Part I

Simple Linear Regression

Contents

1 The simple linear regression model

1.1 The model

We start with the simplest situation where we have one **response variable** Y and one **predictor** (or **explanatory**) **variable** X. In many practical situations we deal with a predictor variable X that can be controlled (known) and a response variable Y which can be observed (unknown). We want to predict (or estimate) the mean value of Y for given values of X working from a sample on n pairs of observations

$$\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}.$$

Mathematically, the regression of a random variable Y on a variable X means finding the expected value μ_x of Y when X takes the specific value x, that is

$$\mu_x = \mathbb{E}(Y|X = x).$$

For example, if X = Day of the week and Y = Sales at a given company, then the regression of Y on X represents the mean (or average) sales on a given day. The regression of Y on X is linear if

$$\mu_x = \beta_0 + \beta_1 x \tag{1.1}$$

where the unknown parameters β_0 and β_1 determine the intercept and the slope of a specific straight line.

Suppose that $Y_1, Y_2, \ldots, Y_n$ are independent realisations of the random variable Y that are observed at values $x_1, x_2, \ldots, x_n$ of X. If the regression of Y on X is linear, then for $i = 1, 2, \ldots, n$

$$Y_i = \mathbb{E}(Y|X = x_i) + \varepsilon_i = \beta_0 + \beta_1 x_i + \varepsilon_i \tag{1.2}$$

where ε_i is the **random error** in Y_i and is such that $\mathbb{E}(\varepsilon_i|X) = 0$ for all values of X. (We should write ε_i in capital since it is a random variable, but the E will be used elsewhere.)

The random error term is there since there will almost certainly be some variation in Y due strictly to random phenomenon that cannot be predicted or explained. In other words, **all unexplained variation is called random error**. Thus, the random error term does not depend on X, nor does it contain any information about Y (otherwise it would be a systematic error).

We will assume that the random errors ε_i are independent identically distributed normal random variables with mean 0 and common variance σ^2

$$\varepsilon_i \overset{ind}{\sim} N(0, \sigma^2).$$

This in turn implies that

$$Y_i \overset{ind}{\sim} N(\mu_i, \sigma^2) \quad \text{with} \quad \mu_i = \beta_0 + \beta_1 x_i.$$

With these assumptions in place the model is called the **normal SLR model**. The parameters β_0 and β_1 are called **regression coefficients**.

1.2 Least squares estimation

Suppose for example that X = height and Y = weight of a randomly selected individual from some population Then for a straight line regression model the mean weight of individuals of a given height would be a linear function of that height. In practice, we usually have a sample of data instead of the whole population. The slope β_1 and intercept β_0 are unknown, since these are the values for the whole population. Thus, we wish to use the sample at hand to estimate the slope and the intercept. This can be achieved by finding the equation of the line which "best" fits our data, that is, choose β_0 and β_1 such that

$$\hat{y}_i = \beta_0 + \beta_1 x_i$$

is as "close" as possible to y_i. Here the notation $\hat{y}_i$ is used to denote the value of the **line of best fit** in order to distinguish it from the observed values y_i of Y_i. We shall refer to $\hat{y}_i$ as the i-th **predicted value** or the **fitted value** of y_i.

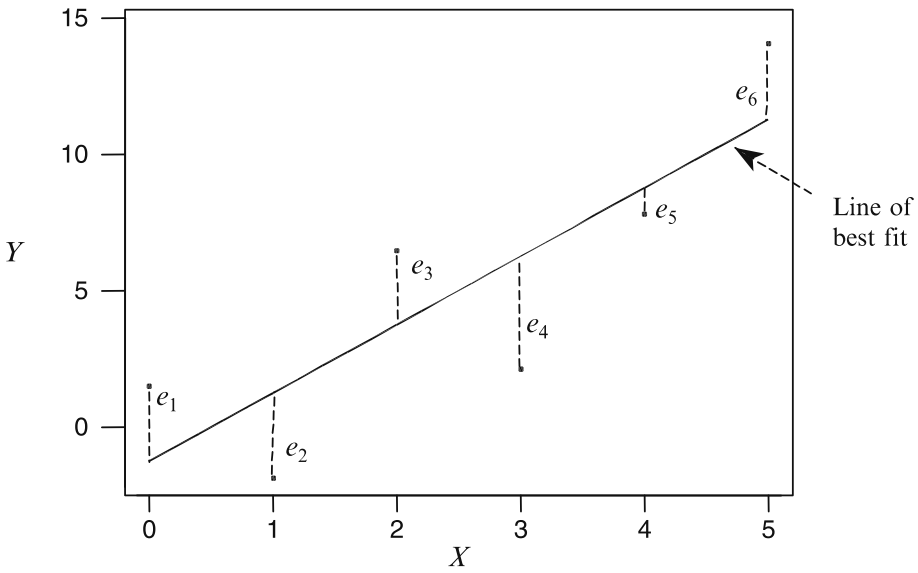

Figure 1.1: A scatter plot of data with a line of best fit and the residuals identified.

In practice, we wish to minimise the difference between the observed value y_i and the predicted value $\hat{y}_i$. This difference is called the **residual**, e_i, that is,

$$e_i = y_i - \hat{y}_i.$$

Figure 1.1 shows a hypothetical situation based on six data points. Marked on this plot is a **line of best fit**, $\hat{y}_i$ along with the residuals.

The most common method of estimating β_0 and β_1 is the method of least squares. As the name suggests, β_0 and β_1 are chosen to minimize the **residual** (or **error**) **sum of squares**, SS_E (also denoted by SS_{Res} or RSS):

$$SS_E = \sum_{i=1}^{n} e_i^2.$$

Claim 1.1:

The least squares estimates of β_0 and β_1 for the SLR model are

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}, \qquad \hat{\beta}_1 = \frac{s_{xy}}{s_{xx}} \tag{1.3}$$

where

$$s_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2, \qquad s_{xy} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}), \tag{1.4}$$

and $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ and $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ are the mean values of y_i and x_i, respectively.

Proof. First, we rewrite SS_E as

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

It is a function of two parameters and so to find its minimum we differentiate it with respect to β_0 and β_1:

$$\frac{\partial SS_E}{\partial \beta_0} = -2\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i),$$
$$\frac{\partial SS_E}{\partial \beta_1} = -2\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)x_i.$$

Since (by definition) $\hat{\beta}_0$ and $\hat{\beta}_1$ minimise SS_E, both partial derivative must equal 0 when $\beta_0 = \hat{\beta}_0$ and $\beta_1 = \hat{\beta}_1$, giving

$$0 = -2\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i),$$
$$0 = -2\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)x_i.$$

They yield the so-called **normal equations**:

$$n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i, \tag{1.5}$$
$$\hat{\beta}_0 \sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i. \tag{1.6}$$

Solving the first normal equation for $\hat{\beta}_0$ we find

$$\hat{\beta}_0 = \frac{1}{n}\sum_{i=1}^{n} y_i - \frac{1}{n}\hat{\beta}_1 \sum_{i=1}^{n} x_i = \bar{y} - \hat{\beta}_1 \bar{x}.$$

We now substitute this result to the second normal equation

$$(\bar{y} - \hat{\beta}_1 \bar{x}) \sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i$$

and solve it for $\hat{\beta}_1$:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2}.$$

We need to rewrite this result in terms of the sums s_{xy} and s_{xx}. Note that $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$ and

$$\sum_{i=1}^{n}(x_i - \bar{x})\bar{y} = \bar{y}\sum_{i=1}^{n}(x_i - \bar{x}) = 0.$$

We may thus write the sums s_{xy} and s_{xx} as

$$s_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n}(x_i - \bar{x})y_i = \sum_{i=1}^{n} x_i y_i - \frac{1}{n}\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i,$$
$$s_{xx} = \sum_{i=1}^{n}(x_i^2 - 2x_i\bar{x} + \bar{x}^2) = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2 = \sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2,$$

leading to

$$\hat{\beta}_1 = \frac{s_{xy}}{s_{xx}}$$

which is the wanted expression. It remains to verify that SS_E attains its minimum at these values. We need to calculate the second derivatives and evaluate the Hessian determinant:

$$H(SS_E) = \begin{vmatrix} \frac{\partial^2 SS_E}{\partial \beta_0^2} & \frac{\partial^2 SS_E}{\partial \beta_0 \partial \beta_1} \\ \frac{\partial^2 SS_E}{\partial \beta_1 \partial \beta_0} & \frac{\partial^2 SS_E}{\partial \beta_1^2} \end{vmatrix} = \begin{vmatrix} 2n & 2\sum_{i=1}^n x_i \\ 2\sum_{i=1}^n x_i & 2\sum_{i=1}^n x_i^2 \end{vmatrix} = 4n\sum_{i=1}^n (x_i - \bar{x})^2.$$

Thus $H(SS_E) > 0$. Moreover,

$$\frac{\partial^2 SS_E}{\partial \beta_0^2} = 2n > 0, \qquad \frac{\partial^2 SS_E}{\partial \beta_1^2} = 2\sum_{i=1}^n x_i^2 > 0$$

for all β_0, β_1, thus the function SS_E indeed attains its minimum at $(\hat{\beta}_0, \hat{\beta}_1)$. □

Example 1: Manufacturer production data

A manufacturer wants to investigate the time it takes (in minutes) to produce individual orders of different sizes. Data from 20 randomly selected orders is given in the table below.

Order	1	2	3	4	5	6	7	8	9	10
Run Time	195	215	243	162	185	231	234	166	253	196
Run Size	175	189	344	88	114	338	271	173	284	277
Order	11	12	13	14	15	16	17	18	19	20
Run Time	220	168	207	225	169	215	147	230	208	172
Run Size	337	58	146	277	123	227	63	337	146	68

Table 1.1: Production data.

The relation between the run time and run size is expected to be linear. The means of observations are $\bar{x} = 201.75$ (Run Size) and $\bar{y} = 202.05$ (Run Time). Thus

$$s_{xx} = \sum_{i=1}^{20} (x_i - 201.75)^2 = 191473.80, \quad s_{xy} = \sum_{i=1}^{20} (x_i - 201.75)(y_i - 202.05) = 49638.25$$

giving

$$\hat{\beta}_1 = \frac{49638.25}{191473.8} = 0.259, \qquad \hat{\beta}_0 = 202.05 - 0.26 \cdot 201.75 = 149.797$$

Therefore the equation of the fitted regression line is (shown in Figure 1.2)

$$\hat{y} = 149.797 + 0.259x.$$

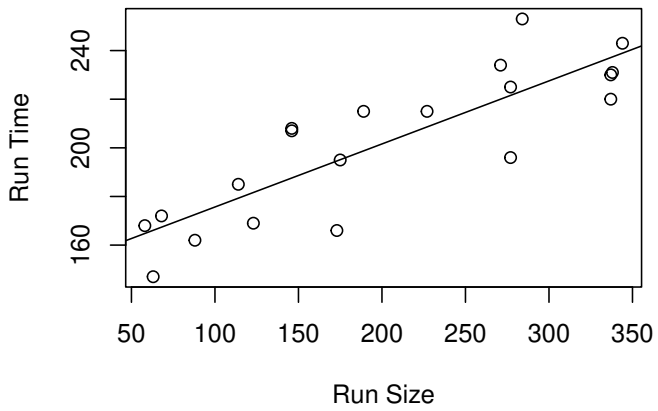

Figure 1.2: A plot of the production data with the fitted regression line.

The intercept is $\hat{\beta}_0 = 149.797$. This is where the regression line crosses the run time axis. We interpret this value as the average set up time, that is 149.80 minutes. The slope of the line is $\hat{\beta}_1 = 0.259$. Thus, we say that each additional unit to be produced is predicted to add 0.259 minutes (15.54 seconds) to the run time.

1.3 Properties of the slope and the intercept

We want to know how well $\hat{\beta}_1$ and $\hat{\beta}_0$ estimate the unknown true values of β_1 and β_0. For this we need to know statistical properties of $\hat{\beta}_1$ and $\hat{\beta}_0$ first.

Recall from (1.3) that

$$\hat{\beta}_1 = \frac{s_{xy}}{s_{xx}} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sum_{i=1}^n (x_i - \bar{x}) y_i}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

since $\sum_{i=1}^n (x_i - \bar{x}) = 0$. We may thus rewrite $\hat{\beta}_1$ as

$$\hat{\beta}_1 = \sum_{i=1}^n c_i y_i \quad \text{with} \quad c_i = \frac{x_i - \bar{x}}{\sum_{i=1}^n (x_i - \bar{x})^2}. \tag{1.7}$$

The $\hat{\beta}_1$ above is a scalar. Replacing observations y_i's with random variables Y_i's we obtain a random variable, also denoted $\hat{\beta}_1$ (due to the lack of a capital β letter). The random variable $\hat{\beta}_1$ is the least squares estimator of β_1.

Claim 1.2:

$\hat{\beta}_1$ is an unbiased estimator of β_1, that is

$$\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{s_{xx}}\right). \tag{1.8}$$

Proof. Recall that $Y_i \overset{ind}{\sim} N(\mu_i, \sigma^2)$ with $\mu_i = \beta_0 + \beta_1 x_i$. Then from (1.7) we find

$$\begin{aligned}\mathbb{E}(\hat{\beta}_1) &= \mathbb{E}\Big(\sum_{i=1}^n c_i Y_i\Big) = \sum_{i=1}^n c_i\, \mathbb{E}(Y_i) \\ &= \sum_{i=1}^n c_i(\beta_0 + \beta_1 x_i) = \beta_0 \sum_{i=1}^n c_i + \beta_1 \sum_{i=1}^n c_i x_i.\end{aligned}$$

But $\sum_{i=1}^n c_i = 0$ and $\sum_{i=1}^n c_i x_i = 1$ since $\sum_{i=1}^n (x_i - \bar{x})x_i = s_{xx}$. Hence $\mathbb{E}(\hat{\beta}_1) = \beta_1$, and so $\hat{\beta}_1$ is an unbiased estimator of β_1.

