Module 4: Residual analysis

4.1 Introduction

The principle of least squares provides a general methodology for fitting straight-line models to regression data. So far, we have fitted such models to any data for which scatterplots between the response variable and the explanatory variables displayed anything resembling straight-line relationships. But we have made no further effort to check the validity of the assumptions of the models. For a multiple linear regression model

$$Y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \cdots + \beta_k x_{i,k} + \varepsilon_i, \quad i = 1, \ldots n,$$

we make the following four **model assumptions**:

(I) **Independence:** The response variables Y_i are independent.

(II) **Normality:** The response variables Y_i are normally distributed.

(III) **Homoscedasticity:** The response variables Y_i all have the same variance σ^2. (The term *homoscedasticity* is from Greek and means 'same variance'.)

(IV) **Linearity:** The true relationship between the mean of the response variable $\mathbb{E}[Y]$ and the explanatory variables $x_1, \ldots, x_k$ is a straight line.

Assumption (I) on independence of the response variables is subject to the design of the study and the way the data have been collected. In this course, we shall assume that all data have been collected independently; that is, we shall assume that Assumption (I) is satisfied.

In order to check the model assumptions, we shall need a new type of residuals: *standardised residuals*. These are introduced in Section 4.2. The remaining sections are concerned with methods for assessing the appropriateness of the model: Section 4.3 concerns the normality assumption, Section 4.4 the homoscedasticity and linearity assumptions, and Section 4.5 the linearity assumption in multiple regression. The module concludes with Section 4.6 which considers situations where a few points differ from the rest of the data.

4.2 Residuals

Rather than checking Assumptions (II)–(IV) on the response variables directly, it is convenient to re-express the assumptions in terms of the random errors

$$\varepsilon_i = Y_i - (\beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \cdots + \beta_k x_{i,k}), \quad i = 1, \ldots, n, \tag{4.1}$$

and check the assumptions on the random errors instead.

The following four assumptions on the random errors are equivalent to the assumptions on the response variables.

(i) The random errors ε_i are independent.

(ii) The random errors ε_i are normally distributed.

(iii) The random errors ε_i have constant variance σ^2.

(iv) The random errors ε_i have zero mean.

If assumptions (i)-(iv) are satisfied, the random errors ε_i are independent, identically distributed random variables with distributions:

$$\varepsilon_i \sim N(0, \sigma^2), \quad i = 1, \ldots, n.$$

Thus, the random errors ε_i can be regarded as a random sample from a $N(0, \sigma^2)$ distribution. We can check the assumptions on the random errors (and thereby the assumptions on the response variables) by analysing an observed sample of the random errors. All we need are observations of the random errors.

The obvious candidates for observations of the random errors are the fitted residuals: the differences between the observed values $y_1, y_2, \ldots, y_n$ of Y, and the values $\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_n$ fitted by the model, where

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i,1} + \hat{\beta}_2 x_{i,2} + \cdots + \hat{\beta}_k x_{i,k}, \quad i = 1, 2, \ldots, n, \tag{4.2}$$

with $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$ denoting the least squares estimates of the regression parameters. That is, the fitted residuals are given by

$$\begin{aligned} \hat{\varepsilon}_i &= y_i - \hat{y}_i \\ &= y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i,1} - \hat{\beta}_2 x_{i,2} - \cdots - \hat{\beta}_k x_{i,k}. \end{aligned}$$

However, as we shall see in Subsection 4.2.1, these residuals are observations of random variables–known as *raw residuals*–which are not independent, and which do not have the same variance. In Subsection 4.2.2, the raw residuals are transformed into *standardised residuals*, for which the issue of non-constant variance is overcome.

4.2.1 Raw residuals

The observed values r_i of the raw residuals are given by the fitted residuals

$$r_i = \hat{\varepsilon}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i,1} - \hat{\beta}_2 x_{i,2} - \cdots - \hat{\beta}_k x_{i,k}, \quad i = 1, \ldots, n,$$

where $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$ are the least squares estimates of the regression parameters. The corresponding random variables, denoted by R_i, are obtained by substituting the observed y_is with the random variables Y_i, and the least squares estimates of $\beta_0, \beta_1, \ldots, \beta_k$ with the corresponding random variables: the least squares estimators. That is, the **raw residuals** are given by

$$R_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i,1} - \hat{\beta}_2 x_{i,2} - \cdots - \hat{\beta}_k x_{i,k}, \quad i = 1, \ldots, n, \tag{4.3}$$

where $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$ are the least squares estimators of the regression parameters.

It can be shown (we shall not do it here) that the ith raw residual R_i has the distribution

$$R_i \sim N\left(0, (1 - h_{ii}) \times \sigma^2\right), \quad i = 1, \ldots, n, \tag{4.4}$$

where h_{ii} is the ith diagonal element of the **hat-matrix h** given by

$$\begin{aligned} \mathbf{h} &= \begin{pmatrix} h_{11} & h_{12} & \cdots & h_{1n} \\ h_{21} & h_{22} & & \vdots \\ \vdots & & \ddots & \vdots \\ h_{n1} & \cdots & \cdots & h_{nn} \end{pmatrix} \\ &= \mathbf{x}\left(\mathbf{x}^T\mathbf{x}\right)^{-1}\mathbf{x}^T, \end{aligned} \tag{4.5}$$

where $\mathbf{x}$ is the design matrix

$$\mathbf{x} = \begin{pmatrix} 1 & x_{1,1} & x_{1,2} & \cdots & x_{1,k} \\ 1 & x_{2,1} & x_{2,2} & \cdots & x_{2,k} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n,1} & x_{n,2} & \cdots & x_{n,k} \end{pmatrix}.$$

The matrix $\mathbf{h}$ is called the *hat-matrix*, because it has the property that it 'puts a hat on the ys', in the sense that the fitted values $\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_n$ in (4.2) are found by matrix-multiplying the hat-matrix on the vector of observed values $y_1, y_2, \ldots, y_n$:

$$\mathbf{hy} = \mathbf{h} \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \mathbf{x} \left(\mathbf{x}^T \mathbf{x}\right)^{-1} \mathbf{x}^T \mathbf{y} = \hat{\mathbf{y}} = \begin{pmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_n \end{pmatrix}.$$

Here $\mathbf{y}$ denotes the column vector of response variables, as defined in Module 3.

