

Module 2: Simple linear regression

2.1	oIntroduction	1
2.2	Simple linear regression models	1
2.3	Fitting the model	3
2.3.1	The principle of least squares	4
2.3.2	Coefficient of determination	9
2.3.3	Estimating the variance	10
2.4	Inference in simple linear regression	11
2.4.1	Inference on the regression parameters	12
2.5	Summary	15

2.1 oIntroduction

This module reviews simple linear regression models. That is, regression models with just one explanatory variable, and where the relationship between the response variable and the explanatory variable is a straight line. Although these models are of a simple nature, they are important for various reasons. Firstly, they are very common (you have already met several examples in Module 1). This is partly due to the fact that non-linear relationships often can be approximated by straight lines, over limited ranges. Secondly, in cases where a scatterplot of the data displays a non-linear relationship between the response variable and the explanatory variable, it is sometimes possible to transform the data into a new pair of variables with a straight-line relationship. That is, we can transform a simple *non-linear* regression model into a simple *linear* regression model, and analyse the data using methodology from linear models. Lastly, the simplicity of these models make them useful in providing an overview of the general methodology. Later in the course, we shall extend the results for simple linear regression models to more complex settings.

A formal definition of the simple linear regression model is given in Section 2.2. In Section 2.3, we discuss how to fit the model, and how to estimate the variation away from the line. Section 2.4 concerns inference on simple linear regression models.

2.2 Simple linear regression models

In most of the examples and exercises in Module 1, there was only one explanatory variable, and the relationship between this variable and the response variable was a straight-line with

some random fluctuation around the line.

Example 2.1 *Mobility of elderly people*

These data concern the relationship between two methods for measuring the mobility of elderly people: the TUG score (x) and the Berg score (Y). A scatterplot of the data is shown in Figure 2.1.

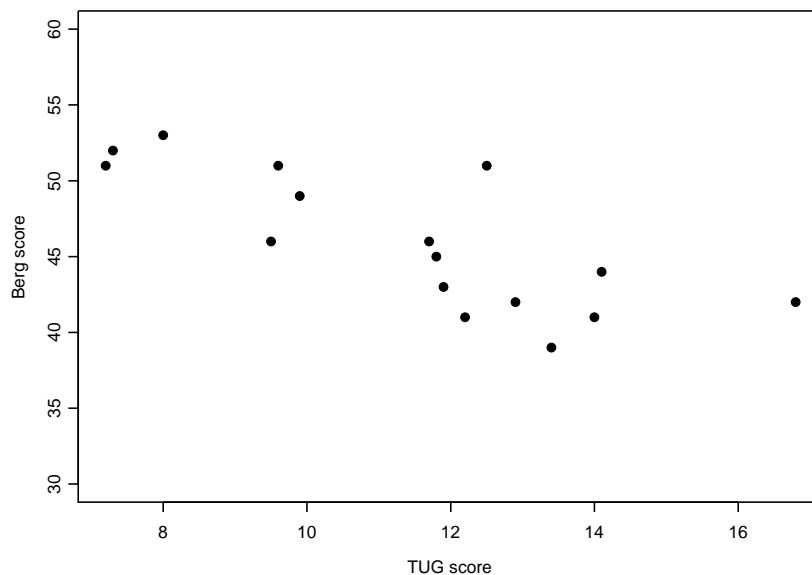


Figure 2.1: Berg score against TUG score

The relationship between the variables could be described as a straight line, and some random fluctuations. Thus, we can use, as a model for the data, the model

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, \dots, 16.$$

This is an example of a simple linear regression model.

Further details on this dataset can be found here.

◇

Suppose that we have a response variable Y and an explanatory variable x , then the **simple linear regression model** for Y on x is given by

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, \dots, n, \quad (2.1)$$

where β_0 and β_1 are unknown parameters, and the ε_i s are independent random variables with zero mean and constant variance for all $i = 1, \dots, n$.