Next, since Y_i's are independent, we find

$$\mathrm{Var}(\hat{\beta}_1) = \mathrm{Var}\Big(\sum_{i=1}^n c_i Y_i\Big) = \sum_{i=1}^n c_i^2 \cdot \mathrm{Var}(Y_i) = \sum_{i=1}^n \frac{(x_i - \bar{x})^2}{s_{xx}^2} \cdot \sigma^2 = \frac{\sigma^2}{s_{xx}}.$$

A linear combination of normally distributed random variables is also a normally distributed random variable, which implies (1.8). □

We now repeat the same analysis for $\hat{\beta}_0$. Recall from (1.3) that

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = \frac{1}{n}\sum_{i=1}^n y_i - \bar{x}\sum_{i=1}^n c_i y_i = \sum_{i=1}^n \Big(\frac{1}{n} - c_i \bar{x}\Big) y_i.$$

Replacing y_i's with Y_i's we find the least squares estimator of β_0 to be

$$\hat{\beta}_0 = \sum_{i=1}^n \Big(\frac{1}{n} - c_i \bar{x}\Big) Y_i \qquad \text{with} \quad c_i = \frac{x_i - \bar{x}}{s_{xx}}. \tag{1.9}$$

Claim 1.3:

$\hat{\beta}_0$ is an unbiased estimator of β_0, that is

$$\hat{\beta}_0 \sim N\Big(\beta_0, \Big(\frac{1}{n} + \frac{\bar{x}^2}{s_{xx}}\Big)\sigma^2\Big). \tag{1.10}$$

Proof. Using $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x}$ and Claim 1.2 we find

$$\mathbb{E}(\hat{\beta}_0) = \mathbb{E}(\bar{Y}) - \mathbb{E}(\hat{\beta}_1 \bar{x}) = \beta_0 + \beta_1 \bar{x} - \bar{x}\,\mathbb{E}(\hat{\beta}_1) = \beta_0 + \beta_1 \bar{x} - \bar{x}\beta_1 = \beta_0.$$

Then from (1.9) we obtain

$$\begin{aligned}\mathrm{Var}(\hat{\beta}_0) &= \mathrm{Var}\Big(\sum_{i=1}^n \Big(\frac{1}{n} - c_i \bar{x}\Big) Y_i\Big) \\ &= \sum_{i=1}^n \Big(\frac{1}{n} - c_i \bar{x}\Big)^2 \mathrm{Var}(Y_i) \\ &= \sum_{i=1}^n \Big(\frac{1}{n^2} - \frac{2c_i \bar{x}}{n} + c_i^2 \bar{x}^2\Big)\sigma^2 \\ &= \Big(\frac{1}{n} - 0 + \sum_{i=1}^n \frac{(x_i - \bar{x})^2}{s_{xx}^2}\bar{x}^2\Big)\sigma^2 = \Big(\frac{1}{n} + \frac{\bar{x}^2}{s_{xx}}\Big)\sigma^2\end{aligned}$$

and the assertion follows. □

Remark 1.1. (i) For large samples, where there is no assumption of normality of Y_i, the sampling distributions of $\hat{\beta}_1$ and $\hat{\beta}_0$ are approximately normal.
(ii) Variance $\mathrm{Var}(\hat{\beta}_1) = \sigma^2/s_{xx}$ decreases as s_{xx} increases (i.e., as the variability in the x's increases). This is an important fact to note if the experimenter has control over the choice of the values of the X variable. □

1.4 Estimating variance of the random error term

In the previous subsection we found that distributions of the estimators of the slope and the intercept are parametrised by the unknown true values of β_1, β_0 and σ^2. This means that making predictions about the slope and the intercept requires knowing σ^2. This is only possible when there are multiple observations on Y_i for at least one value of x_i. When this approach cannot be used, the estimate of σ^2 is obtained from the error sum of squares.

Notice that

$$\varepsilon_i = Y_i - (\beta_0 + \beta_1 x_i) = Y_i - (\textit{unknown regression line at } x_i).$$

Since β_0 and β_1 are unknown all we can do is to estimate the unknown random errors ε_i by replacing β_0 and β_1 with their least squares estimators, $\hat{\beta}_0$ and $\hat{\beta}_1$, giving the residuals

$$E_i = Y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i) = Y_i - (\textit{estimated regression line at } x_i).$$

These residuals can now be used to estimate σ^2. Indeed, let $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$. Then it can be shown that[1]

$$\frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{\sigma^2} \sim \chi^2_{n-2}$$

and

$$\mathbb{E}\Big[\sum_{i=1}^n (Y_i - \hat{Y}_i)^2\Big] = (n-2)\sigma^2.$$

On the other hand,

$$\sum_{i=1}^n (y_i - \hat{y}_i)^2 = SS_E.$$

Therefore

$$\hat{\sigma}^2 \equiv MS_E = \frac{SS_E}{n-2} \tag{1.11}$$

is an unbiased estimate of σ^2. (The estimate $\hat{\sigma}^2$ is also denoted by S^2.) The denominator $n-2$ in (1.11) can be explained by the fact that there are only $n-2$ linearly independent residuals. Indeed, the normal equations (1.5) and (1.6) imply that

$$\sum_{i=1}^n e_i = \sum_{i=1}^n e_i x_i = 0$$

which allow us to express e_{n-1} and e_n as linear combinations of $e_1, \ldots, e_{n-2}$. The quantity $\nu_E = n-2$ is called the **number of degrees of freedom** of SS_E, and the quantity MS_E is called the **mean residual** (or **error**) **sum of squares**.

[1] For a prove of this statement see Appendix C.3 in D. C. Montgomery, E. A. Peck and G. G. Vining, *Introduction to Linear Regression Analysis*, 5th edition, Wiley (2012). However the techniques used in that prove go beyond the scope of the present course.

Example 2: Manufacturer production data

For the production data we have

$$SS_E = \sum_{i=1}^{20}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{20}(y_i - 149.80 - 0.26x_i)^2 = 4754.58$$

giving $\hat{\sigma}^2 = 4754.58/18 = 264.14$.

Remark 1.2. Recall that sample variance is given by a very similar formula to (1.11)

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2.$$

In this case there are $n-1$ independent y_i's around $\bar{y}$. In other words, there are $n-1$ degrees of freedom of variation around $\bar{y}$. □

Remark 1.3. Residuals play an important role in investigating model adequacy and in detecting departures from the underlying assumptions. This will be discussed in subsequent sections. □

1.5 Testing hypotheses for the slope and intercept

Suppose we wish to test the hypothesis that the slope β_1 equals a certain value, say β_1^*, and the true regression model is $Y_i = \beta_0 + \beta_1^* x_i + \varepsilon_i$. The appropriate hypotheses are

$$H_0 : \beta_1 = \beta_1^*, \qquad H_1 : \beta_1 \neq \beta_1^*, \tag{1.12}$$

where H_0 is the null hypothesis and H_1 is a two-sided alternative. If the null hypothesis is true, Claim 1.2 implies that $\hat{\beta}_1 \sim N(\beta_1^*, \sigma^2/s_{xx})$. Standartising gives

$$Z = \frac{\hat{\beta}_1 - \beta_1^*}{\sqrt{\sigma^2/s_{xx}}} \sim N(0,1). \tag{1.13}$$

If σ^2 was known, we could use a Z-test to test the hypothesis. However, typically, σ^2 is unknown. Replacing σ^2 with its estimate $\hat{\sigma}^2 = MS_E$ causes $N(0,1)$ to be replaced with a Student's t-distribution with $\nu = n-2$ d.o.f.

$$T = \frac{\hat{\beta}_1 - \beta_1^*}{\sqrt{\hat{\sigma}^2/s_{xx}}} = \frac{\hat{\beta}_1 - \beta_1^*}{\mathrm{se}(\hat{\beta}_1)} \sim t_{n-2}. \tag{1.14}$$

where $\mathrm{se}(\hat{\beta}_1) = \sqrt{\hat{\sigma}^2/s_{xx}}$ is called the **estimated standard error**. The number of d.o.f. satisfies the following formula:

$$\textit{d.o.f.} = \textit{sample size} - \textit{number of mean parameters estimated}.$$

In the case at hand we have estimated two such parameters, β_0 and β_1, hence $\nu = n-2$.

The test procedure computes the value t of T in (1.14) for a given data set, and compares with the upper $\alpha/2$ percentage point of the t_{n-2} distribution, $t_{\alpha/2,n-2}$, where α is an a priori chosen significance level. This procedure rejects the null hypothesis if

$$|t| > t_{\alpha/2,n-2}.$$

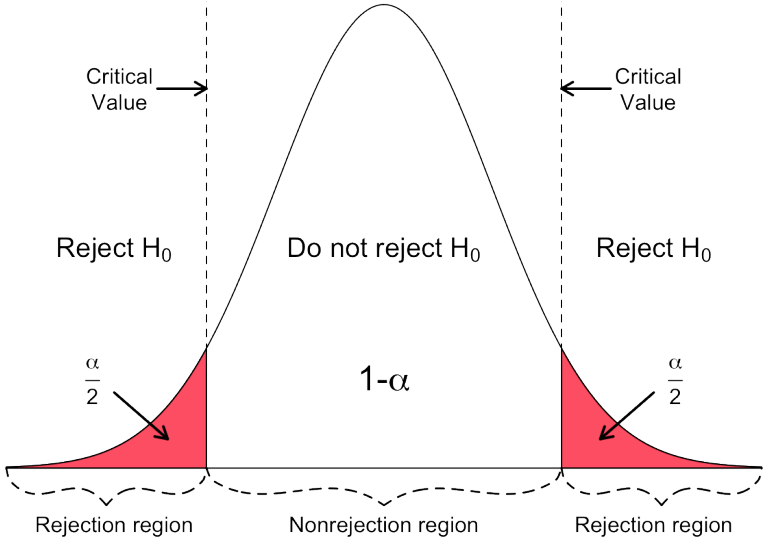

Figure 1.3: Students T-test.

There is not sufficient evidence to reject H_0 otherwise. (See Figure 1.3.)

A similar procedure can be used to test hypotheses about the intercept:

$$H_0 : \beta_0 = \beta_0^*, \qquad H_1 : \beta_0 \neq \beta_0^*. \tag{1.15}$$

The only difference is that the test statistic is now

$$T = \frac{\hat{\beta}_0 - \beta_0^*}{\text{se}(\hat{\beta}_0)} \sim t_{n-2} \tag{1.16}$$

where $\text{se}(\hat{\beta}_0) = \sqrt{\hat{\sigma}^2\left(1/n + \bar{x}^2/s_{xx}\right)}$ is the estimated standard error of $\hat{\beta}_0$.

Remark 1.4. Alternatively, a P-value approach could also be used for decision making. In this case the null hypothesis is rejected if

$$P(t) < \alpha. \qquad \square$$

Remark 1.5. The explanation of $n-2$ d.o.f in (1.14) is as follows. Recall that if $Z \sim N(0,1)$ and $U \sim \chi^2_\nu$, and Z and U are independent, then

$$\frac{Z}{\sqrt{U/\nu}} \sim t_\nu. \tag{1.17}$$

It can be shown that (see Appendix C.3 in D. C. Montgomery, E. A. Peck and G. G. Vining, *Introduction to Linear Regression Analysis*, 5th edition, Wiley (2012))

$$U = \frac{(n-2)\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-2}$$

and that $\hat{\sigma}^2$ and $\hat{\beta}_1$ are independent random variables. Then (1.13) and (1.17) imply that

$$T = \frac{\hat{\beta}_1 - \beta_1^*}{\sqrt{\sigma^2/s_{xx}}} \Big/ \sqrt{\frac{(n-2)\hat{\sigma}^2}{\sigma^2(n-2)}} = \frac{\hat{\beta}_1 - \beta_1^*}{\sqrt{\hat{\sigma}^2/s_{xx}}} \sim t_{n-2}. \qquad \square$$

Remark 1.6. There is an alternate form of the SLR model that is occasionally useful. Suppose that we redefine the predictors x_i as deviations from their own average, $x_i - \bar{x}$. The regression model then becomes

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 x_i - \beta_1 \bar{x} + \beta_1 \bar{x} + \varepsilon_i \\ &= (\beta_0 + \beta_1 \bar{x}) + \beta_1 (x_i - \bar{x}) + \varepsilon_i \\ &= \alpha + \beta (x_i - \bar{x}) + \varepsilon_i, \end{aligned}$$

where $\alpha = \beta_0 + \beta_1 \bar{x}$ and $\beta = \beta_1$.

It is easy to show that the least-squares estimate of the transformed intercept α is $\hat{\alpha} = 0$. The estimate of the slope β is unaffected by the transformation. This alternate form of the model has some advantages – some applications of the model are easier. For instance, the estimators $\hat{\alpha} = \bar{Y}$ and $\hat{\beta} = s_{xy}/s_{xx}$ (with y_i's replaced by Y_i's in s_{xy}) are uncorrelated, $\mathrm{Cov}(\hat{\alpha}, \hat{\beta}) = 0$. □

2 Further inference and significance of regression

2.1 Confidence intervals on the slope and intercept

In addition to point estimates of β_0 and β_1, we may also estimate confidence intervals (CI's) on these parameters. The width of these CI's is a measure of the overall quality of the regression line.

To find a CI on an unknown parameter θ means to find boundaries a and b such that

$$P(a \leq \theta \leq b) = 1 - \alpha$$

for some small α, that corresponds to a high confidence level $(1-\alpha)100\%$. The boundaries will depend on the data and so using the parameter estimates is a natural way of finding the CI. For example, when $-a = b = 1.96$ and θ is an unknown parameter from a normal distribution, we have that

$$P(-1.96 \leq \theta \leq 1.96) = 0.95.$$

In other words, 95% of observations of a normal population are within 1.96 standard deviations of the mean.

Let us now apply this knowledge to β_1. Using (1.14) we write

$$P\left(-t_{\alpha/2,\, n-2} \leq \frac{\hat{\beta}_1 - \beta_1}{\mathrm{se}(\hat{\beta}_1)} \leq t_{\alpha/2,\, n-2} \right) = 1 - \alpha.$$

Rearranging the expression in the brackets above gives

$$P\Big(\hat{\beta}_1 - t_{\alpha/2,\, n-2} \cdot \mathrm{se}(\hat{\beta}_1) \leq \beta_1 \leq \hat{\beta}_1 + t_{\alpha/2,\, n-2} \cdot \mathrm{se}(\hat{\beta}_1) \Big) = 1 - \alpha.$$

This is a probability statement about random variables $\hat{\beta}_1$ and $\hat{\sigma}^2$. When we replace them by their values from the observed data we obtain

$$\mathrm{CI}(\beta_1) = \Big[\hat{\beta}_1 - t_{\alpha/2,\, n-2} \cdot \mathrm{se}(\hat{\beta}_1),\ \hat{\beta}_1 + t_{\alpha/2,\, n-2} \cdot \mathrm{se}(\hat{\beta}_1) \Big]$$

as our $100(1-\alpha)\%$ confidence interval for β_1. In other words, there is a $100(1-\alpha)\%$ probability that the unknown true value of β_1 is within this CI.

By the same arguments, a $100(1-\alpha)\%$ CI on β_0 is

$$\text{CI}(\beta_0) = \Big[\hat{\beta}_0 - t_{\alpha/2,\,n-2} \cdot \text{se}(\hat{\beta}_0),\ \hat{\beta}_0 + t_{\alpha/2,\,n-2} \cdot \text{se}(\hat{\beta}_0)\Big].$$

Remark 2.1. The sampling distribution of SS_R/σ^2 is χ^2_{n-2}. Thus

$$P\Big(\chi^2_{\alpha/2,\,n-2} \le SS_R/\sigma^2 \le \chi^2_{1-\alpha/2,\,n-2}\Big) = 1-\alpha$$

and consequently a $100(1-\alpha)\%$ CI for σ^2 is

$$\text{CI}(\sigma^2) = \Big[SS_R/\chi^2_{\alpha/2,\,n-2},\ SS_R/\chi^2_{1-\alpha/2,\,n-2}\Big]. \qquad \square$$

Example 3: Manufacturer production data

We want to test hypotheses (1.12) for $\beta_1^* = 0$ and (1.15) for $\beta_0^* = 0$ assuming $\alpha = 5\% = 0.05$. We already know that

$$\hat{\beta}_0 = 149.797, \quad \hat{\beta}_1 = 0.259, \quad \hat{\sigma}^2 = 264.14, \quad s_{xx} = 191473.75, \quad \bar{x} = 201.75.$$

Thus the estimated standard errors are

$$\text{se}(\hat{\beta}_0) = \sqrt{264.14 \times (1/20 + (201.75)^2/191473.75)} = 8.328,$$
$$\text{se}(\hat{\beta}_1) = \sqrt{264.14/191473.75} = 0.037.$$

The computed values of the test statistic are

$$T(\hat{\beta}_0) = \frac{149.797}{8.328} = 17.987, \qquad T(\hat{\beta}_1) = \frac{0.259}{0.037} = 7.0.$$

The critical value of the test statistic is $t_{0.025,18} = 2.1$. Since both $T(\hat{\beta}_0)$ and $T(\hat{\beta}_1)$ are greater than the critical value, we reject the null hypothesis in both cases.
Finally, we want to find 95% CIs for both β_0 and β_1:

$$\text{CI}(\beta_0) = [149.797 \pm 2.1 \times 8.328] = [132.3, 167.2],$$
$$\text{CI}(\beta_1) = [0.259 \pm 2.1 \times 0.037] \quad = [0.181, 0.337].$$

2.2 Estimating mean response

A major use of a regression model is to estimate the mean response $\mathbb{E}(Y)$ for a given value of the predictor variable $X = x_0$. Using (1.2) we can write the mean response as

$$\mu_0 = \mathbb{E}(Y|X = x_0) = \beta_0 + \beta_1 x_0.$$

An estimate of this unknown quantity μ_0 is the value of the estimated regression equation at $X = x_0$, namely,

$$\hat{\mu}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0.$$

Viewing $\hat{\beta}_0$ and $\hat{\beta}_1$ as random variables yields an estimator $\hat{\mu}_0$ of μ_0.