You can see from (4.4) that the raw residuals have different variances. Also, notice that none of the raw residuals have the variance we are looking for: σ^2. It can be shown that all the diagonal elements h_{ii} take values between 0 and 1: if h_{ii} is small, the variance $(1 - h_{ii}) \times \sigma^2$ is close to 'right' variance σ^2; however, if h_{ii} is close to one, the variance $(1 - h_{ii}) \times \sigma^2$ is much smaller than σ^2.

A further problem with the raw residuals is that they are not independent. However, it can be shown that if the values of the diagonal elements h_{ii} of the hat-matrix $\mathbf{h}$ are reasonably small, the raw residuals are 'nearly' independent. We shall not go into further details with this problem.

In summary, the raw residuals are not suitable for checking the assumptions on the random errors. The random errors all have the same variance–the raw residuals have different variances; the random errors have variance σ^2–in general, none of the raw residuals have variance σ^2; the random errors are independent–the raw residuals are not.

4.2.2 Standardised residuals

The standardised residuals are designed to overcome the problem of different variances of the raw residuals. The problem is solved by dividing each of the raw residuals by an appropriate term.

Recall that the ith raw residual R_i has a $N(0, (1 - h_{ii}) \times \sigma^2)$-distribution. A standard result on the normal distribution states that if $X \sim N(\mu, \sigma^2)$, then

$$aX \sim N(a\mu, a^2\sigma^2).$$

Therefore, if we multiply R_i by $a_i = 1/\sqrt{1 - h_{ii}}$, we get the **standardised residual,** S_i, with distribution

$$S_i = \frac{R_i}{\sqrt{1 - h_{ii}}} \sim N\left(\frac{0}{\sqrt{1 - h_{ii}}}, \frac{(1 - h_{ii}) \times \sigma^2}{1 - h_{ii}}\right) = N\left(0,\, \sigma^2\right).$$

That is, the standardised residuals $S_1, \ldots, S_n$ are random variables with distributions

$$S_i \sim N\left(0,\ \sigma^2\right), \quad i = 1, \ldots, n. \tag{4.6}$$

The observed value s_i of the ith standardised residual is given by

$$s_i = \frac{r_i}{\sqrt{1-h_{ii}}}. \tag{4.7}$$

The standardisation of the residuals has taken care of the issue of different variances, but nothing has changed with regard to dependence between the residuals. It can be shown that the dependence between the standardised residuals is exactly the same as the dependence between the raw residuals. We shall not go into further details with this problem.

In summary, the standardised residuals are better suited than the raw residuals for checking the assumptions on the random errors. The standardised residuals S_i have the same distributions as the random errors: $N\left(0,\sigma^2\right)$. However, the standardised residuals are not, in general, independent. But if the values of the diagonal elements h_{ii} of the hat-matrix $\mathbf{h}$ are reasonably small, the standardised residuals are 'nearly' independent.

Note that most statistical computer packages (including SAS) calculate the standardised residuals slightly differently from the standardised residuals defined in the module. In most packages, each of the standardised residuals is divided by an estimate of the standard error, to obtain variables which are approximately $N\left(0,1\right)$-distributed, rather than $N\left(0,\sigma^2\right)$-distributed. However, since all the residuals are divided by the same value, the patterns in residual plots and normal probability plots are identical whether one uses the un-scaled version in (4.7) or the scaled version.

4.3 Normality

The first assumption we consider is Assumption (ii): *the random errors ε_i are normally distributed*. Since the random errors can be regarded as a random sample from a $N(0,\sigma^2)$ distribution, we can check Assumption (ii) by checking whether the standardised residuals s_i might have come from a normal distribution. A **normal probability plot** of the standardised residuals will give an indication of whether or not the assumption of normality of the random errors is appropriate. Recall that a normal probability plot is found by plotting the quantiles of the observed sample against the corresponding quantiles of a standard normal distribution $N(0,1)$. If the normal probability plot shows a straight line, it is reasonable to assume that the observed sample comes from a normal distribution. If, on the other hand, the points deviate from a straight line, there is statistical evidence against the assumption that the random errors are an independent sample from a normal distribution.

Example 4.1 *Holiday cottages*
Recall from Module 3 the data on sales prices, ages and livable areas of holiday cottages in Odsherred, Denmark. It was suggested, in Module 3, that a multiple linear regression model might describe the variation in the data well. The least squares line for the relationship between sales price (Y), age (x_1), and livable area (x_2), is given by