The parameters β_0 and β_1 are called **regression parameters** (or **regression coefficients**), and the line $h(x) = \beta_0 + \beta_1 x$ is called the **regression line** or the **linear predictor**. (Recall that a general $h(\cdot)$ is called a regression curve.) The regression parameters β_0 and β_1 are unknown, non-random parameters. They are the intercept and the slope, respectively, of the straight line relating Y to x .

The name simple *linear* regression model refers to the fact that the mean value of the response:

$$\mathbb{E}[Y_i] = \beta_0 + \beta_1 x_i$$

is a linear function of the regression parameters β_0 and β_1 . (Note that $\mathbb{E}[Y_i]$ is an *affine* function of the explanatory variable x_i .)

The terms ε_i in (2.1) are called **random errors** or **random terms**. The random error ε_i is the term which accounts for the variation of the i th response variable Y_i away from the linear predictor $\beta_0 + \beta_1 x_i$ at the point x_i . That is,

$$\varepsilon_i = Y_i - \beta_0 - \beta_1 x_i, \quad i = 1, \dots, n. \quad (2.2)$$

The ε_i s are independent random variables with the same variance and zero mean. Hence, the response variables Y_i are independent with means $\beta_0 + \beta_1 x_i$, and constant variance equal to the variance of ε_i .

Example 2.1 (continued) *Mobility of elderly people*

An interpretation of the regression parameters β_0 and β_1 is as follows:

β_0 : The expected Berg score for a hypothetical patient with TUG score zero.

β_1 : The expected change in the Berg score, when the TUG score is increased by one minute.

Observe that the slope of the line is negative, implying that the Berg score decreases with increasing TUG score.

◇

2.3 Fitting the model

Having decided that a straight line might describe the relationship in the data well, the obvious question is now: which line fits the data best?

In Figure 2.2 four different lines are added to a scatterplot for the data on mobility of elderly people. One or two of the lines may look a little better than others, but it is difficult to decide which line is the best.

The most common criterion for estimating the best fitting line to data is the *principle of least squares*. This criterion is described in Subsection 2.3.1. Subsection 2.3.2 concerns a

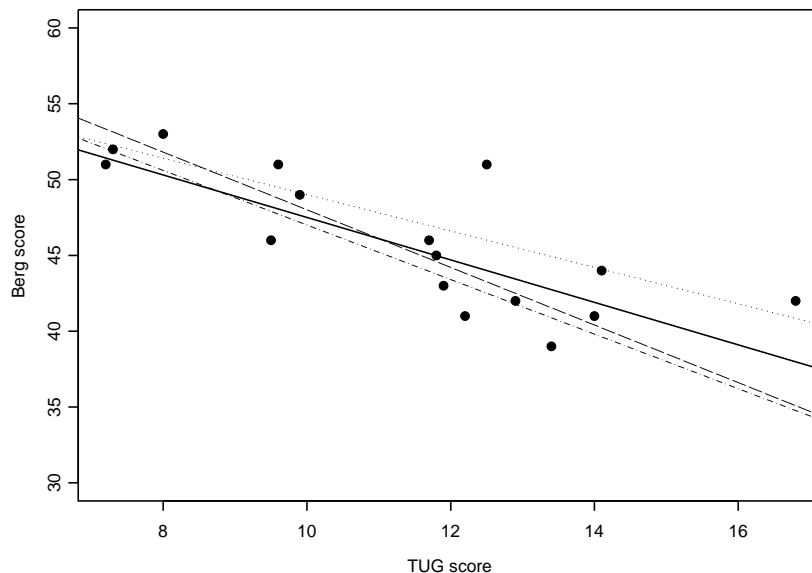


Figure 2.2: Mobility data: Four different regression lines

measure of the strength of the straight-line relationship. When we estimate the regression line, we effectively estimate the two regression parameters β_0 and β_1 . That leaves one remaining parameter in the model: the common variance σ^2 of the response variables. We discuss how to estimate σ^2 in Subsection 2.3.3.