Claim 2.1:

$\hat{\mu}_0$ is an unbiased estimator of μ_0, that is

$$\hat{\mu}_0 \sim N\left(\mu_0, \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{s_{xx}}\right)\sigma^2\right).$$

Proof. We already know that $\mathbb{E}(\hat{\mu}_0) = \mu_0$. Thus we only need to compute $\mathrm{Var}(\hat{\mu}_0)$. We recall that the estimators of β_0 and β_1 are

$$\hat{\beta}_0 = \sum_{i=1}^{n}\left(\frac{1}{n} - c_i\bar{x}\right)Y_i, \qquad \hat{\beta}_1 = \sum_{i=1}^{n} c_i Y_i, \quad \text{where} \quad c_i = \frac{x_i - \bar{x}}{s_{xx}}.$$

Thus

$$\begin{aligned}
\mathrm{Var}(\hat{\mu}_0) &= \mathrm{Var}(\hat{\beta}_0 + \hat{\beta}_1 x_0) \\
&= \mathrm{Var}\left(\sum_{j=1}^{n}\left(\frac{1}{n} - c_j\bar{x}\right)Y_j + \sum_{j=1}^{n} c_j Y_j x_0\right) \\
&= \mathrm{Var}\left(\sum_{j=1}^{n}\left(\frac{1}{n} + c_j(x_0 - \bar{x})\right)Y_j\right) \\
&= \sum_{j=1}^{n}\left(\frac{1}{n} + c_j(x_0 - \bar{x})\right)^2 \mathrm{Var}(Y_j) \\
&= \sum_{j=1}^{n}\left(\frac{1}{n^2} + \frac{2}{n}\cdot\frac{(x_j - \bar{x})(x_0 - \bar{x})}{s_{xx}} + \frac{(x_j - \bar{x})^2}{s_{xx}^2}(x_0 - \bar{x})^2\right)\sigma^2 \\
&= \left(\frac{1}{n} + 0 + \frac{s_{xx}}{s_{xx}^2}(x_0 - \bar{x})^2\right)\sigma^2 = \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{s_{xx}}\right)\sigma^2.
\end{aligned}$$

□

Claim 2.1 allows us to test the hypotheses

$$H_0 : \mu_0 = \mu_0^*, \qquad H_1 : \mu_0 \neq \mu_0^*,$$

where μ_0^* is a particular value. The test statistic is

$$T = \frac{\hat{\mu}_0 - \mu_0^*}{\mathrm{se}(\hat{\mu}_0)} \sim t_{n-2}, \qquad \mathrm{se}(\hat{\mu}_0) = \sqrt{\hat{\sigma}^2\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{s_{xx}}\right)}.$$

Lastly, by the same arguments as before, a $100(1-\alpha)\%$ CI on μ_0 is

$$\mathrm{CI}(\mu_0) = \left[\hat{\mu}_0 - t_{\alpha/2,\,n-2}\cdot\mathrm{se}(\hat{\mu}_0),\ \hat{\mu}_0 + t_{\alpha/2,\,n-2}\cdot\mathrm{se}(\hat{\mu}_0)\right]. \tag{2.1}$$

Remark 2.2. Care is needed when estimating the mean response μ_0 at x_0. It should only be done if x_0 is within the data range, i.e., if $\min\{x_i\} < x_0 < \max\{x_i\}$. Extrapolation beyond the range of the given x-values is not reliable, as there is no evidence that a linear relationship is appropriate there. □

2.3 Prediction of a new observation

An important application of the regression model is prediction of a new observation of Y corresponding to a specified level of the predictor variable X. If x_0 is the value of X of interest, then

$$\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$$

is the point estimate of the future observation of Y. We will denote by $\hat{Y}_0$ its least squares estimator.

We want to obtain an interval estimate of this future observation. The CI on the mean response at $X = x_0$ given by (2.1) is inappropriate for this problem because it is an interval estimate on the mean response, not a probability statement about future observations from that distribution.

We can write $\hat{Y}_0$ as

$$\hat{Y}_0 = \hat{\mu}_0 + \varepsilon_0$$

where $\hat{\mu}_0$ is an estimator of the mean response μ_0 at $X = x_0$, and ε_0 is a random error of the new observation. From Claim 2.1 we know that

$$\hat{\mu}_0 \sim N\left(\mu_0, \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{s_{xx}}\right)\sigma^2\right).$$

We have assumed that $\varepsilon_0 \sim N(0, \sigma^2)$, and that ε_0 is independent from $\hat{\mu}_0$. Thus

$$\hat{Y}_0 = \hat{\mu}_0 + \varepsilon_0 \sim N\left(\mu_0, \left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{s_{xx}}\right)\sigma^2\right)$$

and so

$$\frac{\hat{Y}_0 - \mu_0}{\sqrt{\sigma^2\left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{s_{xx}}\right)}} \sim N(0, 1).$$

Replacing σ^2 by its estimate $\hat{\sigma}^2$ gives

$$\frac{\hat{Y}_0 - \mu_0}{\mathrm{se}(\hat{y}_0)} \sim t_{n-2}, \qquad \mathrm{se}(\hat{y}_0) = \sqrt{\hat{\sigma}^2\left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{s_{xx}}\right)}.$$

Hence a $(1-\alpha)100\%$ PI on a new observation y_0 is

$$\mathrm{PI}(y_0) = \left[\hat{y}_0 - t_{\alpha/2,\, n-2} \cdot \mathrm{se}(\hat{y}_0),\ \hat{y}_0 + t_{\alpha/2,\, n-2} \cdot \mathrm{se}(\hat{y}_0)\right]$$

This interval is usually much wider than the CI for the mean response μ_0. This is because of the random error ε_0 reflecting the random source of variability in the data. Again, we should only make predictions for values of x_0 within the range of the data.

Remark 2.3. Prediction interval relies strongly on the assumption that the residual errors are normally distributed with a constant variance. So, you should only use such intervals if you believe that the assumption is approximately met for the data at hand. □

2.4 Testing significance of regression

A very important special case of the hypotheses (1.12) is

$$H_0 : \beta_1 = 0, \qquad H_1 : \beta_1 \neq 0.$$

These hypotheses relate to the **significance of regression**. Failing to reject $H_0 : \beta_1 = 0$ implies that there is no linear relationship between X and Y. This situation is illustrated in Figure 2.1. Note that this may imply either that X is of little value in explaining the variation in Y and that the best estimator of Y for any value of X is $\hat{Y} = \bar{Y}$ (Figure 2.1a) or that the true relationship between X and Y is not linear (Figure 2.1b). Therefore, failing to reject $H_0 : \beta_1 = 0$ is equivalent to saying that there is no linear relationship between Y and X.

Alternatively, if $H_0 : \beta_1 = 0$ is rejected, this implies that X is of value in explaining the variability in Y. This is illustrated in Figure 2.2. However, rejecting $H_0 : \beta_1 = 0$ could mean either that the straight-line model is adequate (Figure 2.2a) or that even though there is a linear effect of X, better results could be obtained with the addition of higher order polynomial terms in X (Figure 2.2b).

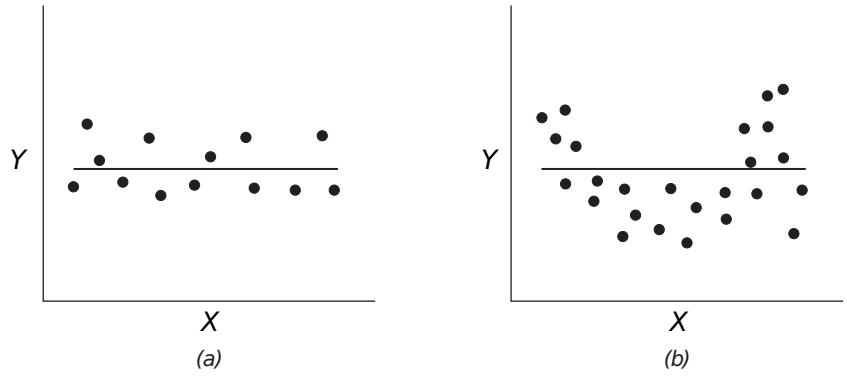

Figure 2.1: Situations where the hypothesis $H_0 : \beta_1 = 0$ is not rejected.

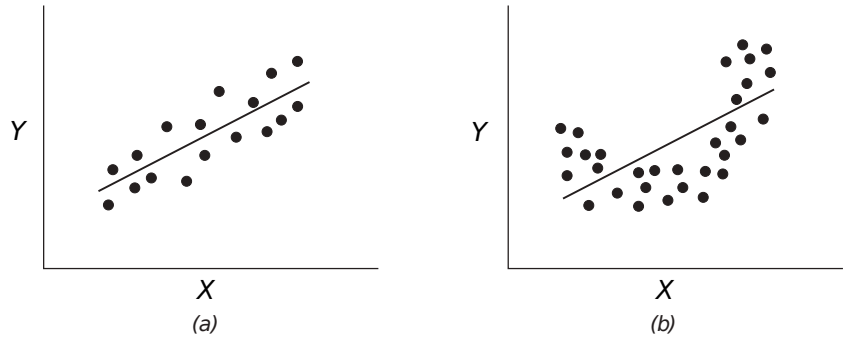

Figure 2.2: Situations where the hypothesis $H_0 : \beta_1 = 0$ is rejected.

2.5 Analysis of variance

We may also use the **analysis of variance** (ANOVA) approach to test the significance of regression. The analysis of variance is based on a partitioning of total variability in the response variable,

$$SS_T = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

into variability explained by the regression model,

$$SS_R = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

and the residual, or error, variability,

$$SS_E = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2.$$

Claim 2.2: The Analysis of Variance identity

In the SLR model the total sum of squares is a sum of the regression sum of squares and the residual sum of squares, that is

$$SS_T = SS_R + SS_E. \tag{2.2}$$

Proof. Partitioning the total variability gives

$$\begin{aligned} SS_T &= \sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} [(y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})]^2 \\ &= \sum_{i=1}^{n} [(y_i - \hat{y}_i)^2 + 2(y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) + (\hat{y}_i - \bar{y})^2] \\ &= SS_E + SS_R + 2\sum_{i=1}^{n} (y_i - \hat{y}_i)(\hat{y}_i - \bar{y}). \end{aligned}$$

However the last sum evaluates to zero,

$$\begin{aligned} \sum_{i=1}^{n} (y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) &= \sum_{i=1}^{n} (y_i - \hat{y}_i)\hat{y}_i - \bar{y}\sum_{i=1}^{n} (y_i - \hat{y}_i) \\ &= \sum_{i=1}^{n} e_i \hat{y}_i - \bar{y}\sum_{i=1}^{n} e_i \\ &= \sum_{i=1}^{n} e_i (\hat{\beta}_0 + \hat{\beta}_1 x_i) \\ &= \hat{\beta}_0 \sum_{i=1}^{n} e_i + \hat{\beta}_1 \sum_{i=1}^{n} e_i x_i = 0 \end{aligned}$$

since $\sum_{i=1}^{n} e_i = 0$ and $\sum_{i=1}^{n} e_i x_i = 0$. □

The Analysis of Variance identity is used to draw the Analysis of Variance table (Table 2.1). It shows the sources of variation, the sums of squares and the statistic, based on the sums of squares, for testing the significance of regression slope.

Source of variation	d.o.f.	SS	MS	F
Regression	$\nu_R = 1$	SS_R	$MS_R = \frac{SS_R}{\nu_R}$	$F = \frac{MS_R}{MS_E}$
Residual (Error)	$\nu_E = n-2$	SS_E	$MS_E = \frac{SS_E}{\nu_E}$	
Total	$\nu_T = n-1$	SS_T		

Table 2.1: ANOVA table.

To understand the numbers of degrees of freedom (d.o.f.), consider observations $y_1, \ldots, y_n$, and assume that their sum is fixed, say equal to $\sum_{i=1}^{n} y_i = a$. For a fixed value of the sum a there are $n-1$ arbitrary y-values but one y-value is determined by the difference of a and the $n-1$ arbitrary y-values. This one value is not free – it depends on the other y-values and on a. We say, that there are $n-1$ independent (free to vary) pieces of information and one piece is taken up by a.

Estimates of parameters can be based upon different amounts of information. The number of independent pieces of information that go into the estimate of a parameter is called the degrees of freedom. This is why in order to calculate

$$SS_T = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

we have $n-1$ free to vary pieces of information from the collected data, that is we have $n-1$ degrees of freedom. The one degree of freedom is taken up by $\bar{y}$. Similarly, for

$$SS_E = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_1 - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

we have two degrees of freedom taken up: one by $\hat{\beta}_0$ and one by $\hat{\beta}_1$ (both depend on $y_1, y_2, \ldots, y_n$). Hence, there are $n-2$ independent pieces of information used to calculate SS_E.

Finally, since $SS_R = SS_T - SS_E$, we can calculate the d.o.f. for SS_R as a difference between d.o.f. for SS_T and for SS_E, that is

$$\nu_R = \nu_T - \nu_E = (n-1) - (n-2) = 1.$$

This is exactly the number of slopes in the model.

In the ANOVA table there are also included the so-called **Mean Squares (MS)**, which can be thought of as **measures of average variation**. The last column of the table contains the variance ratio, or the **F-value**,

$$F = \frac{MS_R}{MS_E} = \frac{SS_R/\nu_R}{SS_E/\nu_E}.$$

It measures the variation explained by the model relative to the variation due to residuals, and can be used to test the hypotheses

$$H_0 : \beta_1 = 0, \qquad H_1 : \beta_1 \neq 0.$$

It can be shown that, provided the null hypothesis is true,

$$\frac{SS_R}{\sigma^2} \sim \chi^2_1, \qquad \frac{SS_E}{\sigma^2} \sim \chi^2_{n-2},$$

and SS_R and SS_E are independent. Then, by the definition of the Fisher's F-statistic,

$$F = \frac{MS_R}{MS_E} \sim F_{1,n-2}. \tag{2.3}$$

where $F_{1,n-2}$ is the Fisher's F-statistic with 1 and $n-2$ degrees of freedom.

The test procedure computes the value F^* of F in (2.3) for a given data set, and compares with $F_{\alpha,1,n-2}$, the percentile of the $F_{1,\,n-2}$ distribution corresponding to the cumulative probability of $(1-\alpha)$. We reject the null hypothesis if

$$F^* > F_{\alpha,1,n-2}.$$

Rejecting H_0 means that the slope $\beta_1 \neq 0$ and the full regression model

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

is better than the constant model

$$Y_i = \beta_0 + \varepsilon_i.$$

Example 4: Manufacturer production data

Source of variation	d.o.f.	SS	MS	F
Regression	1	12868.37	12868.37	48.72
Residual (Error)	18	4754.58	264.14	
Total	19	17622.95		

Table 2.2: ANOVA table for the production data.

Assuming $\alpha = 5\%$, we have $F_{\alpha,1,18} = 4.41$. Since $F^* = 48.72 > F_{\alpha,1,18}$, we conclude that regression is significant.