$$\hat{y} = -281.43 - 7.611x_1 + 19.01x_2.$$

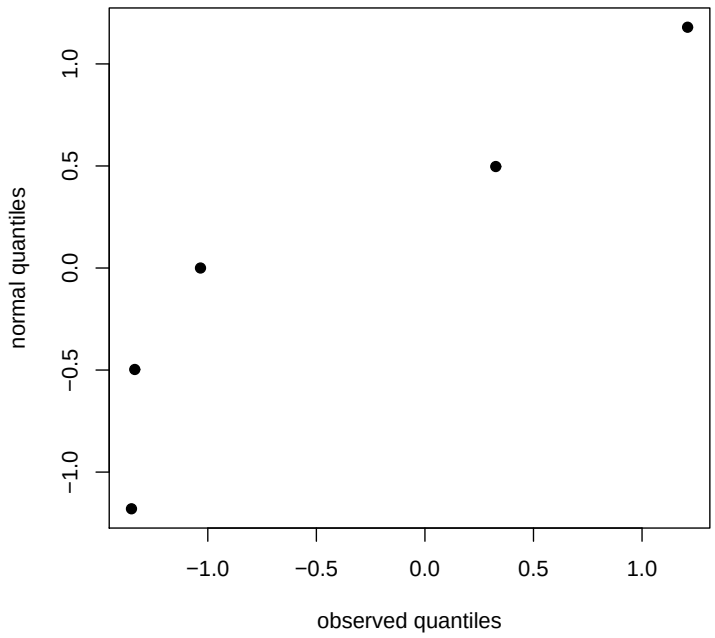

Figure 4.1: Normal probability plot of standardised residuals for Odsherred data

Figure 4.1 shows a normal probability plot of the residuals There are very few data points, so one should be careful in concluding too much from the plot. Nevertheless, the points deviate quite a bit from a straight line, so the normality assumption might not be satisfied for these data.

Further details on this dataset can be found here.

♢

Example 4.2 *Ice cream consumption*

In Module 3, we considered how the ice cream consumption (Y) is related to temperature (x_1), ice cream price (x_2), average annual family income (x_3), and the year (x_4). In Module 3, a possible outlier was removed from the dataset before we fitted a multiple linear regression model to the data. In this module, we consider the full dataset–including the outlying point. The least squares line, relating the ice cream consumption to the four explanatory variables, is given by

$$\hat{y} = 0.714 + 0.00315\,x_1 - 1.29\,x_2 - 0.00237\,x_3 + 0.0508x_4.$$

A normal probability plot of the standardised residuals is shown in Figure 4.2. The nor-

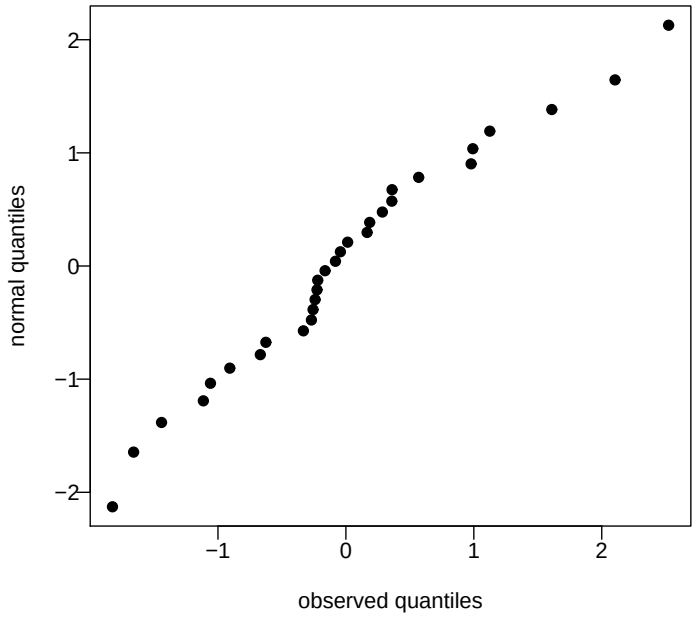

Figure 4.2: Normal probability plot of standardised residuals for ice cream data

mal probability plot is not too far from a straight line. (Although the line is not entirely convincing.) It seems that the normality assumption might be satisfied for these data.

Further details on this dataset can be found here.

♢

The two most common ways to deal with failure of the normality assumption are either to transform the data into a new set of data for which the assumption is satisfied (transforming data is discussed in Module 6), or to use a distribution different from the normal. A general framework to dealing with non-normal (and/or non-linear) models is that of *generalised linear models*. Generalised linear models are studied in ST112.

Note that, it can affect the normal probability plot if one or more of the other assumptions are broken, for instance, if the response variables are dependent, or if the variances of the response variables differ.

4.4 Homoscedasticity and linearity

The two assumptions Assumption (iii): *the random errors ε_i have constant variation*, and Assumption (iv): *the random errors ε_i have zero mean*, can be checked at the same time. To

do this, we use a *residual plot*. A **residual plot** is a scatterplot of the standardised residuals s_i against the fitted values $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i,1} + \hat{\beta}_2 x_{i,2} + \cdots + \hat{\beta}_k x_{i,k}$. Recall that the (standardised) residuals are the deviations of the observations away from the fitted values. If Assumptions (iii) and (iv) are satisfied we would expect the residuals to vary randomly around zero and we would expect the spread of the residuals to be about the same throughout the plot.

Example 4.2(continued) *Ice cream consumption*

A residual plot for the data on the relationship between ice cream consumption and temperature, ice cream price, average annual family income, and the year is shown in Figure 4.3. The points in the plot seem to be fluctuating randomly around zero in an un-patterned

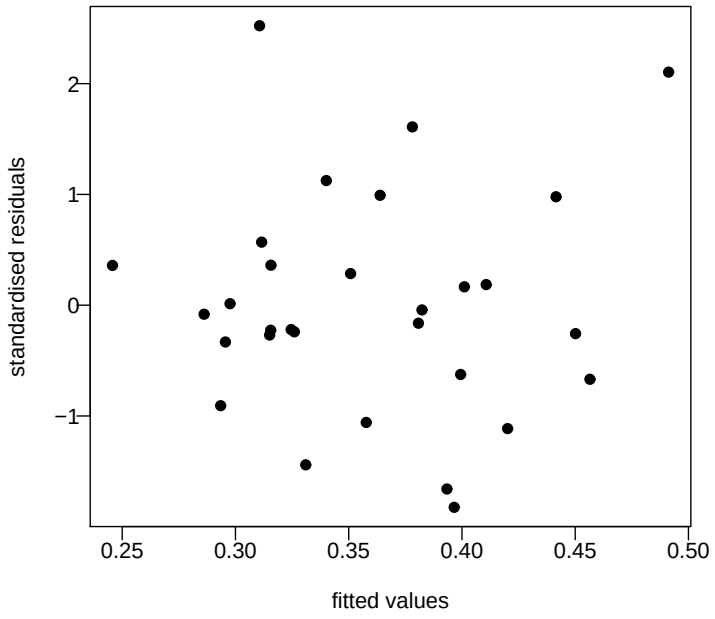

Figure 4.3: Residual plot for the ice cream data

fashion. Thus, the plot does not suggest violations of the assumptions of zero means and constant variance of the random errors.