2.3.1 The principle of least squares

The principle of least squares is based on the *residuals*. For any line, the **residuals** are the deviations of the response variables Y_i away from the line. (Note that residuals always refer to a given line or curve.) The residuals are usually denoted by ε_i like the random errors in (2.2). The reason for this notation is that, if the line is the true regression line of the model, then the residuals are exactly the random errors ε_i in (2.2). For a given line $\tilde{h}(x) = \tilde{\beta}_0 + \tilde{\beta}_1 x$, the observed value of ε_i is the difference between the i th observation y_i and the linear predictor $\tilde{\beta}_0 + \tilde{\beta}_1 x_i$ at the point x_i . That is,

$$\varepsilon_i = y_i - \tilde{\beta}_0 - \tilde{\beta}_1 x_i, \quad i = 1, \dots, n. \quad (2.3)$$

The observed values of ε_i are called **observed residuals** (or just **residuals**). In figure 2.3, a possible regression line has been drawn in a scatterplot of the data on mobility of elderly people. The residuals are indicated as vertical lines in the plot.

Note that, the better the line fits the data, the smaller the residuals will be. Thus, we can use the ‘sizes’ of the residuals as a measure of how well a proposed line fits the data. If we

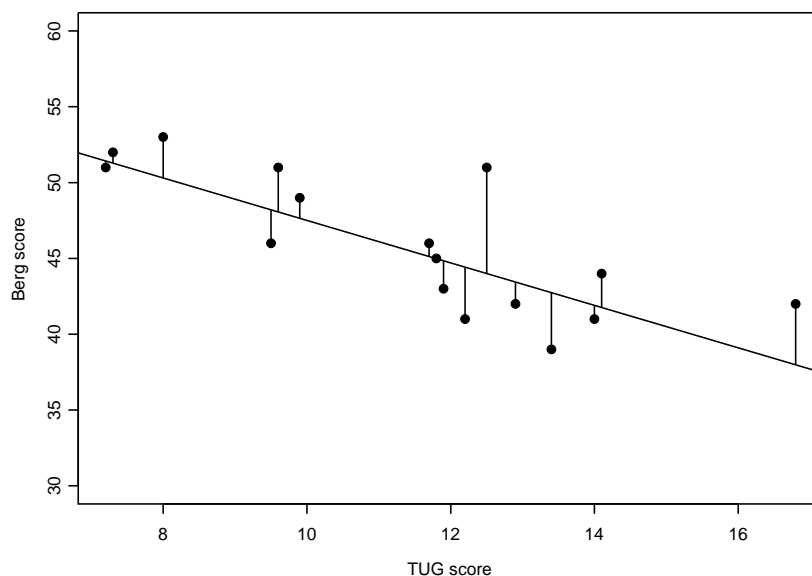


Figure 2.3: Mobility data: the observed residuals

simply used the sum of the residuals, we would get a problem with large positive and large negative values cancelling out; this problem can be avoided by using the sum of the *squared* residuals instead. If this measure—the sum of squared residuals—is small, the line explains the variation in the data well; if it is large, the line explains the variation in the data poorly. The **principle of least squares** is to estimate the regression line by the line which *minimises the sum of squared residuals*. Or, equivalently: estimate the regression parameters β_0 and β_1 by the values which minimise the sum of squared residuals.

The **sum of squared residuals**, or, as it is usually called, the **residual sum of squares**, is denoted by RSS (or $RSS(\beta_0, \beta_1)$ to emphasise that it is a function of β_0 and β_1), and is given by

$$RSS = RSS(\beta_0, \beta_1) = \sum_{i=1}^n \varepsilon_i^2 = \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2. \quad (2.4)$$

(For simplicity, we omit the limits $i = 1$ and $i = n$ on the summation symbols in the following.)

