in the plot (random plot), then there is likely no violation of the assumption (see Fig. 11.10a). However, if there is a systematic trend, i.e. higher/lower variability for higher/lower $\hat{y}_i$, then this indicates heteroscedasticity (Fig. 11.10b, trumpet plot). The consequences are again that confidence intervals and tests may no longer be correct; again, in some situations, a transformation of $\mathbf{y}$ can help.

*Example 11.9.1* Recall Example 11.8.3 where we identified a good model to describe the delivery time for the pizza data. Branch, bill, temperature, number of pizzas ordered, and driver were found to be associated with delivery time. To explore whether the normality assumption is fulfilled for this model, we can create a histogram for the standardized residuals (using the *R* function `rstandard()`). A QQ-plot is contained in the various model diagnostic plots of the `plot()` function.

```
fm <- lm(time~branch+bill+driver+temperature
+pizzas)
hist(rstandard(fm))
plot(fm, which=2)
```

Figure 11.9 shows a reasonably symmetric distribution of the residuals, maybe with a slightly longer tail to the right. As expected for a standard normal distribution, not many observations are larger than 2 or smaller than $-2$ (which are close to the 2.5 and 97.5 % quantiles). The QQ-plot also points towards a normal error distribution



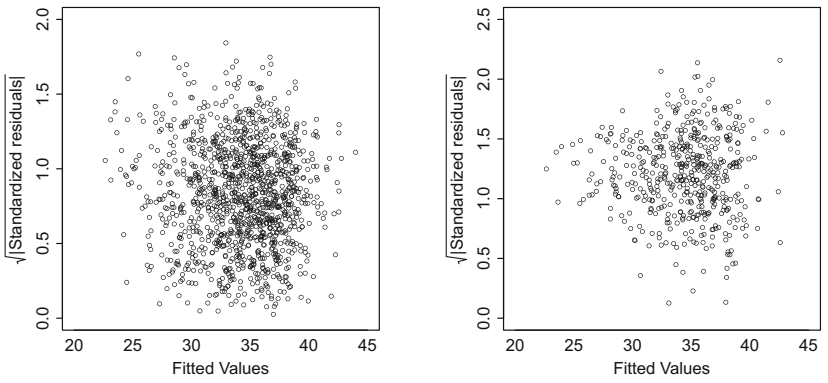

(a) Via the histogram of the standardized residuals

(b) Via a QQ-plot of the standardized residuals

**Fig. 11.9** Checking the normality assumption in Example 11.9.1

because the observed quantiles lie approximately on the bisecting line. It seems that the lowest residuals deviate a bit from the expected normal distribution, but not extremely. The normality assumption does not seem to be violated.

Plotting the fitted values $\hat{y}_i$ against the square root of the absolute values of the standardized residuals ($= \sqrt{|\hat{e}_i^*|}$, used by *R* for stability reasons) yields a plot with no visible structure (see Fig. 11.10a). There is no indication of heteroscedasticity.




(a) Example 11.9.1: chaos plot

(b) Violation: $\sigma^2$ depends on $i$

**Fig. 11.10** Checking the heteroscedasticity assumption

```
plot(fm, which=3)
```

Figure 11.10b shows an artificial example where heteroscedasticity occurs: the higher the fitted values, the higher is the variation of residuals. This is also called a **trumpet plot**.

## 11.10 Association Versus Causation

There is a difference between association and causation: association says that higher values of $X$ relate to higher values of $Y$ (or vice versa), whereas causation says that *because* values of $X$ are higher, values of $Y$ are higher too. Regression analysis, as introduced in this chapter, establishes associations between $X_i$'s and $Y$, not causation, unless strong assumptions are made.

For example, recall the pizza delivery example from Appendix A.4. We have seen in several examples and exercises that the delivery time varies by driver. In fact, in a multiple linear regression model where $Y$ is the delivery time, and the covariates are branch, bill, operator, driver, and temperature, we get significant differences of the mean delivery times of some drivers (e.g. Domenico and Bruno). Does this mean that *because* Domenico is the driver, the pizzas are delivered faster? Not necessarily. If Domenico drives the fastest scooter, then this may be the cause of his speedy deliveries. However, we have not measured the variable "type of scooter" and thus cannot take it into account. The result we obtain from the regression model is still useful because it provides the manager with useful hypotheses and predictions about his delivery times, but the interpretation that the driver's abilities cause shorter delivery times may not necessarily be appropriate.