♢

In general, any systematic pattern in a residual plot suggests that one or more of Assumptions (i)–(iv) are violated. Since we have assumed independence of the random errors, and since a normal probability plot is better for assessing the assumption of normality, we shall concentrate on breaches of Assumptions (iii) and (iv). When looking for patterns in

residual plots, there are three main features which are important. If the residuals seem to increase or decrease in average magnitude with the fitted values, it is an indication that the variance of the residuals is not constant. That is, Assumption (iii) is broken. If the points in the plot lie on a curve around zero, rather than fluctuating randomly, it is an indication that Assumption (iv) is broken. If a few points in the plot lie a long way from the rest of the points, they might be outliers, that is, data points for which the model is not appropriate. (Outliers are considered further in Section 4.6.) Figure 4.4 illustrates the most important features to look for in a residual plot. Figure 4.4(a) shows a residual plot with no systematic

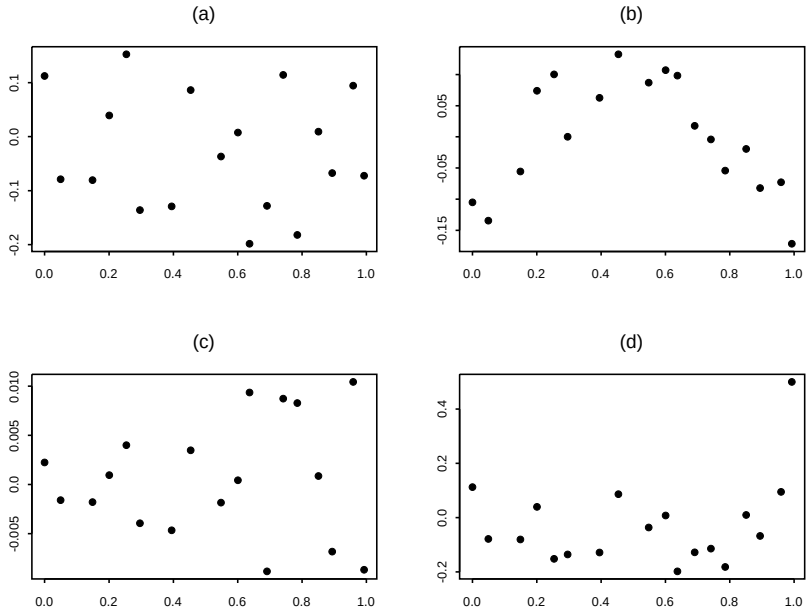

Figure 4.4: Different features in residual plots

pattern. It seems that Assumptions (iii) and (iv) are satisfied for the data associated with this residual plot. In Figure 4.4(b) there is a clear curved pattern: Assumption (iv) may be broken. In Figure 4.4(c) the random variation of the residuals increases as the fitted values increase. This pattern indicates that the variance σ^2 is not constant. Finally, in Figure 4.4(d) most of the residuals are randomly scattered around 0, but one observation has produced a residual which is much larger than any of the other residuals. The point may be an outlier.

In Module 6, we shall consider ways to analyse data for which Assumption (iii) and/or Assumption (iv) are broken.

Example 4.3 *Wind power*

In Module 1, we considered a study into how the direct current output from a wind power generator changes with wind speed. A scatterplot of the data is reproduced in Figure 4.5.

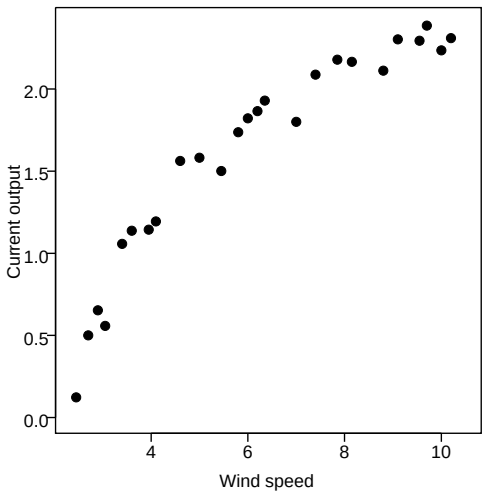

Figure 4.5: Direct current outpur against wind speed

The data points seem to lie along a slightly curved line, but it is not too far from a straight line, so perhaps a simple linear regression model might be a reasonable model for the data after all. The least squares line for the data is given by

$$\hat{y} = 0.131 + 0.241\ x.$$

Figure 4.6 shows (a) a residual plot and (b) a normal probability plot for the data. The normal probability plot in Figure 4.6 (b) is not very convincing: the residuals appear to come from a skew distribution. However, it is the residual plot in Figure 4.6(a) that provides the strongest argument against using a simple linear regression model for these data. There is a very clear pattern in the residual plot: the residuals go from being negative to positive and then negative again. Thus, it seems that Assumption (iv) is broken. In Module 6, we shall return to this example and find a better model for the data.

Further details on this dataset can be found here.

♢

In the case of simple models (with only one explanatory variable), a residual plot is useful for assessing both the assumption on constant variance of the response variables, and the assumption that the relationship between the response variable and the explanatory variable is a straight line. For example, in Example 4.3 the residual plot gives a very explicit indication of how the model assumptions are broken: the relationship between wind speed and current output is not a straight line–it is curved. However, when there are more than one explanatory variable in the model, the residual plot is less informative regarding the linearity assumption.