In order to minimise RSS with respect to β_0 and β_1 , we differentiate (2.4), and get

$$\begin{aligned} \frac{\partial RSS}{\partial \beta_0}(\beta_0, \beta_1) &= -2 \sum (y_i - \beta_0 - \beta_1 x_i), \\ \frac{\partial RSS}{\partial \beta_1}(\beta_0, \beta_1) &= -2 \sum x_i (y_i - \beta_0 - \beta_1 x_i). \end{aligned}$$

Putting the derivatives equal to zero and re-arranging the terms, yields the following equations

$$\begin{aligned}\sum y_i &= \beta_0 n + \beta_1 \sum x_i, \\ \sum x_i y_i &= \beta_0 \sum x_i + \beta_1 \sum x_i^2.\end{aligned}$$

Solving the equations for β_0 and β_1 provides the **least squares estimates** $\hat{\beta}_0$ (reads beta-naught-hat) and $\hat{\beta}_1$ (beta-one-hat) of β_0 and β_1 , respectively. They are given by

$$\begin{aligned}\hat{\beta}_0 &= \bar{y} - \beta_1 \bar{x}, \\ \hat{\beta}_1 &= \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2},\end{aligned}$$

where $\bar{y} = \sum y_i/n$ and $\bar{x} = \sum x_i/n$ denote the sample means of the response and explanatory variable, respectively.

The estimated regression line is called the **least squares line** or the **fitted regression line** and is given by

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x. \quad (2.5)$$

The values $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ are called the **fitted values** or the **predicted values**. The fitted value \hat{y}_i is an estimate of the expected response for a given value x_i of the explanatory variable. The residuals corresponding to the fitted regression line, are called the **fitted residuals**, or simply the **residuals**. They are given by

$$\begin{aligned}\hat{\varepsilon}_i &= y_i - \hat{y}_i \\ &= y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \quad i = 1, \dots, n.\end{aligned} \quad (2.6)$$

The fitted residuals can be thought of as observations of the random errors ε_i in the simple linear regression model (2.1).

It is convenient to use the following shorthand notation for the sums involved in the expressions for the parameter estimates (all summations are for $i = 1, \dots, n$):

$$\begin{aligned}s_{xx} &= \sum (x_i - \bar{x})^2 = \sum x_i^2 - \frac{(\sum x_i)^2}{n}, \\ s_{yy} &= \sum (y_i - \bar{y})^2 = \sum y_i^2 - \frac{(\sum y_i)^2}{n}, \\ s_{xy} = s_{yx} &= \sum (x_i - \bar{x})(y_i - \bar{y}) = \sum x_i y_i - \frac{\sum y_i \sum x_i}{n}.\end{aligned}$$

The sums s_{xx} and s_{yy} are called **corrected sums of squares**, and the sums s_{xy} and s_{yx} are called **corrected sums of cross products**. (The corresponding sums involving the random variables Y_i rather than the observations y_i are denoted by upper-case letters: S_{yy} , S_{xy} and S_{yx} .) In this notation, the least squares estimates of the regression parameters β_1 and β_0 of the slope and intercept of the regression line are given by

$$\hat{\beta}_1 = \frac{s_{xy}}{s_{xx}}, \quad (2.7)$$

and

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}, \quad (2.8)$$

respectively.

Note that the estimate of $\hat{\beta}_1$ is undefined if $s_{xx} = 0$ (division by zero). But this is not a problem in practice: if $s_{xx} = 0$ the explanatory variable only takes one value, and there can be no best line. Note also that the least squares line passes through the **centroid** (the point (\bar{x}, \bar{y})) of the data.

Example 2.1 (continued) *Mobility of elderly people*

For the data on mobility of elderly people, the least squares estimates of the regression parameters are given by

$$\begin{aligned} \hat{\beta}_1 &= -1.340 \\ \hat{\beta}_0 &= 61.314. \end{aligned}$$

So, the fitted least squares line has equation

$$\hat{y} = 61.314 - 1.340 x.$$

The least squares line is shown in Figure 2.4. The line appears to fit the data reasonably well.