This example highlights that one of the most important assumptions to interpret regression parameters in a causal way is to have measured (and used) all variables $X_i$ which affect both the outcome $Y$ and the variable of interest $A$ (e.g. the variable "driver" above). Another assumption is that the relationship between all $X_i$'s and $Y$ is modelled correctly, for example non-linear if appropriate. Moreover, we need some *a priori* assumptions about how the variables relate to each other, and some technical assumptions need to be met as well. The interested reader is referred to Hernan and Robins (2017) for more details.

## 11.11 Key Points and Further Issues

**Note:**

✓ If $X$ is a continuous variable, the parameter $\beta$ in the linear regression model $Y = \alpha + \beta X + e$ can be interpreted as follows:

"For each unit of increase in $X$, there is an increase of $\beta$ units in $\mathrm{E}(Y)$."

✓ If $X$ can take a value of zero, then it can be said that

$$\text{"For } X = 0, \mathrm{E}(Y) \text{ equals } \alpha\text{."}$$

If $X = 0$ is out of the data range or implausible, it does not make sense to interpret $\alpha$.

✓ A model is said to be linear when it is linear in its parameters.

✓ It is possible to include many different covariates in the regression model: they can be binary, log-transformed, or take any other form, and it is still a *linear* model. The interpretation of these variables is always conditional on the other variables; i.e. the reported association assumes that all other variables in the model are held fixed.

✓ To evaluate whether $\mathbf{x_j}$ is associated with $\mathbf{y}$, it is useful to test whether $\beta_j = 0$. If the null hypothesis $H_0 : \beta_j = 0$ is rejected (i.e. the $p$-value is smaller than $\alpha$), one says that there is a *significant* association of $\mathbf{x_j}$ and $\mathbf{y}$. However, note that a non-significant result does not necessarily mean there is no association, it just means we could not show an association.

✓ An important assumption in the multivariate linear model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$ relates to the errors:

$$\mathbf{e} \sim N(\mathbf{0}, \sigma^2\mathbf{I}).$$

To check whether the errors are (approximately) normally distributed, one can plot the standardized residuals in a histogram or a QQ-plot. The heteroscedasticity assumption (i.e. $\sigma^2$ does not depend on $i$) is tested by plotting the fitted values ($\hat{y}_i$) against the standardized residuals ($\hat{e}_i$). The plot should show no pattern (random plot).

✓ Different models can be compared by systematic hypothesis testing, $R^2_{adj}$, AIC, and other methods. Different methods may yield different results, and there is no unique best option.

## 11.12 Exercises

*Exercise 11.1* The body mass index (BMI) and the systolic blood pressure of 6 people were measured to study a cardiovascular disease. The data are as follows:

| Body mass index | 26 | 23 | 27 | 28 | 24 | 25 |
|---|---|---|---|---|---|---|
| Systolic blood pressure | 170 | 150 | 160 | 175 | 155 | 150 |

(a) The research hypothesis is that a high BMI relates to a high blood pressure. Estimate the linear model where blood pressure is the outcome and BMI is the covariate. Interpret the coefficients.
(b) Calculate $R^2$ to judge the goodness of fit of the model.

*Exercise 11.2* A psychologist speculates that spending a lot of time on the internet has a negative effect on children's sleep. Consider the following data on hours of deep sleep ($Y$) and hours spent on the internet ($X$) where $x_i$ and $y_i$ are the observations on internet time and deep sleep time of the $i$th ($i = 1, 2, \ldots, 9$) child respectively:

| Child $i$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|
| Internet time $x_i$ (in h) | 0.3 | 2.2 | 0.5 | 0.7 | 1.0 | 1.8 | 3.0 | 0.2 | 2.3 |
| Sleep time $y_i$ (in h) | 5.8 | 4.4 | 6.5 | 5.8 | 5.6 | 5.0 | 4.8 | 6.0 | 6.1 |

(a) Estimate the linear regression model for the given data and interpret the coefficients.
(b) Calculate $R^2$ to judge the goodness of fit of the model.
(c) Reproduce the results of a) and b) in *R* and plot the regression line.
(d) Now assume that we only distinguish between spending more than 1 hour on the internet ($X = 1$) and spending less than (or equal to) one hour on the internet ($X = 0$). Estimate the linear model again and compare the results. How can $\hat{\beta}$ now be interpreted? Describe how $\hat{\beta}$ changes if those who spend more than one hour on the internet are coded as 0 and the others as 1.