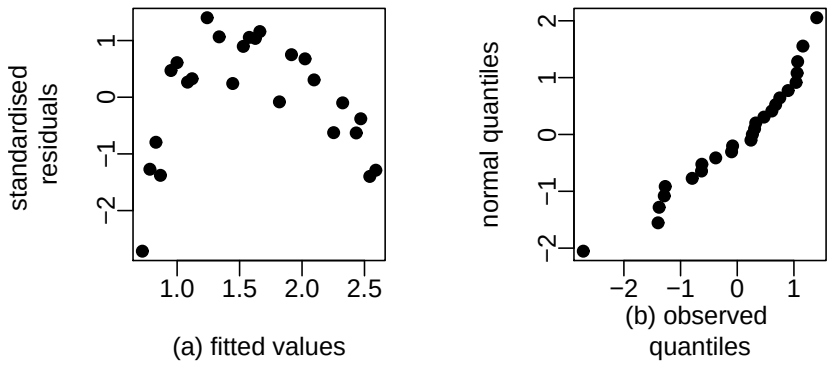

Figure 4.6: Wind power data: (a) residual plot, (b) probability plot.

For instance, although the scatterplot for the ice cream data in Figure 4.3 does not indicate violations of the assumption that the mean response is of the form $\mathbb{E}[Y] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$, it is still possible that one or two of the explanatory variables enter the true relationship in a non-linear fashion. In general, when there are several explanatory variables, a non-linear relationship between the response and one (or more) of the explanatory variables can easily be concealed in a residual plot–in particular if the explanatory variables are correlated. In order to check whether each of the explanatory variables enters the model linearly, we need a different type of plot: *partial residual plots*. These are discussed in the next section.

4.5 Linearity in multiple regression

In a multiple linear regression model, it is assumed that each of the explanatory variables $x_1, \ldots, x_k$ affects the mean of the response in a linear way. That is, we assume that

$$\mathbb{E}[Y] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k. \tag{4.8}$$

How can we check this assumption? An obvious suggestion would be to look for straight-line relationships in scatterplots of the observed response variables against each of the explanatory variables, one at the time.

Example 4.2(continued) *Ice cream consumption*

Scatterplots of the ice cream consumption against the four explanatory variables temperature, ice cream price, average annual family income, and the year are displayed in Figure 4.7. There seems to be straight-line relationships between the ice cream consumption and

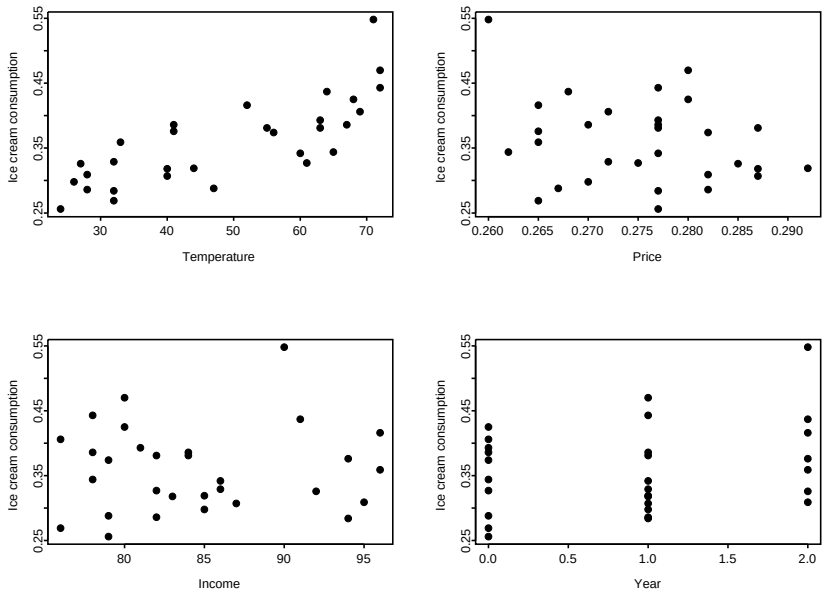

Figure 4.7: Scatterplots of the ice cream consumption against the four explanatory variables

temperature, and between the ice cream consumption and the year. The two remaining plots (against price and income) have a lot of scatter. The relationships might be straight-line, but the plots are hardly convincing.

♢

When we investigate the scatterplots, we essentially consider each of the simple models

$$\begin{aligned} \mathbb{E}[Y] &= \beta_0 + \beta_1 x_1, \\ \mathbb{E}[Y] &= \beta_0 + \beta_2 x_2, \\ &\vdots \\ \mathbb{E}[Y] &= \beta_0 + \beta_k x_k, \end{aligned}$$

separately. For example, that the relationship between ice cream consumption and temperature (ignoring all other variables) is a straight line, and the relationship between ice cream consumption and average annual income (ignoring all other variables) is a straight line, *etc.*

But the assumption we wish to check is (4.8), rather than each of the simple models. That is, for each $l = 1, \ldots, k$ we wish to check whether x_l enters the model linearly, *taking all the other variables into account*. If all the explanatory variables are uncorrelated, there is no difference between checking (4.8) and checking the simple models separately. However, it is usually the case that the explanatory variables are correlated. For instance, in Example 4.2 it is likely that the average annual income will increase from one year to the next; thus the variables 'income' and 'year' are likely to be correlated.