Remark 2.4. The F-test and the t-test for the hypotheses $H_0 : \beta_1 = 0$ and $H_1 : \beta_1 \neq 0$ are in fact equivalent since if a random variable $W \sim t_\nu$, then $W^2 \sim F_{1,\nu}$. Indeed, for the production data we found $T(\beta_1) = 6.98$ and $F = 48.72 = 6.98^2$. □

2.6 Coefficient of determination R^2

The coefficient of determination, denoted by R^2, is the percentage of total variation in y_i explained by the fitted model, that is

$$R^2 = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} = \frac{SS_R}{SS_T} = \frac{SS_T - SS_E}{SS_T} = \left(1 - \frac{SS_E}{SS_T}\right).$$

For the simple linear regression model R^2 is equal to a square of the Pearson correlation coefficient,

$$r^2 = \frac{s_{xy}^2}{s_{xx} s_{yy}}.$$

Note that:

- $R^2 \in [0,1]$ (or $[0,100]\%$).
- $R^2 = 0$ (0%) indicates that none of the variability in the data (y) is explained by the regression model.
- $R^2 = 1$ (or 100%) indicates that $SS_E = 0$ and all observations fall on the fitted line exactly.

Remark 2.5. R^2 is a measure of the linear association between Y and X. A small R^2 does not always imply a poor relationship between Y and X, which may, for example, follow a quadratic model. □

3 Diagnostics and transformations

In the previous sections we assumed that the normal simple linear regression model was a valid model for the data, that is,

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad \text{with} \quad \varepsilon_i \overset{ind}{\sim} N(0, \sigma^2).$$

In this section we will explain a series of tools known as **regression diagnostics** to check if the model under consideration is indeed a valid model. Our aim will be to understand what actually happens when the assumptions associated with the regression model are violated, and what should be done in response to this violation.

3.1 Valid and invalid regression models: Anscombe's four data sets

We shall consider four data sets constructed by the statistician Francis Anscombe in 1973. This example illustrates dramatically the point that looking only at the numerical regression output may lead to very misleading conclusions about the data, and lead to adopting the wrong model. The data are given in Table 3.1 and are plotted in Figure 3.1. Notice that the Y-values differ in each of the four data sets, while the X-values are the same for data sets 1, 2 and 3.

Case	x1	x2	x3	x4	y1	y2	y3	y4
1	10	10	10	8	8.04	9.14	7.46	6.58
2	8	8	8	8	6.95	8.14	6.77	5.76
3	13	13	13	8	7.58	8.74	12.74	7.71
4	9	9	9	8	8.81	8.77	7.11	8.84
5	11	11	11	8	8.33	9.26	7.81	8.47
6	14	14	14	8	9.96	8.10	8.84	7.04
7	6	6	6	8	7.24	6.13	6.08	5.25
8	4	4	4	19	4.26	3.10	5.39	12.5
9	12	12	12	8	10.84	9.13	8.15	5.56
10	7	7	7	8	4.82	7.26	6.42	7.91
11	5	5	5	8	5.68	4.74	5.73	6.89

Table 3.1: Anscombe's four data sets.

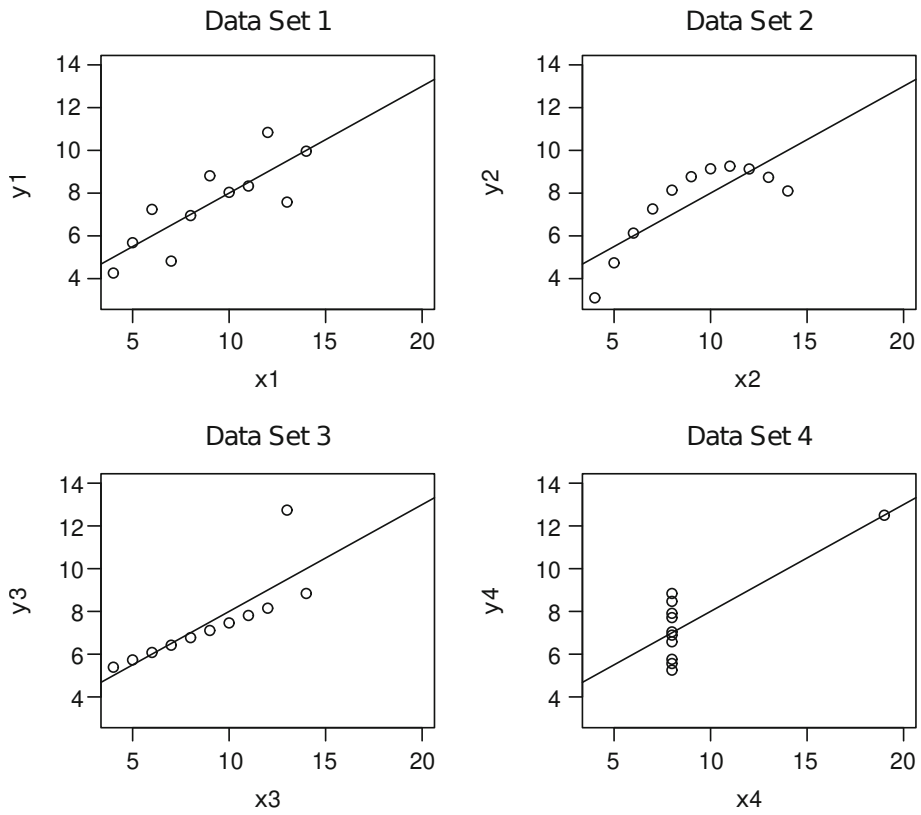

Figure 3.1: Plots of Anscombe's four data sets.

When a regression model is fitted to each data set, in each case the fitted regression model is

$$\hat{y} = 3.0 + 0.5x$$

The regression output for the four constructed data sets is identical (to two decimal places) in every respect:

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.0001     1.1247   2.667  0.02573 *
x1            0.5001     0.1179   4.241  0.00217 **

Residual standard error: 1.237 on 9 degrees of freedom
Multiple R-squared:  0.6665,    Adjusted R-squared:  0.6295
F-statistic: 17.99 on 1 and 9 DF,  p-value: 0.00217

            Estimate Std. Error t value Pr(>|t|)
(Intercept)    3.001      1.125   2.667  0.02576 *
x2             0.500      0.118   4.239  0.00218 **

Residual standard error: 1.237 on 9 degrees of freedom
Multiple R-squared:  0.6662,    Adjusted R-squared:  0.6292
F-statistic: 17.97 on 1 and 9 DF,  p-value: 0.002179

            Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.0025     1.1245   2.670  0.02562 *
x3            0.4997     0.1179   4.239  0.00218 **
```

```
Residual standard error: 1.236 on 9 degrees of freedom
Multiple R-squared:  0.6663,    Adjusted R-squared:  0.6292
F-statistic: 17.97 on 1 and 9 DF,  p-value: 0.002176

            Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.0017     1.1239   2.671  0.02559 *
x4            0.4999     0.1178   4.243  0.00216 **