*Exercise 11.3* Consider the following data on weight and height of 17 female students:

| Student $i$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|
| Weight $y_i$ | 68 | 58 | 53 | 60 | 59 | 60 | 55 | 62 | 58 |
| Height $x_i$ | 174 | 164 | 164 | 165 | 170 | 168 | 167 | 166 | 160 |
| Student $i$ | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | |
| Weight $y$ | 53 | 53 | 50 | 64 | 77 | 60 | 63 | 69 | |
| Height $x$ | 160 | 163 | 157 | 168 | 179 | 170 | 168 | 170 | |

(a) Calculate the correlation coefficient of Bravais–Pearson (use $\sum_{i=1}^{n} x_i y_i = 170,821$, $\bar{x} = 166.65$, $\bar{y} = 60.12$, $\sum_{i=1}^{n} y_i^2 = 62,184$, $\sum_{i=1}^{n} x_i^2 = 472,569$). What does this imply with respect to a linear regression of height on weight?

(b) Now estimate and interpret the linear regression model where "weight" is the outcome.
(c) Predict the weight of a student with a height 175 cm.
(d) Now produce a scatter plot of the data (manually or by using $R$) and interpret it.
(e) Add the following two points to the scatter plot $(x_{18},\ y_{18}) = (175,\ 55)$ and $(x_{19},\ y_{19}) = (150,\ 75)$. Speculate how the linear regression estimate will change after adding these points.
(f) Re-estimate the model using all 19 observations and $\sum x_i y_i = 191,696$ and $\sum x_i^2 = 525,694$.
(g) Given the results of the two regression models: What are the general implications with respect to the least squares estimator of $\beta$?

*Exercise 11.4* To study the association of the monthly average temperature (in °C, $X$) and hotel occupation (in %, $Y$), we consider data from three cities: Polenca (Mallorca, Spain) as a summer holiday destination, Davos (Switzerland) as a winter skiing destination, and Basel (Switzerland) as a business destination.

| Month | Davos | | Polenca | | Basel | |
|---|---|---|---|---|---|---|
| | $X$ | $Y$ | $X$ | $Y$ | $X$ | $Y$ |
| Jan | −6 | 91 | 10 | 13 | 1 | 23 |
| Feb | −5 | 89 | 10 | 21 | 0 | 82 |
| Mar | 2 | 76 | 14 | 42 | 5 | 40 |
| Apr | 4 | 52 | 17 | 64 | 9 | 45 |
| May | 7 | 42 | 22 | 79 | 14 | 39 |
| Jun | 15 | 36 | 24 | 81 | 20 | 43 |
| Jul | 17 | 37 | 26 | 86 | 23 | 50 |
| Aug | 19 | 39 | 27 | 92 | 24 | 95 |
| Sep | 13 | 26 | 22 | 36 | 21 | 64 |
| Oct | 9 | 27 | 19 | 23 | 14 | 78 |
| Nov | 4 | 68 | 14 | 13 | 9 | 9 |
| Dec | 0 | 92 | 12 | 41 | 4 | 12 |

(a) Interpret the following regression model output where the outcome is "hotel occupation" and "temperature" is the covariate.

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 50.33459    7.81792   6.438 2.34e-07 ***
X            0.07717    0.51966   0.149    0.883
```

(b) Interpret the following output where "city" is treated as a covariate and "hotel occupation" is the outcome.

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  48.3333     7.9457   6.083 7.56e-07 ***
cityDavos     7.9167    11.2369   0.705    0.486
cityPolenca   0.9167    11.2369   0.082    0.935
```

(c) Interpret the following output and compare it with the output from b):

```
Analysis of Variance Table
Response: Y
          Df  Sum Sq Mean Sq F value Pr(>F)
city       2   450.1  225.03   0.297  0.745
Residuals 33 25001.2  757.61
```

(d) In the following multiple linear regression model, both "city" and "temperature" are treated as covariates. How can the coefficients be interpreted?