The idea behind the method for checking whether x_l enters linearly in (4.8), taking all the other variables into account, is the following. We want to know how x_l affects the response variable, Y, if all the other explanatory variables $x_1, \ldots, x_{l-1}, x_{l+1}, \ldots, x_k$ affect the response variable linearly. That is, we consider the following form of the response variables

$$Y_i \approx \beta_0 + \beta_1 x_{i,1} + \cdots + \beta_{l-1} x_{i,l-1} + p_l\left(x_{i,l}\right) + \beta_{l+1} x_{i,l+1} + \cdots + \beta_k x_{i,k},$$

for some function $p_l(\cdot)$ which we wish to determine. (If we can show that $p_l(\cdot)$ is linear, the assumption is satisfied for x_l.) Since the true regression parameters $\beta_0, \beta_1, \ldots, \beta_{l-1}, \beta_{l+1}, \ldots, \beta_k$ are unknown, we substitute by the least squares estimators $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_{l-1}, \hat{\beta}_{l+1}, \ldots, \hat{\beta}_k$, obtaining

$$Y_i \approx \hat{\beta}_0 + \hat{\beta}_1 x_{i,1} + \cdots + \hat{\beta}_{l-1} x_{i,l-1} + p_l\left(x_{i,l}\right) + \hat{\beta}_{l+1} x_{i,l+1} + \cdots + \hat{\beta}_k x_{i,k}. \tag{4.9}$$

The next step is to use the definition of the raw residual R_i in (4.3): $R_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i,1} - \hat{\beta}_2 x_{i,2} - \cdots - \hat{\beta}_k x_{i,k}$. We can rewrite this as

$$Y_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i,1} + \cdots + \hat{\beta}_{l-1} x_{i,l-1} + \hat{\beta}_l x_{i,l} + \hat{\beta}_{l+1} x_{i,l+1} + \cdots + \hat{\beta}_k x_{i,k} + R_i.$$

If we substitute this expression for Y_i into (4.9) and cancel out, we get

$$p_l\left(x_{i,l}\right) \approx \hat{\beta}_l x_{i,l} + R_i.$$

That is, the true function $p_l(\cdot)$ for how x_l affects Y is approximately equal to

$$p_l\left(x_{i,l}\right) \approx \hat{\beta}_l x_{i,l} + R_i = P_{i,l}, \quad i = 1, \ldots, n. \tag{4.10}$$

The terms $P_{i,l}$, $i = 1, \ldots, n$, are called the lth **partial residuals**. (Note that, for each explanatory variable, we get a new set of partial residuals: the 1st partial residuals refer to x_1, the 2nd to x_2, *etc.*) The partial residuals are random variables since both the least squares estimator $\hat{\beta}_l$ and the raw residuals R_i are random variables. Observations of the partial residuals are given by

$$p_{i,l} = \hat{\beta}_l x_{i,l} + r_i, \quad i = 1, \ldots, n,$$

where $\hat{\beta}_l$ is the least squares estimate of β_l.

We know from (4.10) that, for each $i = 1, \ldots, n$, we have that $p_l(x_{i,l}) \approx P_{i,l}$. Thus, if we plot the values of the lth explanatory variable, $x_{1,l}, x_{2,l}, \ldots, x_{n,l}$, against the observed lth partial residuals $p_{1,l}, p_{2,l}, \ldots, p_{n,l}$, the plot will indicate the true function $p_l(\cdot)$. This plot is called the lth **partial residual plot**. (Note that we get a different plot for each explanatory variable: the 1st partial residual plot refers to x_1, the 2nd to x_2, *etc.*) If the partial residual plot shows a straight line, it is an indication that the true relationship between the response variable and the lth explanatory variable x_l is straight-line, when all other variables are taken into account. If the plot shows a non-linear relationship, it is an indication that x_l affects the response variable in a non-linear fashion.

Example 4.2(continued) *Ice cream consumption*

Figure 4.8 shows partial residual plots for each of the four explanatory variables in the ice cream data. The four plots indicate clear relationships between the ice cream consumption

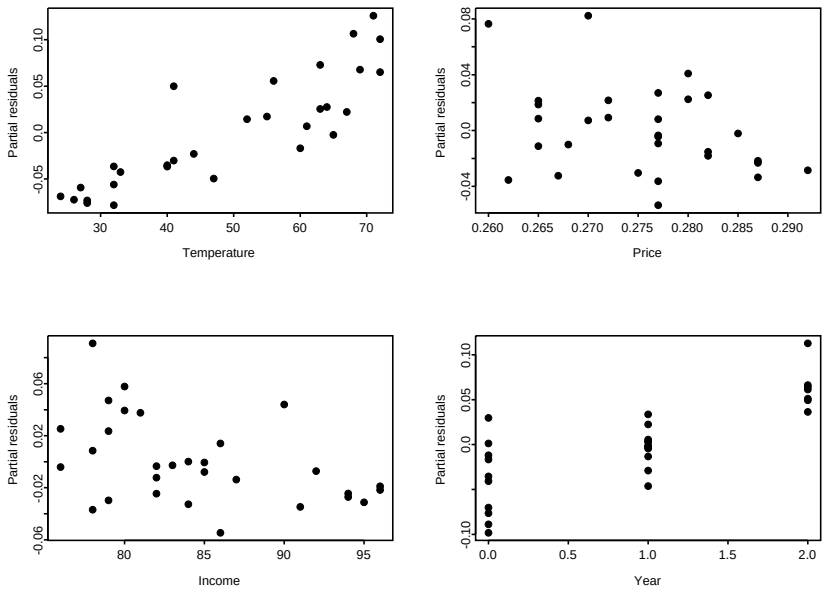

Figure 4.8: Partial residual plots for the ice cream data

and each of the explanatory variables. The relationships seem to be more-or-less straight-line, although there is some indication of possible slight curves in the plots against temperature and income. Also, in the plot against temperature, a single point appears to deviate from the trend of the rest of the points. This point could be an outlier.

You can see that the plots in Figure 4.8 are quite different from the scatterplots in Figure 4.7. This is because the partial residual plots take the other variables into account.

♢

4.6 Outliers and leverage points

This section concerns situations where one or a few observations are different–in some way–from the rest of the data. We distinguish between two ways a few points may differ from the remaining points. A data point might lie far from the general trend in the rest of the data: such a point is called an outlier. Outliers are discussed in Subsection 4.5.1.