Residual standard error: 1.236 on 9 degrees of freedom
Multiple R-squared:  0.6667,    Adjusted R-squared:  0.6297
F-statistic:    18 on 1 and 9 DF,  p-value: 0.002165
```

Looking at Figure 3.1 it is clear that a straight-line regression model is appropriate only for Data Set 1, since it is the only data set for which $\mathbb{E}(Y) = \beta_0 + \beta_1 x$ and $\text{Var}(Y) = \sigma^2$ seem reasonable assumptions. The second data set seems to have a curved rather than a straight-line relationship. The third data set has an extreme outlier that should be investigated. For the fourth data set, the slope of the regression line is solely determined by a single point, namely, the point with the largest X-value.

This example demonstrates that the numerical regression output should always be supplemented by an analysis to ensure that an appropriate model has been fitted to the data. In this case it is sufficient to look at the scatter plots in Figure 3.1 to determine whether an appropriate model has been fit. However, when we consider situations in which there is more than one predictor variable, we shall need some additional tools in order to check the appropriateness of the fitted model.

One of such tools is to plot residuals versus predictor and look for patterns. If no pattern is found then this indicates that the estimated regression model provides an adequate summary of the data, i.e., is a valid model. If a pattern is found then the shape of the pattern provides information on the functional dependence on the predictor (or predictors) that is missing from the model.

Figure 3.2 provides plots of residuals against X for each of Anscombe's four data sets. There is no discernible pattern for Data Set 1. This indicates that an appropriate model has been fit to the data. The residuals for Data Set 2 have a quadratic pattern. Hence there is need for a quadratic term to be added to the original straight-line regression model, i.e., $Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon$. The residuals for Data Set 3 indicate an outlier, and the residuals for Data Set 4 indicate a leverage point. If these were real-life data sets, one would need to investigate further the origin of these abnormalities.

3.2 Regression diagnostics

When fitting a regression model it is important to:

1. Determine whether the proposed regression model is a valid model (i.e., determine whether it provides an adequate fit to the data). This can be done with the help of **standardised residuals**, that we will define a bit later. Plotting standardised residuals enables us to assess visually whether the assumptions are being violated and point to what should be done to overcome these violations.

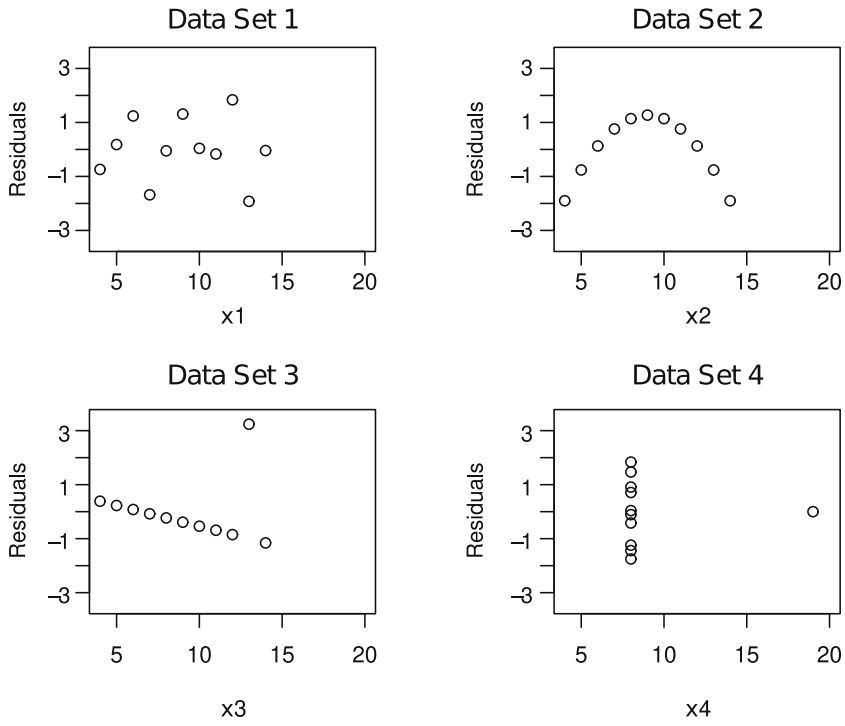

Figure 3.2: Residual plots for Anscombe's data sets.

2. Determine which (if any) of the data points have X-values that have an unusually large effect on the estimated regression model. Such points are called **leverage points**.
3. Determine which (if any) of the data points are **outliers**, that is, points which do not follow the pattern set by the bulk of the data, when one takes into account the given model.
4. If leverage points exist, determine whether each is a **bad leverage point**. If a bad leverage point exists we shall assess its influence on the fitted model.
5. Examine whether the assumption of constant variance of the errors is reasonable. If not, we shall look at how to overcome this problem.
6. If the sample size is small or prediction intervals are of interest, examine whether the assumption that the errors are normally distributed is reasonable.

We begin by looking at the second item of the above list, leverage points, as these will be needed in the explanation of standardised residuals.

3.2.1 Leverage points

Data points which exercise considerable influence on the fitted model are called leverage points. In other words, a leverage point is a point whose X-value is distant from the other X-values. To make things as simple as possible, we shall begin somewhat unrealistically, by describing leverage points as either "good" or "bad":

- A leverage point is a **bad leverage point** if its Y-value does not follow the pattern set by the other data points, see Figure 3.3(b). That is, a bad leverage point is a leverage point which is also an **outlier**.

- A leverage point is a **good leverage point** if it is not also an outlier, see Figure 3.3(a).

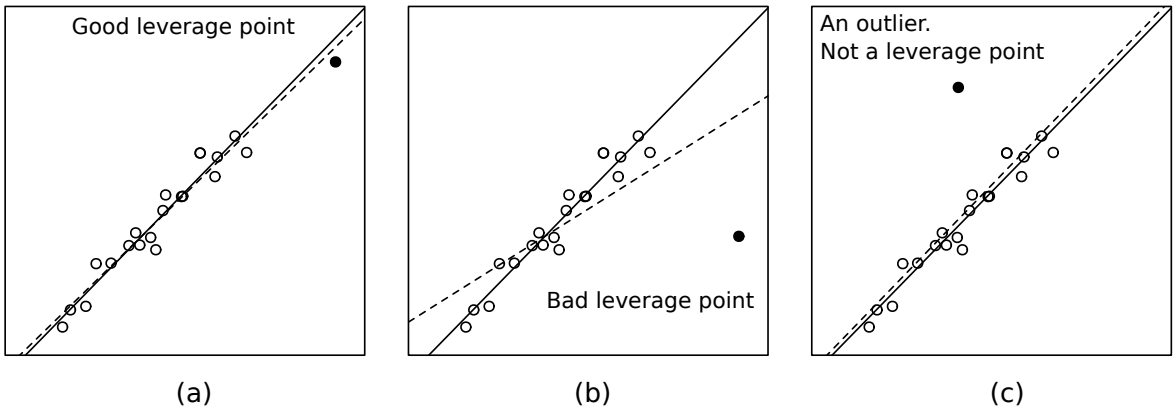

Figure 3.3: Leverage points.

Not that the indicated point in Figure 3.3(c) is not a leverage point, since its X-value is not distant from the other X-values. Since it is much further away from the fitted regression line than the other points, it is an outlier.

We would like to have a numerical rule that will identify leverage points. Recall that

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

where

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}, \qquad \hat{\beta}_1 = \sum_{j=1}^{n} c_j y_j, \qquad c_j = \frac{x_j - \bar{x}}{s_{xx}}.$$

We can thus rewrite $\hat{y}_i$ as

$$\begin{aligned}
\hat{y}_i &= \bar{y} - \hat{\beta}_1 \bar{x} + \hat{\beta}_1 x_i \\
&= \bar{y} + \hat{\beta}_1 (x_i - \bar{x}) \\
&= \frac{1}{n} \sum_{j=1}^{n} y_j + \sum_{j=1}^{n} \frac{x_j - \bar{x}}{s_{xx}} \cdot y_j \cdot (x_i - \bar{x}) \\
&= \sum_{j=1}^{n} \left(\frac{1}{n} + \frac{(x_i - \bar{x})(x_j - \bar{x})}{s_{xx}} \right) y_j \\
&= \sum_{j=1}^{n} h_{ij} y_j
\end{aligned}$$

where

$$h_{ij} = \frac{1}{n} + \frac{(x_i - \bar{x})(x_j - \bar{x})}{s_{xx}} \tag{3.1}$$

Notice that

$$\sum_{j=1}^{n} h_{ij} = \frac{n}{n} + \frac{(x_i - \bar{x}) \sum_{j=1}^{n} (x_j - \bar{x})}{s_{xx}} = 1 \tag{3.2}$$

We can predict the value $\hat{y}_i$ as

$$\hat{y}_i = h_{ii} y_i + \sum_{j \neq i} h_{ij} y_j$$

where

$$h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{s_{xx}}.$$

The term h_{ii} is commonly called the **leverage** (or the hat-value) of the i-th case. The term $(x_i - \bar{x})^2$ measures the distance x_i is away from the bulk of the x's, and so a high value of h_{ii} indicates a point distant from the other points. For instance, if $h_{ii} \approx 1$, then then other h_{ij} because of (3.2) are close to zero, and so

$$\hat{y}_i = 1 \times y_i + \text{other terms} \approx y_i.$$

In this situation, the estimated value, $\hat{y}_i$, will be close to the actual value, y_i, no matter what values of the rest of the data take. Notice also that h_{ii} depends only on the x's. Thus a point of high leverage (or a leverage point) can be found by looking at just the values of the x's and not at the values of the y's.

A popular rule, which we shall adopt, to classify x_i as a **point of high leverage** (i.e., a **leverage point**) in a simple linear regression model is if

$$h_{ii} > 2 \times \text{average}(h_{ii}) = 2 \times \frac{2}{n} = \frac{4}{n}.$$

It remains to determine if a leverage point is "good" or "bad". This can done by inspecting the value of its standardised residual, which we shall discuss in the next subsection.

Example 5: "Good" vs "bad" leverage

Statistician Peter Huber in 1981 constructed two data sets to illustrate "good" and "bad" leverage points. Both data sets have the same X-values, while the Y's, $YBad$ and $YGood$, only differ when $X = 10$, see Table 3.2.

Case	X	$YBad$	$YGood$	Leverage
1	-4	2.48	2.48	0.2897
2	-3	0.73	0.73	0.2359
3	-2	-0.04	-0.04	0.1974
4	-1	-1.44	-1.44	0.1744
5	0	-1.32	-1.32	0.1667
6	10	0	-11.4	0.9359

Table 3.2: Huber's data set.

The latter point is a leverage point: it is far way away from the other values of X and the values of $YBad$ and $YGood$ have very large effects on the respective least squares regression lines. Also notice that the leverage values for both sets are the same. Moreover, $h_{66} = 0.9359 > 2 \times \text{average}(h_{ii}) = 4/n = 4/6 = 0.67$. Thus, the point $X = 10$, is a point of high leverage (or a leverage point), while the other points have leverage values

much below the cut-off of 0.67. It is evident that the $X = 10$ point is a "good" leverage point for the second data set, however it is not evident that $X = 10$ is "bad" leverage point for the first data set. It fits very well the a quadratic regression model, the dashed regression line in Figure 3.4.

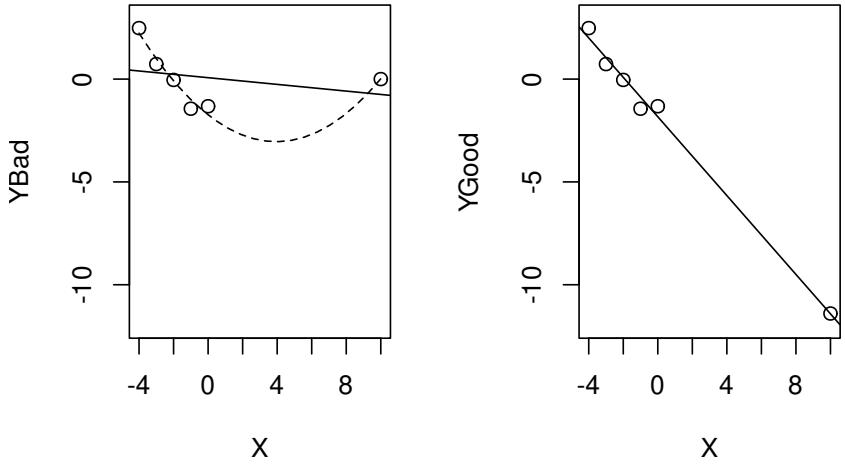

Figure 3.4: Plots of $YGood$ and $YBad$ against X with the fitted regression lines.

This example illustrates the fact that leverage points can not always be classified simply as "good" or "bad". There is a large gray area between these two extrema and a further analysis of the data needs to be conducted. For example, one may check if the data points of high leverage are unusual or different in some way from the rest of the data? If so, consider removing these points and refitting the model without them. You should also question the validity of the regression model that has been fitted? If so, consider trying a different model by including extra predictor variables (e.g., polynomial terms) or by transforming the response and/or predictor variables. We will discuss transformations further ahead.

3.2.2 Standardised residuals

We have discussed at the beginning of this section that residuals e_i can be used to detect any problems with the proposed model. However, as we shall show below, there is a complication that we need to consider, namely, that residuals do not have the same variance. In fact, we shall show that LSE of e_i have

$$\text{Var}(E_i) = \sigma^2(1-h_{ii})$$

where h_{ii} are given by (3.1),

$$h_{ij} = \frac{1}{n} + \frac{(x_i-\bar{x})(x_j-\bar{x})}{s_{xx}}.$$

Thus E_i's do not quite mimic the properties of ε_i's, since $\text{Var}(\varepsilon) = \sigma^2$, but can be thought of as their proxies. (Hence e_i's are also called the **crude residuals**.)

The problem of E_i's having different variances can be overcome by a standardisation,

$$\frac{E_i}{\sqrt{\sigma^2(1-h_{ii})}} \sim N(0,1).$$

Consequently, we introduce the **standardised residuals**, r_i, by

$$r_i = \frac{e_i}{\sqrt{\hat{\sigma}^2(1-h_{ii})}}.$$

Their advantage over the crude residuals is explained below.