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  44.1731    10.9949   4.018 0.000333 ***
X             0.3467     0.6258   0.554 0.583453
cityDavos     9.7946    11.8520   0.826 0.414692
cityPolenca  -1.1924    11.9780  -0.100 0.921326
```

(e) Now consider the regression model for hotel occupation and temperature fitted separately for each city: How can the results be interpreted and what are the implications with respect to the models estimated in (a)–(d)? How can the models be improved?

```
Davos:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  73.9397     4.9462  14.949 3.61e-08 ***
X            -2.6870     0.4806  -5.591 0.000231 ***

Polenca:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -22.6469    16.7849  -1.349  0.20701
X             3.9759     0.8831   4.502  0.00114 **

Basel:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   32.574     13.245   2.459   0.0337 *
X              1.313      0.910   1.443   0.1796
```

(f) Describe what the design matrix will look like if city, temperature, and the interaction between them are included in a regression model.

(g) If the model described in (f) is fitted the output is as follows:

```
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    32.5741    10.0657   3.236 0.002950 **
X               1.3133     0.6916   1.899 0.067230 .
cityDavos      41.3656    12.4993   3.309 0.002439 **
cityPolenca   -55.2210    21.0616  -2.622 0.013603 *
X:cityDavos    -4.0003     0.9984  -4.007 0.000375 ***
X:cityPolenca   2.6626     1.1941   2.230 0.033388 *
```

Interpret the results.

(h) Summarize the association of temperature and hotel occupation by city—including 95 % confidence intervals—using the interaction model. The covariance matrix is as follows:

| | (Int.) | X | Davos | Polenca | X:Davos | X:Polenca |
|---|---|---|---|---|---|---|
| (Int.) | 101.31 | -5.73 | -101.31 | -101.31 | 5.73 | 5.73 |
| X | -5.73 | 0.47 | 5.73 | 5.73 | -0.47 | -0.47 |
| Davos | -101.31 | 5.73 | 156.23 | 101.31 | -9.15 | -5.73 |
| Polenca | -101.31 | 5.73 | 101.31 | 443.59 | -5.73 | -22.87 |
| X:Davos | 5.73 | -0.47 | -9.15 | -5.73 | 0.99 | 0.47 |
| X:Polenca | 5.73 | -0.47 | -5.73 | -22.87 | 0.47 | 1.42 |

*Exercise 11.5* The theatre data (see Appendix A.4) describes the monthly expenditure on theatre visits of 699 residents of a Swiss city. It captures not only the expenditure on theatre visits (in SFR) but also age, gender, yearly income (in 1000 SFR), and expenditure on cultural activities in general as well as expenditure on theatre visits in the preceding year.

(a) The summary of the multiple linear model where expenditure on theatre visits is the outcome is as follows:

| | Estimate | Std. Error | t value | Pr(>\|t\|) | |
|---|---|---|---|---|---|
| (Intercept) | -127.22271 | 19.15459 | -6.642 | 6.26e-11 | *** |
| Age | 0.39757 | 0.19689 | [1] | [2] | |
| Sex | 22.22059 | 5.22693 | 4.251 | 2.42e-05 | *** |