Sometimes, a statistical analysis is very sensitive to a single (or a few) data point(s), in the sense that if the value of this point is changed even slightly, the outcome of the analysis alters greatly. Such points are called leverage points, and are discussed in Subsection 4.5.2.

4.6.1 Outliers

An **outlier** is an observation which differs from the main trend in the data. The reason might be due to (unforeseen) special circumstances about the particular observation (for example, imagine that one of the holiday cottages in Example 4.1 was designed by a famous architect–adding extra value to the sales price), or it might be due to a measurement error. But there is also the possibility that the unusual observation is simply due to random variation in the data: since the data are observations of random variables, there will be some variation away from the true relationship. Most points will lie closely around the true relationship, some will lie a little away, and a few might lie a bit further away.

Suppose that a point lies a bit away from the main trend in the data, and that we wish assess whether this is due to random variation in the data, or whether the observation actually differs–in some way–from the rest of the data. There are various methods for doing this; here we shall use *studentised residuals*.

The idea behind this method is as follows. If a data point lies far from the general trend in the data, it is equivalent to the point having a large (raw) residual. Thus, we can re-phrase the issue of whether a point lies too far from the main trend to have happened by chance, into an issue of whether the corresponding residual is too large to have happened by chance. We know from (4.4) that the ith raw residual R_i has a normal distribution with zero mean and variance $(1 - h_{ii}) \times \sigma^2$; so, in order to check whether the observed value r_i is too large to have happened by chance, we can compare r_i to the distribution of R_i: $N\left(0, (1 - h_{ii}) \times \sigma^2\right)$. This is a basic statistical problem: we have a normal distribution with unknown variance (since σ^2 is unknown), and we wish to test whether or not the observation r_i might have come from this distribution. To do this, we use a t-test. The t-statistic is given by

$$T_i = \frac{R_i}{\sqrt{(1 - h_{ii})\,\tilde{\sigma}_i^2}} = \frac{S_i}{\sqrt{\tilde{\sigma}_i^2}},$$

where $\tilde{\sigma}_i^2$ is an appropriate estimate of the variance of the standardised residual S_i. It can be shown that an appropriate unbiased estimate is given by $\tilde{\sigma}_i^2 = \left((n - k)\,S^2 - S_i^2\right) / (n - k - 1)$,

where $S^2 = \sum_{i=1}^{n} \left(Y_i - \hat{Y}\right)^2 / (n-k)$. The test statistic T_i has a $t\,(n-k-1)$-distribution. That is,

$$T_i = S_i \sqrt{\frac{n-k-1}{(n-k)\,S^2 - S_i^2}} \sim t\,(n-k-1)\,, \quad i = 1, \ldots, n, . \tag{4.11}$$

The variables T_i are called **studentised residuals** (because they are t-distributed; or, more precisely, *Student's* t-distributed). If the numerical value $|t_i|$ of a studentised residual is (much) larger than the rest, it is an indication that the corresponding observation y_i may be an outlier.

There is no fixed value (or quantile) for which a point is an outlier if it exceeds this value (quantile). If the model is correct, we expect around 5% of the studentised residuals to lie outside the interval between the 2.5%- and 97.5%-quantiles of a $t\,(n-k-1)$-distribution, 1% to lie outside the interval between the 0.5%- and 99.5%-quantiles, and so on. For example, an observation with a studentised residual corresponding to the 99.9%-quantile may be an outlier, if the dataset only contains 20 observations, but it is not an outlier in a dataset of 1000 observations. We would expect around 0.1% of the residuals to exceed the 99.9%-quantile; in a dataset of 20 observations, this corresponds to 0.02 observations out of the 20–it is very unlikely, that an extreme residual like this would have occurred by chance. However, if the dataset contains 1000 observations, we would expect 1 observation (0.1% of 1000) to have a residual exceeding the 99.9%-quantile. Hence, the point is not an outlier–in fact, it would be suspicious if there were no residuals around or beyond the 99.9%-quantile!

When a possible outlier is detected, one should always try and find out if there is a reason why this point may be different from the rest. For example, is the particular measurement taken by a different person, or on a different day/in a different place, or does the particular subject differ in some way from the rest? In the example on ice cream sales, one data point lies away from the trend–could this be because the particular period coincided with the summer holidays? Or because there was a fun fair in the town? Or ...? Or could it simply be a misprint? If you have collected the data yourself, or have access to additional information about the data collection, you might be able to avoid the outlier (*e.g.* by correcting a misprint, or introducing an extra explanatory variable). In this course, however, we cannot investigate the background of outlying points, as there is no further information available on the collection and validation of the datasets that are used.

In situations where no explanation can be found to why a point is outlying, one has to decide whether to leave the corresponding observation in the dataset, or whether to omit the observation, when the data are analysed. (Alternatively, it is sometimes possible to transform the data in such a way that outlying points are pulled closer towards the general trend in the transformed data. Transforming data is discussed in Module 6.) Whether an unexplained outlier should be left in or omitted from the dataset depends both on its extremity and on its leverage. The next subsection concerns leverage, and how to check for outliers and leverage points in a diagnostic plot.

Note that sometimes studentised residuals are also used for checking normality of the random errors (Assumption (ii)). But since the T_is in (4.11) are t-distributed rather than

normally distributed, this is not strictly correct. (In order to assess this assumption using the studentised residuals, the quantiles of the observed sample $t_1, t_2, \ldots, t_n$ should be plotted against the corresponding quantiles of an appropriate t-distribution.) However, if the dataset is sufficiently large, the t-distribution is very close to a normal distribution, and the quantiles are almost identical. Thus, for large datasets, one can use a normal probability plot as a good approximation to a t-distribution probability plot.

4.6.2 Leverage points

A **leverage point** is a point for which the observed value of this particular point has a great influence on the analysis. An illustration of a leverage point is shown in Figure 4.9: suppose you have a cluster of data with x-values not too far apart; also, you have one observation corresponding to an x-value further away. The value of this isolated point is disproportionately influential on the least squares line: one might say that it works as a *lever*–if the value of this observation is changed, the least squares line changes considerably (as illustrated in the figure). In contrast, if the value of one of the points within the cluster is changed, the least squares line will not be affected to the same extent.