When points of **high leverage** exist, instead of looking at crude residual plots, it is generally more informative to look at plots of standardised residuals since plots of the residuals will have nonconstant variance even if the errors have constant variance. (When points of high leverage do not exist, there is generally little difference in the patterns seen in plots of residuals when compared with those in plots of standardised residuals.) The other advantage of standardised residuals is that they immediately tell us how many estimated standard deviations any point is away from the fitted regression model. For example, suppose that the 6th point has a standardised residual of 4.3, then this means that the 6th point is an estimated 4.3 standard deviations away from the fitted regression line. If the errors are normally distributed, then observing a point 4.3 standard deviations away from the fitted regression line is highly unusual. Such a point would commonly be referred to as an outlier and as such it should be investigated. It is a common practice of labelling points as **outliers** in small- to moderate-size data sets if the standardised residual for the point falls outside the interval from $\mathbf{-2}$ **to** $\mathbf{2}$. In very large data sets the interval is from $\mathbf{-4}$ **to** $\mathbf{4}$. (Otherwise, many points will be flagged as potential outliers.) Identification and examination of any outliers is a key part of regression analysis.

Recall that a bad leverage point is a leverage point which is also an outlier. Thus, a bad leverage point is a leverage point whose standardised residual falls outside the interval from -2 to 2 (or -4 to 4). On the other hand, a good leverage point is a leverage point whose standardised residual falls inside the interval from -2 to 2 (or -4 to 4).

There is a small amount of correlation present in standardised residuals, even if the errors are independent. In fact it can be shown that

$$\text{Cov}(E_i, E_j) = -\sigma^2 h_{ij}, \qquad \text{Corr}(E_i, E_j) = \frac{-h_{ij}}{\sqrt{(1-h_{ii})(1-h_{jj})}} \qquad (i \neq j).$$

However, the size of the correlations inherent in the least squares residuals are generally so small in situations in which correlated errors is an issue (e.g., data collected over time) that they can be effectively ignored in practice.

To compute the variance $\text{Var}(E_i)$ we recall that $\hat{Y}_i = h_{ii}Y_i + \sum_{j\neq i} h_{ij}Y_j$ with h_{ij} given by (3.1). Thus

$$E_i = Y_i - \hat{Y}_i = Y_i - h_{ii}Y_i - \sum_{j\neq i} h_{ij}Y_j = (1-h_{ii})Y_i - \sum_{j\neq i} h_{ij}Y_j$$

and

$$\begin{aligned}
\text{Var}(E_i) &= \text{Var}\left((1-h_{ii})Y_i - \sum_{j\neq i} h_{ij}Y_j\right) \\
&= (1-h_{ii})^2\sigma^2 + \sum_{j\neq i} h_{ij}^2\sigma^2 \\
&= \sigma^2\left(1 - 2h_{ii} + h_{ii}^2 + \sum_{j\neq i} h_{ij}^2\right) = \sigma^2\left(1 - 2h_{ii} + \sum_{j} h_{ij}^2\right).
\end{aligned}$$

Now notice that

$$\begin{aligned}\sum_{j=1}^{n} h_{ij}^2 &= \sum_{j=1}^{n}\left(\frac{1}{n}+\frac{(x_i-\bar{x})(x_j-\bar{x})}{s_{xx}}\right)^2 \\ &= \frac{1}{n}+2\sum_{j=1}^{n}\frac{1}{n}\times\frac{(x_i-\bar{x})(x_j-\bar{x})}{s_{xx}}+\sum_{j=1}^{n}\frac{(x_i-\bar{x})^2(x_j-\bar{x})^2}{s_{xx}^2} \\ &= \frac{1}{n}+0+\frac{(x_i-\bar{x})^2}{s_{xx}} \\ &= h_{ii}\end{aligned}$$

so that

$$\text{Var}(E_i) = \sigma^2(1-2h_{ii}+h_{ii}) = \sigma^2(1-h_{ii})$$

which shows that standardised residuals indeed have a non-constant variance.

Example 6: US Treasure bonds data

This example illustrates that a relatively small number of outlying points can have a relatively large effect on the fitted model. We shall look at effect of removing these outliers and refitting the model. The example is created by the statistician Andrew Siegel in 1997 based on the data of US Treasury bond prices published in the November 9, 1988 edition of The Wall Street Journal (p. C19) consisting of 35 cases.[a]

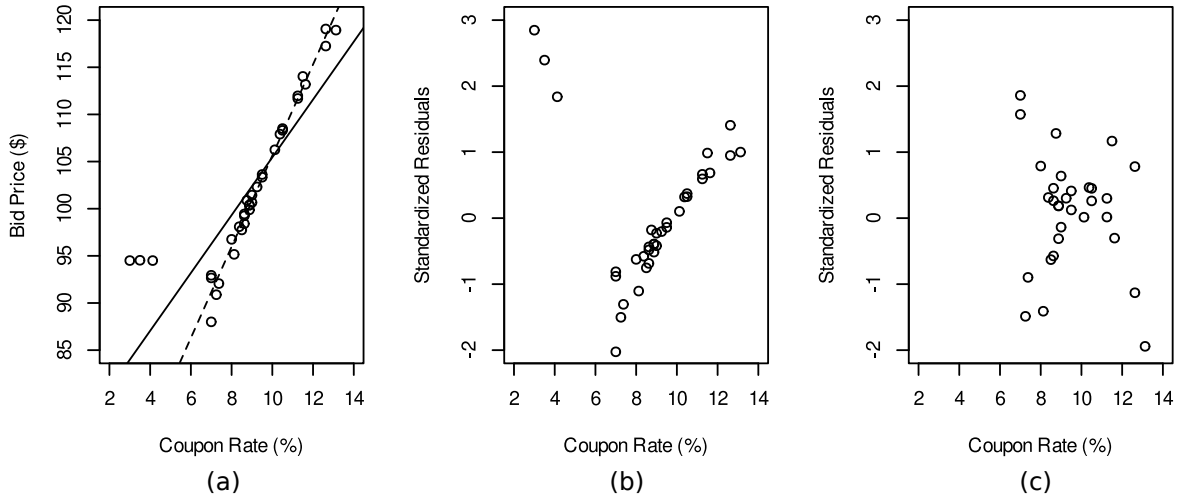

Figure 3.5: Regression diagnostics for bonds data.

The data are plotted in Figure 3.5(a). The solid line is the estimated regression line. We see that the regression model does not describe the data well. The three points on the left of the scatter plot obviously stand out. The regression line is dragged away from the bulk of the points towards these three points. Recall that the rule for simple linear regression for classifying a point as a leverage point is $h_{ii} > 4/n = 4/35 = 0.11$. For the three points in question have leverage values equal to 0.153, 0.218 and 0.0.187. Their standardised residuals are 1.84, 2.85 and 2.39 (Figure 3.5(b)). Thus these points can be classified as bad leverage points. (Strictly speaking only the first two points are bad leverage points, but after removing the first two points and repeating the same regression

diagnostic we would find that the third point is also a bad leverage point.)
Upon removing the three bad leverage points and re-estimating the regression model (the dashed line in Figure 3.5(a)) we get a very good fit. Moreover, the standardised residuals are now more randomly distributed (Figure 3.5(c)). The three points that were removed correspond to "flower" bonds, which have definite tax advantages compared to the other bonds. Given this information, it is clear that there will be different relationship between coupon rate and bid price for "flower" bonds. Given a low coupon rate the bid price is higher for "flower" bonds than regular bonds. Thus, a reasonable strategy is to remove the cases corresponding to "flower" bonds from the data and only consider regular bonds. An alternative way to cope with points such as "flower" bonds is to add one or more dummy variables to the regression model. Dummy variables will be discussed in further sections.

[a] US Treasury bonds are among the least risky investments, in terms of the likelihood of your receiving the promised payments. In addition to the primary market auctions by the Treasury, there is an active secondary market in which all outstanding issues can be traded. You would expect to see an increasing relationship between the coupon of the bond, which indicates the size of its periodic payment (twice a year), and the current selling price. The ... data set of coupons and bid prices [are] for US Treasury bonds maturing between 1994 and 1998... The bid prices are listed per "face value" of $100 to be paid at maturity. Half of the coupon rate is paid every six months. For example, the first one listed pays $3.50 (half of the 7% coupon rate) every six months until maturity, at which time it pays an additional $100.

Remark 3.1. (i) Data points should not be routinely deleted from an analysis just because they do not fit the model. Outliers and bad leverage points are signals, flagging potential problems with the model.
(ii) Outliers often point out an important feature of the problem not considered before. They may point to an alternative model in which the points are not an outlier. In this case it is then worth considering fitting an alternative model. □

3.2.3 Assessing the influence of certain cases

One or more cases can strongly control or influence the estimated regression model. For example, in the previous example on US Treasury bond prices, three cases dramatically influenced the model. In this subsection we look at summary statistics that measure the influence of a single case on the estimated regression model.

We shall use the notation where subscript (i) means that the i-th case has been deleted from the fit. In other words, the fit is then based on the remaining $n-1$ cases indexed 1, 2, ..., $i-1$, $i+1$, ... , n. Thus, $\hat{y}_{j(i)}$ denotes the j-th fitted value based on the fit obtained when the i-th case has been deleted from the fit.

An American statistician Ralph Dennis Cook in 1977 proposed a widely used measure of the influence of individual cases which in the case of the simple linear regression is given by

$$D_i = \frac{\sum_{j=1}^{n}(\hat{y}_{j(i)} - \hat{y}_j)^2}{2\hat{\sigma}^2}.$$

It can be shown that

$$D_i = \frac{r_i^2}{2} \cdot \frac{h_{ii}}{1-h_{ii}}$$

where r_i is the i-th standardised residual and h_{ii} is the i-th leverage value, both of which were defined earlier. Thus, Cook's distance can be obtained by multiplying two quantities, namely, the square of the i-th standardised residual divided by two and a monotonic function that increases as the i-th leverage value increases. The first quantity measures the extent to which the i-th case is outlying while the second quantity measures the leverage of the i-th case. Thus, a large value of D_i may be due to a large value of r_i, a large value of h_{ii} or both. If the largest value of D_i is substantially less than one, deletion of a case will not change the estimate by much. A recommend rough cut-off for noteworthy values of D_i for simple linear regression is $4/(n-2)$. In practice, it is important to look for gaps in the values of Cook's distance and not just whether values exceed the suggested cut-off.

Example 7: US Treasure bonds data

Figure 3.6 contains a plot of Cook's distance against Coupon Rate, x, for the US Treasury bond prices regression model. Marked on the plot as a horizontal dashed line is the cut-off value of $4/(35-2) = 0.121$. The earlier discussed points (cases 13, 35, 4) exceed this value and as such are worthy of investigation. Note that the Cook's distance for point 13 exceeds 1, which means it deserves special attention.

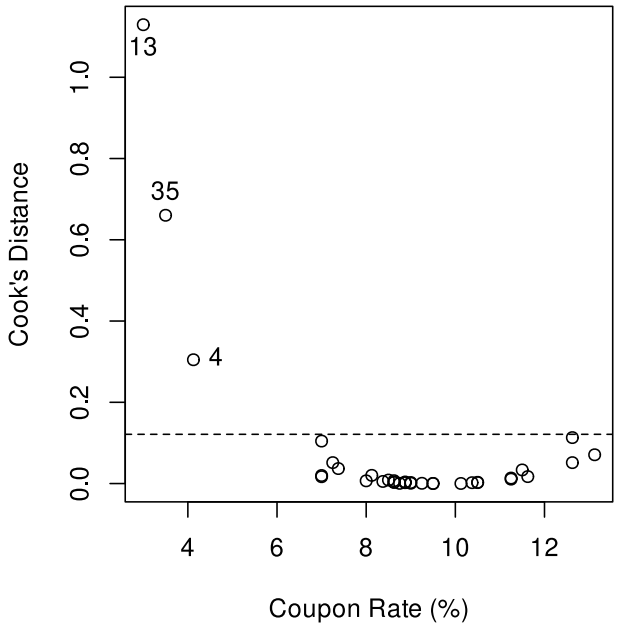

Figure 3.6: A plot of Cook's distance against Coupon Rate.

3.2.4 Normality of the errors

The assumption of normality of the errors is needed in small samples for the validity of t-distribution based hypothesis tests and confidence intervals and for all sample sizes for prediction intervals. This assumption is generally checked by looking at the distribution of the residuals or standardised residuals.

Recall that $e_i = y_i - \hat{y}_i$ and $\hat{y}_i = \sum_{j=1}^{n} h_{ij} y_j$. Then

$$\begin{aligned} e_i &= y_i - \sum_{j=1}^{n} h_{ij} y_j \\ &= \beta_0 + \beta_1 x_i + \varepsilon_i - \sum_{j=1}^{n} h_{ij}(\beta_0 + \beta_1 x_j + \varepsilon_j) \\ &= \beta_0 + \beta_1 x_i + \varepsilon_i - \beta_0 - \beta_1 x_i - \sum_{j=1}^{n} h_{ij}\varepsilon_j \\ &= \varepsilon_i - \sum_{j=1}^{n} h_{ij}\varepsilon_j \end{aligned}$$

since $\sum_{j=1}^{n} h_{ij} = 1$ and