| Income | 1.34817 | 0.20947 | 6.436 | 2.29e-10 | *** |
| Culture | 0.53664 | 0.05053 | 10.620 | <2e-16 | *** |
| Theatre_ly | 0.17191 | 0.11711 | 1.468 | 0.1426 | |

How can the missing values [1] and [2] be calculated?

(b) Interpret the model diagnostics in Fig. 11.11.

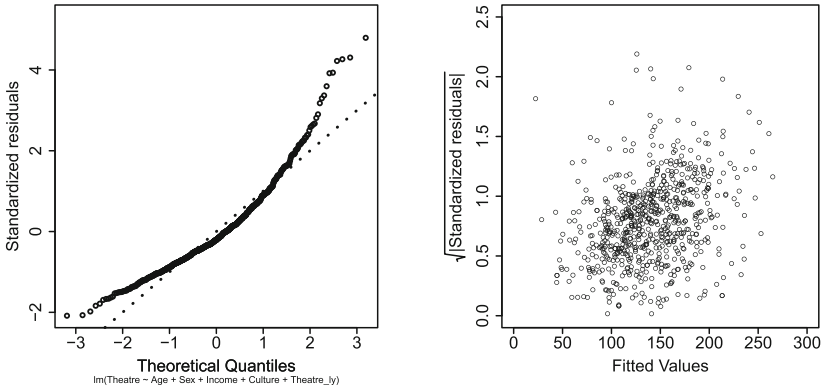
(c) Given the diagnostics in (b), how can the model be improved? Plot a histogram of theatre expenditure in *R* if you need further insight.

(d) Consider the model where theatre expenditure is log-transformed:

Coefficients:

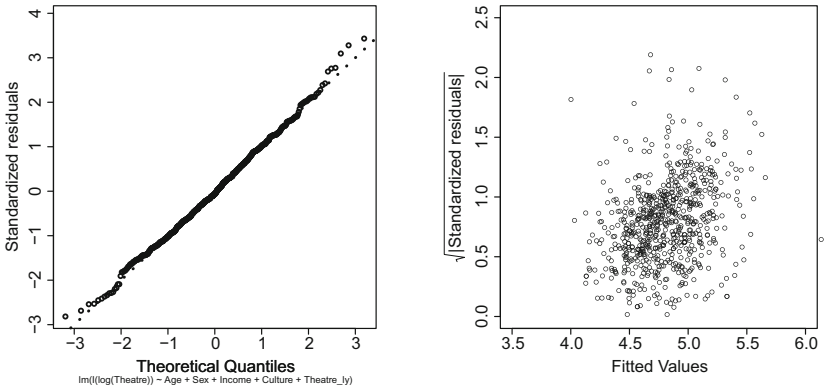
| | Estimate | Std. Error | t value | Pr(>\|t\|) | |
|---|---|---|---|---|---|
| (Intercept) | 2.9541546 | 0.1266802 | 23.320 | < 2e-16 | *** |
| Age | 0.0038690 | 0.0013022 | 2.971 | 0.00307 | ** |
| Sex | 0.1794468 | 0.0345687 | 5.191 | 2.75e-07 | *** |
| Income | 0.0087906 | 0.0013853 | 6.346 | 4.00e-10 | *** |
| Culture | 0.0035360 | 0.0003342 | 10.581 | < 2e-16 | *** |
| Theatre_ly | 0.0013492 | 0.0007745 | 1.742 | 0.08197 | . |

How can the coefficients be interpreted?




(a) QQ-plot (b) Fitted values versus residuals

**Fig. 11.11** Checking the model assumptions




(a) QQ-plot (b) Fitted values vs. residuals

**Fig. 11.12** Checking the model assumptions

(e) Judge the quality of the model from d) by means of Figs. 11.12a and 11.12b. What do they look like compared with those from b)?

*Exercise 11.6* Consider the pizza delivery data described in Appendix A.4.

(a) Read the data into *R*. Fit a multiple linear regression model with delivery time as the outcome and temperature, branch, day, operator, driver, bill, number of ordered pizzas, and discount customer as covariates. Give a summary of the coefficients.

(b) Use $R$ to calculate the 95 % confidence intervals of all coefficients. Hint: the standard errors of the coefficients can be accessed either via the covariance matrix or the model summary.
(c) Reproduce the least squares estimate of $\sigma^2$. Hint: use `residuals()` to obtain the residuals.
(d) Now use $R$ to estimate both $R^2$ and $R^2_{adj}$. Compare the results with the model output from a).
(e) Use backward selection by means of the `stepAIC` function from the library `MASS` to find the best model according to AIC.
(f) Obtain $R^2_{adj}$ from the model identified in e) and compare it to the full model from a).
(g) Identify whether the model assumptions are satisfied or not.
(h) Are all variables from the model in (e) causing the delivery time to be either delayed or improved?
(i) Test whether it is useful to add a quadratic polynomial of temperature to the model.
(j) Use the model identified in (e) to predict the delivery time of the last captured delivery (i.e. number 1266). Use the `predict()` command to ease the calculation of the prediction.

→ Solutions of all exercises in this chapter can be found on p. 409

*Source* Toutenburg, H., Heumann, C., *Deskriptive Statistik*, 7th edition, 2009, Springer, Heidelberg