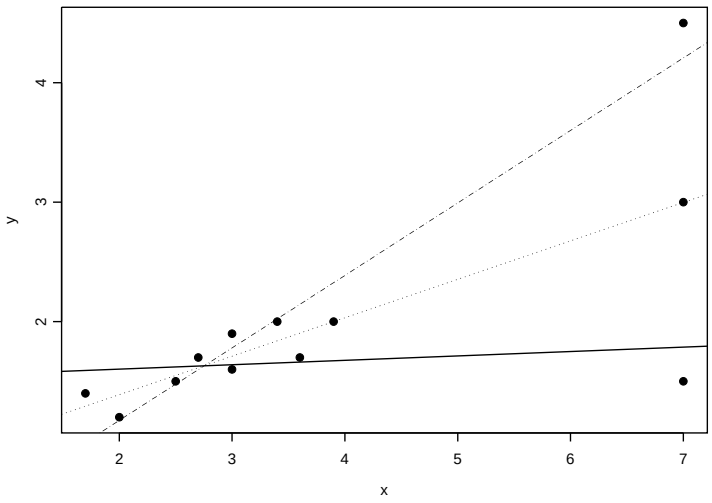

Figure 4.9: A leverage point in regression

It can be shown that the diagonal element h_{ii} of the hat-matrix in (4.5) indicates the amount of *leverage*, or influence, the ith observation has on the least squares line. The larger the value of h_{ii}, the more influence the observation has on the least squares line. (Recall that the largest value h_{ii} can take is 1.) It can be shown that the average value of the h_{ii}s is $(k+1)/n$; a rule of thumb says that an observation is a leverage point if it has a hat-diagonal h_{ii} greater than $2(k+1)/n$. Recall that the hat-matrix, $\mathbf{h} = \mathbf{x}\left(\mathbf{x}^T\mathbf{x}\right)^{-1}\mathbf{x}^T$, only depends

on the design matrix and not on the response variables Y_i. That is, the observed value of the response variable is irrelevant with regard to whether or not a point (x_i, y_i) is a leverage point.

Note that leverage points do not necessarily constitute a problem. If the observation y_i corresponding to a leverage point lies close to the general trend in the data, the point is called a **good leverage point**, and there is no reason to do anything about the data point. However, if y_i differs from the main trend–in particular, if y_i corresponds to an outlier–the point is called a **bad leverage point**, and should be removed from the dataset.

Example 4.2(continued) *Ice cream consumption*
In Figure 4.10 the studentised residuals are plotted against the values h_{ii} of the hat-matrix for the ice cream data. The h_{ii}s are plotted along the horizontal axis. In this example $k = 4$

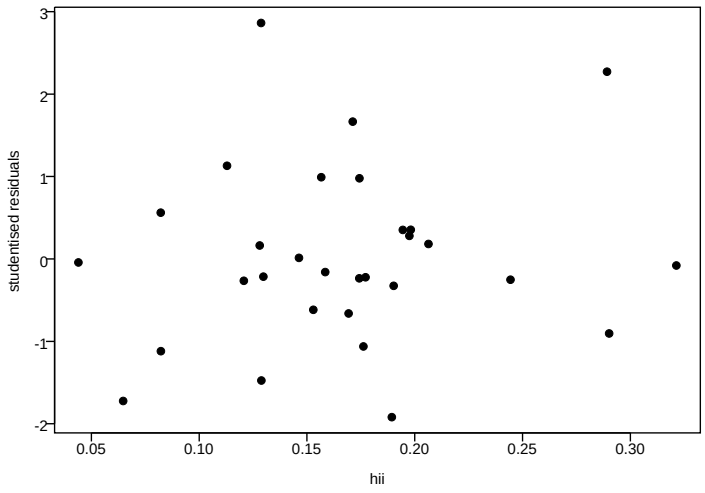

Figure 4.10: Studentised residuals and hat-diagonals for the ice cream data

and $n = 30$, so $2(k+1)/n = 10/30 = 1/3$, that is, the rule of thumb suggests that we should investigate observations for which $h_{ii} > 1/3$. There is one observation with an h_{ii} around $1/3$, but since the studentised residual for the point is close to zero, it seems to be a good leverage point. Two more points have high leverage ($h_{ii} \approx 0.29$), one of which has a high studentised residual ($t_i = 2.27$, corresponding to the 98.4%-quantile). We could have considered omitting this point from the dataset before analysing the data in Module 3.

There is one point in Figure 4.10 for which the studentised residual is a fair bit larger than the rest ($t_i = 2.68$, corresponding to the 99.6%-quantile); this point corresponds to the outlier that was removed from this dataset in Module 3. (It is not a very extreme outlier and it has low leverage, so we could have chosen to leave it in the dataset.)

♢

4.7 Summary

The assumptions of multiple linear regression models are that the response variables are independent normally distributed random variables with constant variance and means depending linearly on the explanatory variables. These assumptions are equivalent to the random errors being independent normally distributed random variables with zero mean and constant variance. The assumptions on the response variables are checked by assessing the assumptions on the random errors. The normality assumption is checked by means of a normal probability plot of the standardised residuals. The assumption on constant variance is checked by means of a residual plot of the standardised residuals. The linearity assumption can be checked through partial residual plots. Finally, we can check for outliers by considering the studentised residuals, and for leverage points by considering the diagonal elements of the hat-matrix.

Keywords: model assumptions, independence assumption, normality assumption, homoscedasticity assumption, linearity assumption, raw residual, hat-matrix, standardised residual, normal probability plot, residual plot, partial residual, partial residual plot, outlier, studentised residual, leverage point, good leverage point, bad leverage point.