$$\sum_{j=1}^{n} h_{ij} x_j = \sum_{j=1}^{n} \left(\frac{x_j}{n} + \frac{(x_i - \bar{x})(x_j - \bar{x})}{s_{xx}} \right) = \bar{x} + \frac{(x_i - \bar{x}) s_{xx}}{s_{xx}} = x_i.$$

Thus, the i-th least squares residual is equal to ε_i minus a weighted sum of all of the ε_i's. In small to moderate samples, the second term in the last equation can dominate the first and the residuals can look like they come from a normal distribution even if the errors do not. As n increases, the second term in the last equation has a much smaller variance than that of the first term and as such the first term dominates the last equation. This implies that for large samples the residuals can be used to assess normality of the errors.

In spite of what we have just discovered, a common way to assess normality of the errors is to look at what is commonly referred to as a **normal probability plot** or a **normal Q-Q plot** of the standardised residuals. A normal probability plot of the standardised residuals is obtained by plotting the ordered standardised residuals on the vertical axis against the expected order statistics from a standard normal distribution on the horizontal axes. If the resulting plot produces points "close" to a straight line then the data are said to be consistent with that from a normal distribution. On the other hand, departures from linearity provide evidence of non-normality.

3.2.5 Constant variance

A crucial assumption in any regression analysis is that errors ε_i have a **constant variance**. This is necessary for all the inferential tools (i.e., P-values, confidence intervals, prediction intervals, etc.) to be valid. When the variance is found to be nonconstant, there are two main methods for overcoming this: transformations and weighted least squares. We will discuss the former method only.[2] We will illustrate the constant variance analysis with a couple of examples.

[2]For the method of weighted least squares see Chapter 4 in Simon J. Sheather, *A Modern Appraoch to Regression with R*, Springer 2009.

Example 8: Manufacturer production data

Recall the example on the timing of production runs for which we fit a straight-line regression model to run time from run size. Figure 3.7 provides diagnostic plots produced by R when the function `plot()` is applied to a linear model. The top right plot is a normal Q-Q plot. The bottom right plot of standardised residuals against leverage enables one to readily identify any "bad" leverage points. We shall see shortly that the bottom left-hand plot provides diagnostic information about whether the variance of the error term appears to be constant.

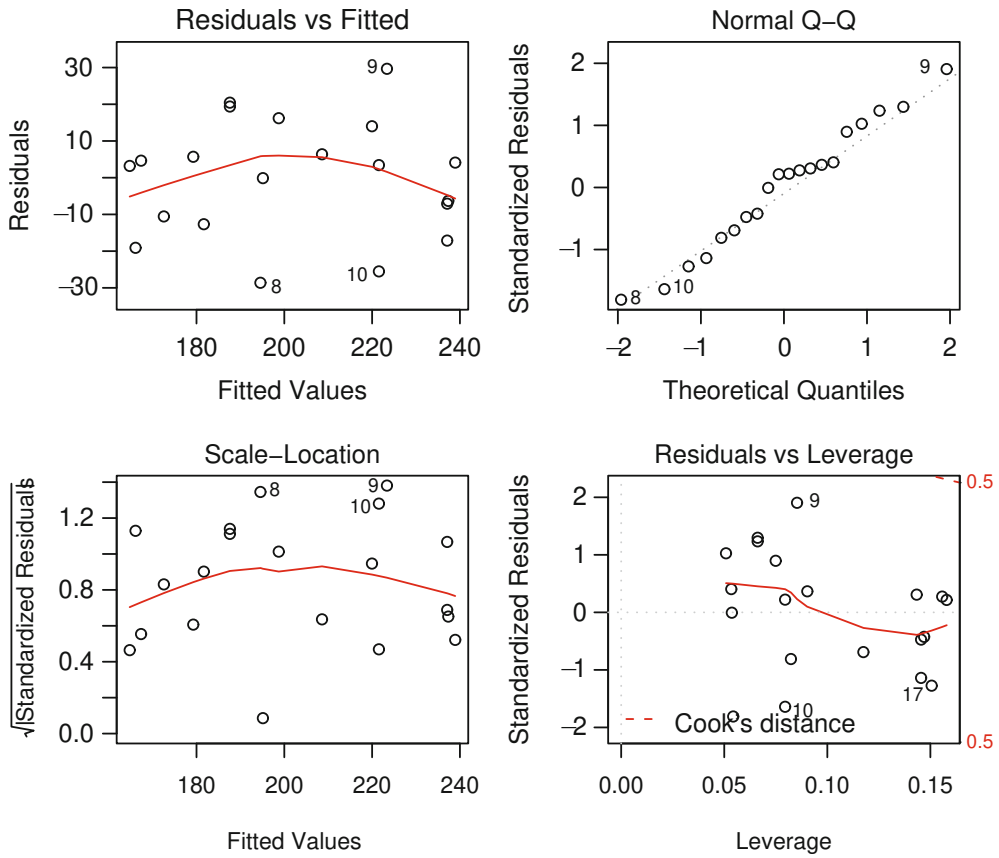

Figure 3.7: Regression diagnostics of a SLR model for the production data.

Example 9: Office cleaning data

A building maintenance company is planning to submit a bid on a contract to clean corporate offices scattered throughout an office complex. The costs incurred by the maintenance company are proportional to the number of cleaning crews needed for this task. Recent data are available for the number of rooms that were cleaned by varying numbers of crews. For a sample of 53 days, records were kept of the number of crews used and the number of rooms that were cleaned by those crews. We want to build a regression model to estimate relationship between the number of rooms cleaned and the number of crews to be able to predict the number of rooms that can be cleaned by 4 crews and by 16 crews. Regression diagnostics is shown in Figure 3.8. It is evident from this figure

that variability in the standardised residuals tends to increase with the number of crews, i.e., there is evidence that the error variance is not constant. Hence the choice of the model must be re-evaluated.

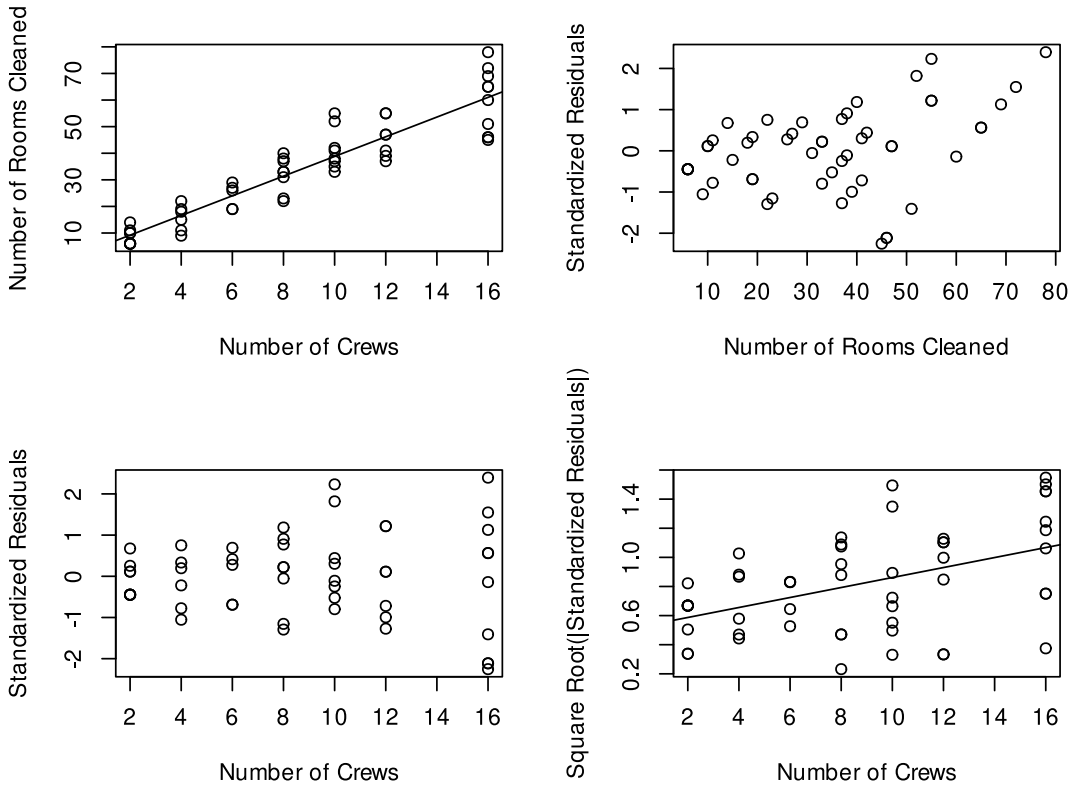

Figure 3.8: Regression diagnostics of a SLR model for the room cleaning data.

3.3 Transformations

When nonconstant variance exists, it is often possible to transform one or both of the regression variables to produce a model in which the error variance is constant. We will provide examples of such scenarios.

3.3.1 Square root transformation

Consider again the room cleaning data from the previous subsection. Count data are often modelled using the Poisson distribution. Suppose that Y follows a Poisson distribution with mean and variance λ, i.e., $\mathrm{P}(Y = y) = \lambda^y e^{-\lambda}/y!$ and $\mathbb{E}(Y) = \mathrm{Var}(Y) = \lambda$. In such a case, the appropriate transformation of Y for stabilizing variance is square root. We shall try the square root transformation for both the predictor and response variables. (*When both Y and X are measured in the same units then it is often natural to consider the same transformation for both X and Y.*) We thus fit the model

$$\sqrt{Y} = \beta_0 + \beta_1 \sqrt{x} + \varepsilon.$$

Figure 3.9 shows regression diagnostics of the transformed data. The standardised residuals do not display the funnel shape, as they did for the nontransformed data. We conclude that

taking the square root of both variables has stabilized the variance of the random errors and hence produced a valid model.

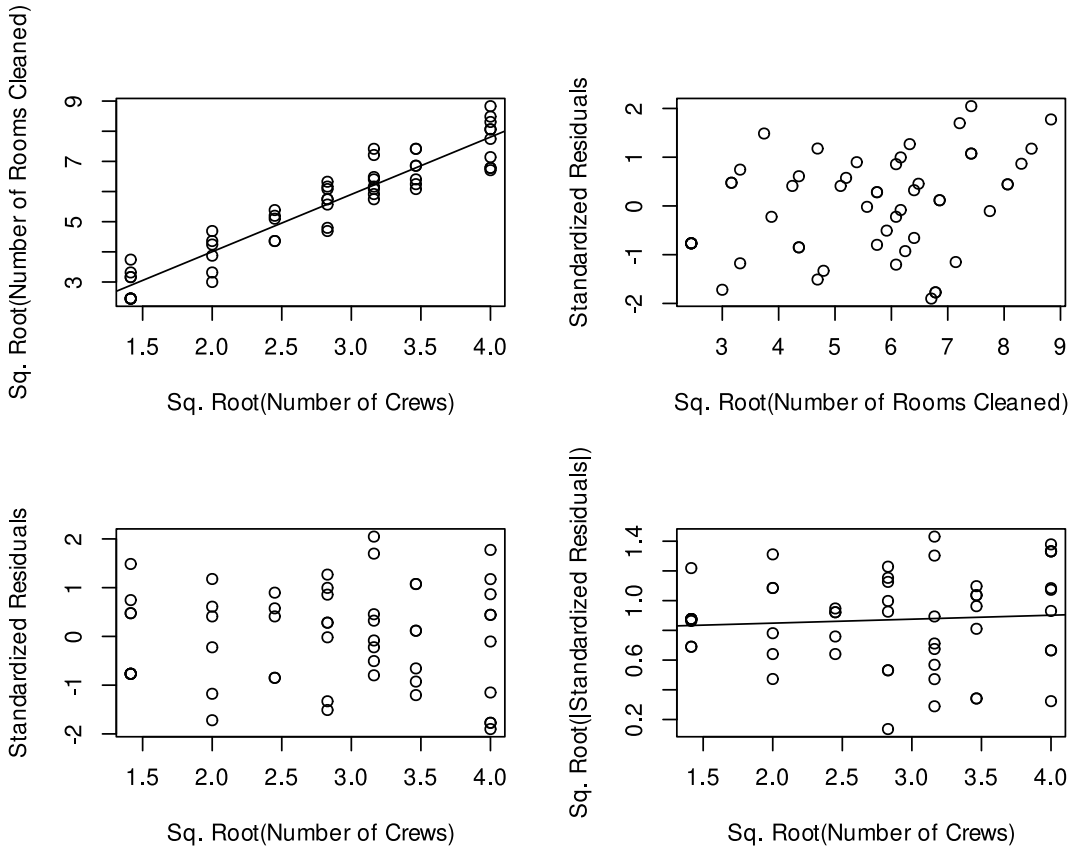

Figure 3.9: Regression diagnostics of a SLR model for transformed room cleaning data.

3.3.2 Log transformation

We consider data of maximum salary for 495 nonunionized job classes in a Midwestern Government unit in US in 1986. We shall focus on developing a regression model to predict MaxSalary, the maximum salary (in $) for employees in this job class, using just one predictor variable, Score, the score for job class based on difficulty, skill level, training requirements and level of responsibility as determined by a consultant to the government unit. (This example is taken from Weisberg (2005).)

We begin by considering a simple linear regression model for the salary data,

$$\text{MaxSalary} = \beta_0 + \beta_1 \text{Score} + \varepsilon.$$

Figure 3.10 shows a plot of the data and various plots of the standardised residuals.

There is a clear evidence of nonlinearity and nonconstant variance in these plots. We shall try log-transforming the response variable, i.e., we fit the model

$$\log(\text{MaxSalary}) = \beta_0 + \beta_1 \text{Score} + \varepsilon.$$

Figure 3.11 shows a plot of the partially transformed data and various plots of the standardised residuals. The relationship between the response and predictor variables is more linear now

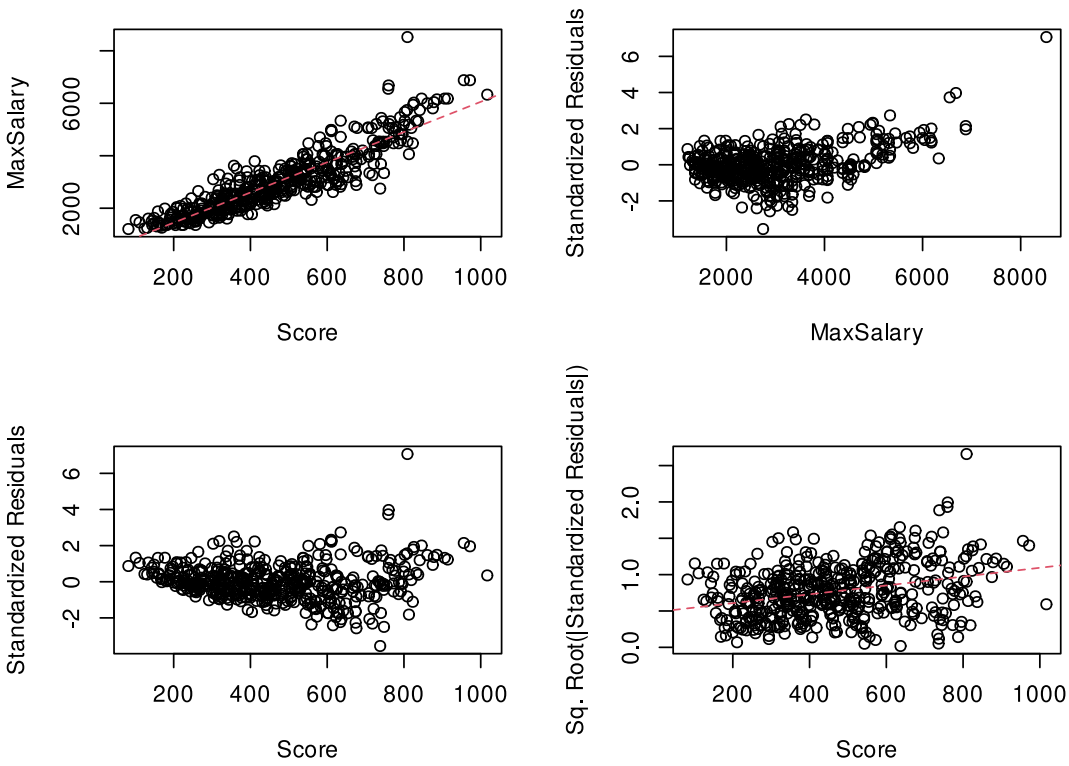

Figure 3.10: Regression diagnostics of the salary data.

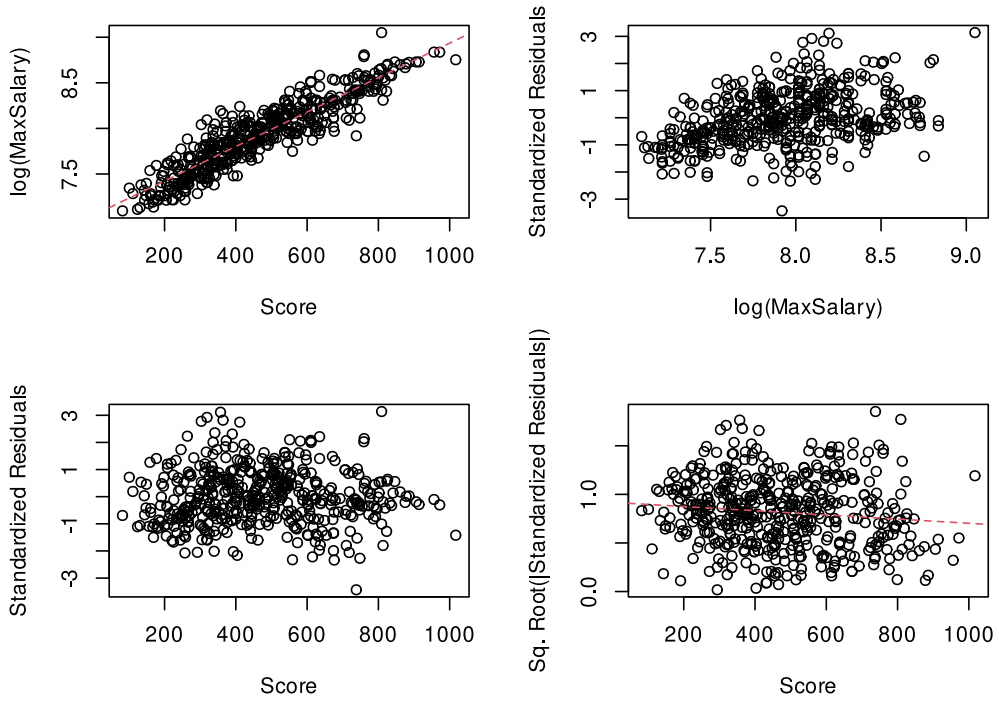

Figure 3.11: Regression diagnostics of the partially transformed salary data.

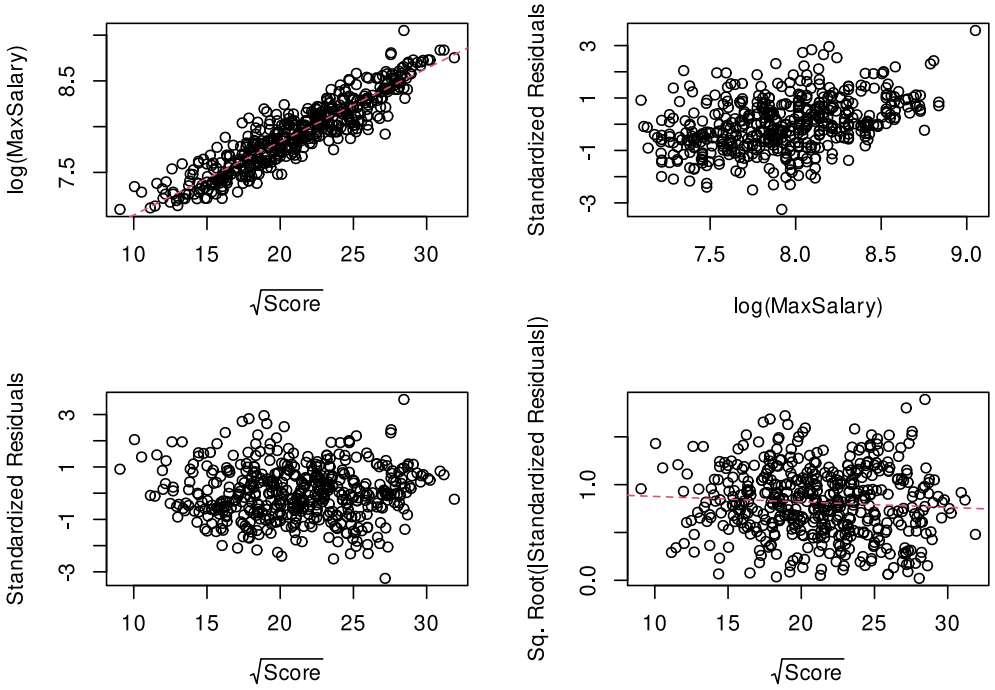

Figure 3.12: Regression diagnostics of the fully transformed salary data.

and the standardised residuals are more monotonic. However, some signs of nonlinearity and nonconstant variance still remain. Hence we need to further transform the data. We try taking a square root of the predictor variable, i.e., we fit the model

$$\log(\text{MaxSalary}) = \beta_0 + \beta_1 \sqrt{\text{Score}} + \varepsilon.$$

Figure 3.12 shows a plot of the fully transformed data and various plots of the standardised residuals. The signs of nonlinearity and nonconstant variance are now even less visible and we conclude that the third model is our preferred model.

3.3.3 Power transformation and the Box-Cox method

Statisticians Box and Cox in 1964 considered a modified family of power transformations

$$Y^{(\lambda)} = \begin{cases} \text{gm}(Y)^{1-\lambda}(Y^\lambda - 1)/\lambda & \text{if } \lambda \neq 0 \\ \text{gm}(Y)\log(Y) & \text{if } \lambda = 0 \end{cases}$$

where

$$\text{gm}(Y) = \prod_{i=1}^{n} Y_i^{1/n} = \exp\left(\frac{1}{n}\sum_{i=1}^{n}\log(Y_i)\right)$$

is the geometric mean of Y. The Box-Cox method is based on the notion that for some value of λ the transformed version of Y, namely, $Y^{(\lambda)}$, is normally distributed. Likelihood methods can then be used to find the wanted value of λ.

As an example, we consider a data set consisting of 250 data points (from Sheather (2009)). The estimated simple linear regression model for this data is shown in Figure 3.13. It is evident that there is a power dependence between Y and x. Thus a better model for this data should

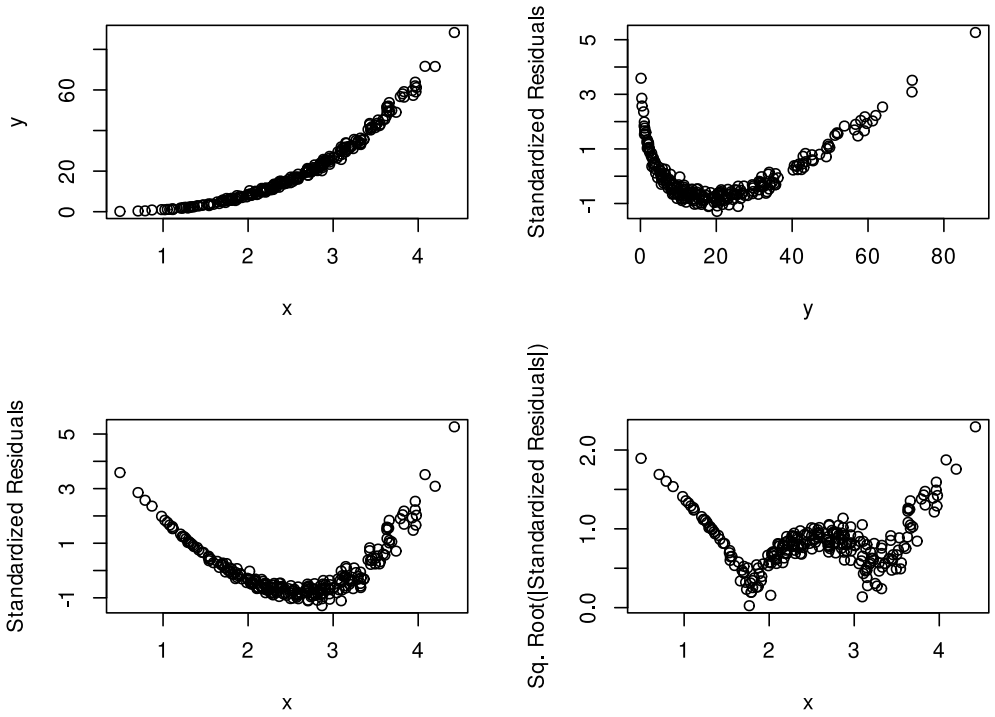

Figure 3.13: Regression diagnostics for the original data.

be $Y^{\lambda} = \beta_0 + \beta_1 x + \varepsilon$ for some λ. Log-likelihood for the Box-Cox transformation is shown in Figure 3.14. The value of λ that maximizes the log-likelihood and 95% confidence limits for λ are marked on each plot. This shows that $\lambda = 0.333$ provides the closest fit to the data. Upon transforming the data we get a very good fit shown in Figure 3.15. We conclude that the right model is $Y^{1/3} = \beta_0 + \beta_1 x + \varepsilon$.

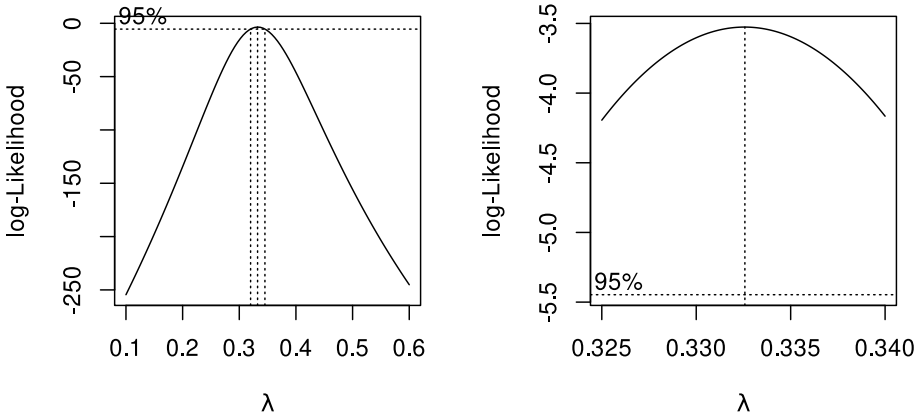

Figure 3.14: Log-likelihood for the Box-Cox transformation method.

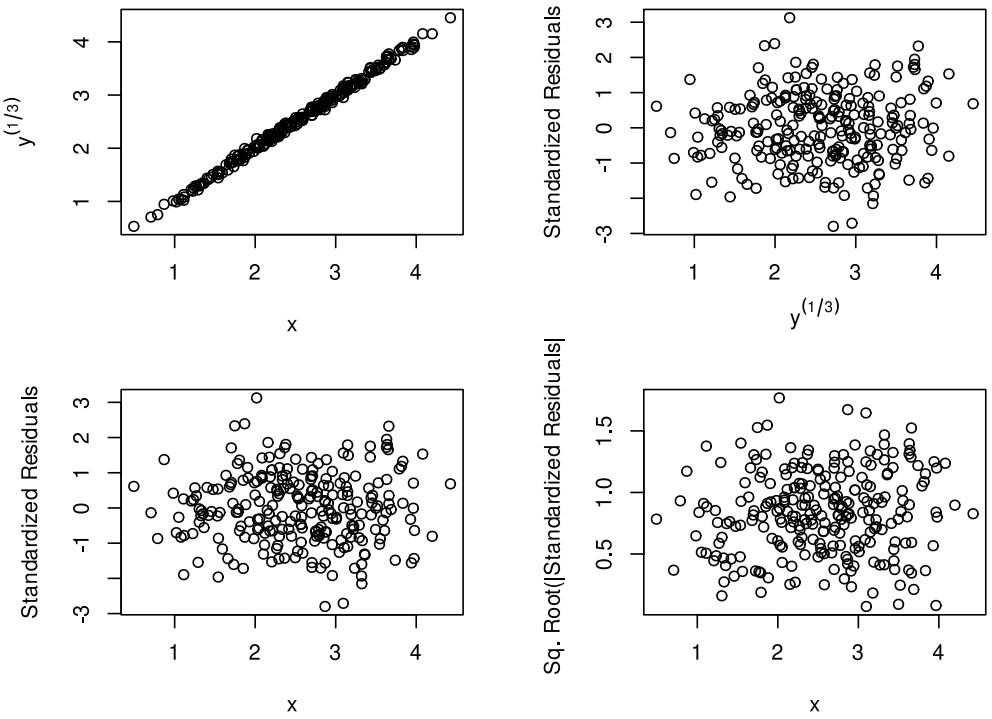

Figure 3.15: Regression diagnostics for the transformed data.

4 Matrix approach to simple linear regression

In this section we will discuss a matrix approach to the simple linear regression model. This will help us to transition to the multiple linear regression models.

Note that any set of linear equations can be re-written in a matrix and vector form. For the simple linear regression model we have n equations,

$$\begin{aligned} y_1 &= \beta_0 + \beta_1 x_1 + \varepsilon_1, \\ y_2 &= \beta_0 + \beta_1 x_2 + \varepsilon_2, \\ &\vdots \qquad\qquad \vdots \\ y_n &= \beta_0 + \beta_1 x_n + \varepsilon_n. \end{aligned}$$

We can write these equations in matrix formulation as

$$\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{4.1}$$

where

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \qquad X = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}, \qquad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}, \qquad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}.$$

The matrix X is known as the **design matrix**. This formulation (4.1) is usually called the **General Linear Model**.

4.1 Vectors of random variables

Vectors $\boldsymbol{y}$ and $\boldsymbol{\varepsilon}$ in equation (4.1) are random vectors as their elements are random variables. Below we show some properties of random vectors.

The expected value of a random vector is the vector of the respective expected values. That is, for a random vector $\boldsymbol{z} = (z_1, \ldots, z_n)^T$ we write

$$\mathbb{E}(\boldsymbol{z}) = \mathbb{E}\begin{bmatrix} z_1 \\ z_2 \\ \vdots \\ z_n \end{bmatrix} = \begin{bmatrix} \mathbb{E}(z_1) \\ \mathbb{E}(z_2) \\ \vdots \\ \mathbb{E}(z_n) \end{bmatrix}.$$

We have analogous properties of the expectation for random vectors as for single random variables. Namely, for a random vector $\boldsymbol{z}$, a constant scalar a, a constant vector $\boldsymbol{b}$ and for matrices of constants A and B we have

$$\begin{aligned} \mathbb{E}(a\boldsymbol{z} + \boldsymbol{b}) &= a\,\mathbb{E}(\boldsymbol{z}) + \boldsymbol{b}, \\ \mathbb{E}(A\boldsymbol{z}) &= A\mathbb{E}(\boldsymbol{z}), \\ \mathbb{E}(\boldsymbol{z}^T B) &= \mathbb{E}(\boldsymbol{z})^T B. \end{aligned}$$

Variances and covariances of the random variables z_i are put together to form the so-called variance-covariance (dispersion) matrix,

$$\mathrm{Var}(\boldsymbol{z}) = \begin{bmatrix} \mathrm{Var}(z_1) & \mathrm{Cov}(z_1, z_2) & \cdots & \mathrm{Cov}(z_1, z_n) \\ \mathrm{Cov}(z_2, z_1) & \mathrm{Var}(z_2) & \cdots & \mathrm{Cov}(z_2, z_n) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}(z_n, z_1) & \mathrm{Cov}(z_n, z_2) & \cdots & \mathrm{Var}(z_n) \end{bmatrix}.$$

The dispersion matrix has the following properties:

(a) The matrix $\mathrm{Var}(\boldsymbol{z})$ is symmetric since $\mathrm{Cov}(z_i, z_j) = \mathrm{Cov}(z_j, z_i)$.

(b) For mutually uncorrelated random variables the matrix is diagonal, since $\mathrm{Cov}(z_i, z_j) = 0$ for all $i \neq j$.

(c) It can be written as $\mathrm{Var}(\boldsymbol{z}) = \mathbb{E}\big[(\boldsymbol{z} - \mathbb{E}(\boldsymbol{z}))(\boldsymbol{z} - \mathbb{E}(\boldsymbol{z}))^T\big]$.

(d) For a transformed variable $\boldsymbol{u} = A\boldsymbol{z}$ we have $\mathrm{Var}(\boldsymbol{u}) = A\,\mathrm{Var}(\boldsymbol{z})A^T$.

To see why (c) is true, denote $\boldsymbol{\mu} = \mathbb{E}(\boldsymbol{z})$. Then

$$\mathbb{E}\big[(\boldsymbol{z} - \mathbb{E}(\boldsymbol{z}))(\boldsymbol{z} - \mathbb{E}(\boldsymbol{z}))^T\big] = \mathbb{E}\left[\begin{bmatrix} z_1 - \mu_1 \\ z_2 - \mu_2 \\ \vdots \\ z_n - \mu_n \end{bmatrix} (z_1 - \mu_1, z_2 - \mu_2, \ldots, z_n - \mu_n)\right]$$

$$= \begin{bmatrix} \mathbb{E}((z_1 - \mu_1)^2) & \mathbb{E}((z_1 - \mu_1)(z_2 - \mu_2)) & \cdots & \mathbb{E}((z_1 - \mu_1)(z_n - \mu_n)) \\ \mathbb{E}((z_2 - \mu_2)(z_1 - \mu_1)) & \mathbb{E}((z_2 - \mu_2)^2) & \ldots & \mathbb{E}((z_2 - \mu_2)(z_n - \mu_n)) \\ \vdots & \vdots & \ddots & \vdots \\ \mathbb{E}((z_n - \mu_n)(z_1 - \mu_1)) & \mathbb{E}((z_n - \mu_n)(z_2 - \mu_2)) & \ldots & \mathbb{E}((z_n - \mu_n)^2) \end{bmatrix}$$

$$= \mathrm{Var}(\boldsymbol{z}).$$

To see (d) we use the notation of (c):

$$\begin{aligned}\mathrm{Var}(\boldsymbol{u}) &= \mathbb{E}\left[(\boldsymbol{u}-\mathbb{E}(\boldsymbol{u}))(\boldsymbol{u}-\mathbb{E}(\boldsymbol{u}))^T\right]\\ &= \mathbb{E}\left[(A\boldsymbol{z}-A\boldsymbol{\mu})(A\boldsymbol{z}-A\boldsymbol{\mu})^T\right]\\ &= \mathbb{E}\left[A(\boldsymbol{z}-\boldsymbol{\mu})(\boldsymbol{z}-\boldsymbol{\mu})^T A^T\right]\\ &= A\mathbb{E}\left[(\boldsymbol{z}-\boldsymbol{\mu})(\boldsymbol{z}-\boldsymbol{\mu})^T\right]A^T = A\mathrm{Var}(\boldsymbol{z})A^T.\end{aligned}$$

Note that the property (c) gives the expression for the dispersion matrix of a random vector analogous to the expression for the variance of a single random variable, that is

$$\mathrm{Var}(\boldsymbol{z}) = \mathbb{E}(\boldsymbol{z}\boldsymbol{z}^T) - \boldsymbol{\mu}\boldsymbol{\mu}^T.$$

4.2 Derivatives in the matrix form

Let $\boldsymbol{z} = (z_1,\ldots,z_r)^T$ and let $f(z_1,\ldots,z_r)$ be a function of $\boldsymbol{z}$. We define

$$\frac{\partial f(z_1,\ldots,z_r)}{\partial \boldsymbol{z}} = \begin{pmatrix} \frac{f(z_1,\ldots,z_r)}{\partial z_1} \\ \vdots \\ \frac{\partial f(z_1,\ldots,z_r)}{\partial z_n} \end{pmatrix}.$$

Then for any vector $\boldsymbol{a} = (a_1,\ldots,a_r)^T$ we have

$$\frac{\partial \boldsymbol{a}^T\boldsymbol{z}}{\partial \boldsymbol{z}} = \frac{\partial(a_1z_1+\ldots+a_rz_r)}{\partial \boldsymbol{z}} = (a_1,\ldots,a_r)^T = \boldsymbol{a}.$$

If M is a square $r\times r$ matrix then

$$\frac{\partial \boldsymbol{z}^T M\boldsymbol{z}}{\partial \boldsymbol{z}} = (M+M^T)\boldsymbol{z}.$$

Indeed, let m_{ij} represent the ij-th element of M. Now $(M+M^T)\boldsymbol{z}$ is a column-vector with the k-th element given by $\sum_{i=1}^r m_{ki}z_i + \sum_{i=1}^r m_{ik}z_i$. Hence we need to show that

$$\frac{\partial \boldsymbol{z}^T M\boldsymbol{z}}{\partial z_k} = \sum_{i=1}^r m_{ki}z_i + \sum_{i=1}^r m_{ik}z_i.$$

From the product rule,

$$\begin{aligned}\frac{\partial \boldsymbol{z}^T M\boldsymbol{z}}{\partial z_k} &= \boldsymbol{z}^T\frac{\partial(M\boldsymbol{z})}{\partial z_k} + \frac{\partial \boldsymbol{z}^T}{\partial z_k}M\boldsymbol{z}\\ &= (z_1,\ldots,z_r)\begin{pmatrix}\partial_{z_k}\sum_{i=1}^r m_{1i}z_i\\ \vdots\\ \partial_{z_k}\sum_{i=1}^r m_{ri}z_i\end{pmatrix} + (0,\ldots,0,1,0,\ldots,0)\begin{pmatrix}\sum_{i=1}^r m_{1i}z_i\\ \vdots\\ \sum_{i=1}^r m_{1i}z_i\end{pmatrix}\\ &= (z_1,\ldots,z_r)\begin{pmatrix}m_{1k}\\ \vdots\\ m_{rk}\end{pmatrix} + (0,\ldots,0,1,0,\ldots,0)\begin{pmatrix}\sum_{i=1}^r m_{1i}z_i\\ \vdots\\ \sum_{i=1}^r m_{1i}z_i\end{pmatrix}\\ &= \sum_{i=1}^r m_{ik}z_i + \sum_{i=1}^r m_{ki}z_i,\end{aligned}$$

as required. Here $(0,\ldots,0,1,0,\ldots,0)$ denotes a row-vector of zeros with the k-th element replaced by 1.

4.3 Least squares estimation

Claim 4.1:

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

$$\hat{\boldsymbol{\beta}} = (X^T X)^{-1} X^T \boldsymbol{y}. \tag{4.2}$$

Proof. First, note that

$$SS_E = \sum_{i=1}^{n} e_i e_i = \boldsymbol{\varepsilon}^T \boldsymbol{\varepsilon}.$$

Hence in vector notation, to minimise SS_E we must solve the equation

$$\begin{pmatrix} \partial_{\beta_0} \boldsymbol{e}^T \boldsymbol{e} \\ \partial_{\beta_1} \boldsymbol{e}^T \boldsymbol{e} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

In other words, we must solve $\dfrac{\partial \boldsymbol{\varepsilon}^T \boldsymbol{\varepsilon}}{\partial \boldsymbol{\beta}} = \mathbf{0}$:

$$\begin{aligned} \frac{\partial \boldsymbol{e}^T \boldsymbol{e}}{\partial \boldsymbol{\beta}} &= \frac{\partial}{\partial \boldsymbol{\beta}} (\boldsymbol{y} - X\boldsymbol{\beta})^T (\boldsymbol{y} - X\boldsymbol{\beta}) \\ &= \frac{\partial}{\partial \boldsymbol{\beta}} (\boldsymbol{y}^T \boldsymbol{y} - \boldsymbol{\beta}^T X^T \boldsymbol{y} - \boldsymbol{y}^T X \boldsymbol{\beta} + \boldsymbol{\beta}^T (X^T X) \boldsymbol{\beta}) \\ &= -X^T \boldsymbol{y} - (\boldsymbol{y}^T X)^T + \{(X^T X)^T + (X^T X)\} \boldsymbol{\beta} \\ &= -2X^T \boldsymbol{y} + 2(X^T X) \boldsymbol{\beta}. \end{aligned}$$

When $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}$, the least squares estimator, we require the derivative above to be be zero,

$$\left. \frac{\partial \boldsymbol{e}^T \boldsymbol{e}}{\partial \boldsymbol{\beta}} \right|_{\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}} = \mathbf{0} \quad \Longrightarrow \quad -2X^T \boldsymbol{y} + 2(X^T X) \hat{\boldsymbol{\beta}} = \mathbf{0}.$$

The last equation is a matrix form of the **normal equations**,

$$(X^T X) \hat{\boldsymbol{\beta}} = X^T \boldsymbol{y}. \tag{4.3}$$

It remains to multiply both sides with the inverse $(X^T X)^{-1}$ giving the required result. □

Why bother with matrices given that we already have the least squares estimates of β_0 and β_1? The crucial feature of the result in (4.2) is that it applies to **any** linear model. In particular, the matrix form of normal equations (4.3) is valid for any linear model, and thus the result (4.2) will allow us to estimate parameters of any linear model that we will encounter in this course.

Let us verify that the matrix form gives us the results we have already obtained. Recall that the design matrix for the simple linear regression model is

$$X = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \\ 1 & x_n \end{bmatrix}.$$

Then

$$X^T\boldsymbol{y} = \begin{bmatrix} 1 & 1 & \dots & 1 \\ x_1 & x_2 & \dots & x_n \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n x_i y_i \end{bmatrix}$$

and

$$X^T X = \begin{bmatrix} 1 & 1 & \dots & 1 \\ x_1 & x_2 & \dots & x_n \end{bmatrix} \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} = \begin{bmatrix} n & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \end{bmatrix}.$$

The determinant of $X^T X$ is given by

$$|X^T X| = n\sum_{i=1}^n x_i^2 - \Big(\sum_{i=1}^n x_i\Big)^2 = n s_{xx}.$$

Thus the inverse of $X^T X$ is

$$(X^T X)^{-1} = \frac{1}{n s_{xx}} \begin{bmatrix} \sum_{i=1}^n x_i^2 & -\sum_{i=1}^n x_i \\ -\sum_{i=1}^n x_i & n \end{bmatrix} = \frac{1}{s_{xx}} \begin{bmatrix} \frac{1}{n}\sum_{i=1}^n x_i^2 & -\bar{x} \\ -\bar{x} & 1 \end{bmatrix}.$$

So the solution to the normal equations is given by

$$\begin{aligned} \hat{\boldsymbol{\beta}} = (X^T X)^{-1} X^T \boldsymbol{y} &= \frac{1}{s_{xx}} \begin{bmatrix} \frac{1}{n}\sum_{i=1}^n x_i^2 & -\bar{x} \\ -\bar{x} & 1 \end{bmatrix} \begin{bmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n x_i y_i \end{bmatrix} \\ &= \frac{1}{s_{xx}} \begin{bmatrix} \frac{1}{n}\sum_{i=1}^n x_i^2 \sum_{i=1}^n y_i - \bar{x}\sum_{i=1}^n x_i y_i \\ \sum_{i=1}^n x_i y_i - \bar{x}\sum_{i=1}^n y_i \end{bmatrix} \\ &= \frac{1}{s_{xx}} \begin{bmatrix} \frac{1}{n}\sum_{i=1}^n x_i^2 \sum_{i=1}^n y_i - \bar{x}^2\sum_{i=1}^n y_i - \bar{x}(\sum_{i=1}^n x_i y_i - \bar{x}\sum_{i=1}^n y_i) \\ s_{xy} \end{bmatrix} \\ &= \frac{1}{s_{xx}} \begin{bmatrix} s_{xx}\bar{y} - \bar{x}s_{xy} \\ s_{xy} \end{bmatrix} \\ &= \begin{bmatrix} \bar{y} - \hat{\beta}_1\bar{x} \\ \hat{\beta}_1 \end{bmatrix}, \end{aligned}$$

which is the same result we found earlier, that is $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$ and $\hat{\beta}_1 = s_{xy}/s_{xx}$.

These notes are a compilation on the following textbooks:

- D. C. Montgomery, E. A. Peck and G. G. Vining, *Introduction to Linear Regression Analysis*, 5th Edition, Wiley (2012). [Chapters 2 and 4–6]
- S. J. Sheather, *A Modern Approach to Regression with R*, Springer (2009). [Sections 2, 3]