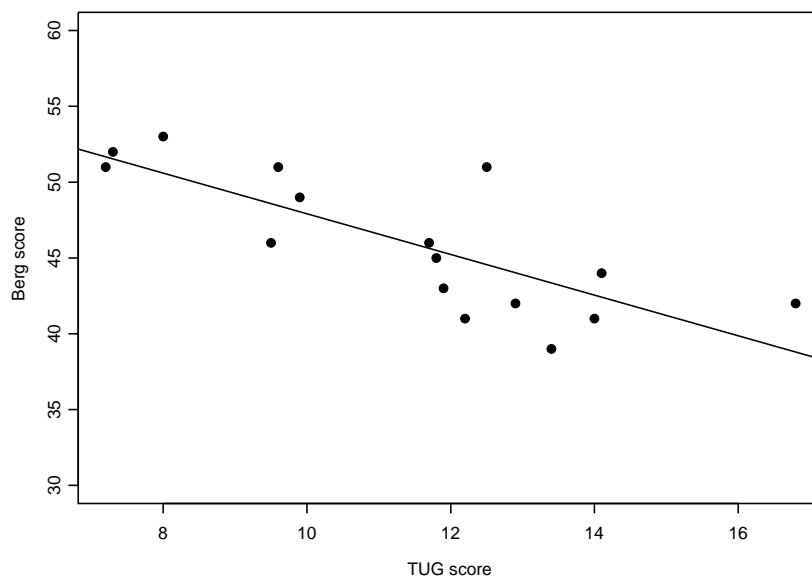


Figure 2.4: Mobility data; the least squares line

**Example 2.2** *Age and height of children*

In the example from Module 1 on age and height of children from an Egyptian village, the interest was in the overall growth pattern of the children. The least squares line relating average height to age has the equation

$$\hat{y} = 64.927 + 0.635x.$$

That is,

$$\text{Height} = 64.927 + 0.635 \times \text{Age},$$

where height is measured in cm, and age in months. Figure 2.5 shows the least squares line in a scatterplot of the data. You can see that the line fits the data very well.

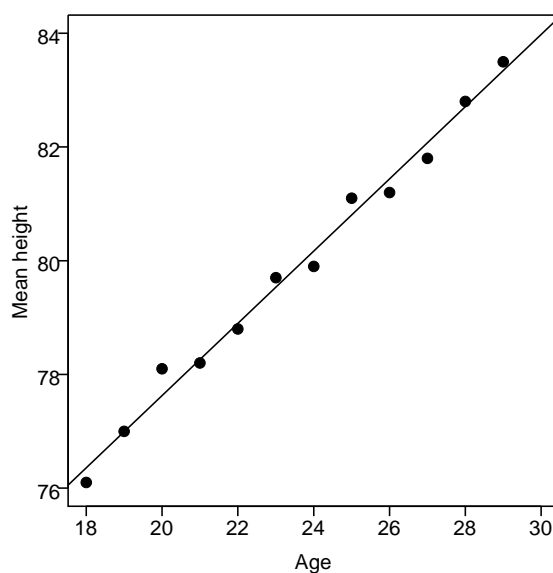


Figure 2.5: Age and height data; the least squares line

Further details on this dataset can be found here.



The least squares principle is the traditional and most common method for estimating the regression parameters. But there exists other estimating criteria: *e.g.* estimating the parameters by the values that minimise the sum of absolute values of the residuals, or by

the values that minimise the sum of orthogonal distances between the observed values and the fitted line. The principle of least squares has various advantages to the other methods. For example, it can be shown that, if the response variables are normally distributed (which is often the case), the least squares estimates of the regression parameters are exactly the maximum likelihood estimates of the parameters.

2.3.2 Coefficient of determination

In the previous subsection we used the principle of least squares to fit the ‘best’ straight line to data. But how well does the least squares line explain the variation in the data? In this subsection we describe a measure for roughly assessing how well a fitted line describes the variation in data: the *coefficient of determination*.

The coefficient of determination compares the amount of variation in the data away from the fitted line with the total amount of variation in the data. The argument is as follows: if we did not have the linear model we would have to use the ‘naïve’ model $\hat{y} = \bar{y}$ instead. The variation away from the naïve model is $S_{yy} = \sum_{i=1}^n (Y_i - \bar{y})^2$: the total amount of variation in the data. However, if we use the least squares line (2.5) as model, the variation away from model is only

$$RSS(\hat{\beta}_0, \hat{\beta}_1) = \sum_{i=1}^n (Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 = S_{yy} - \frac{S_{xy}^2}{s_{xx}}.$$

A measure of the strength of the linear relationship between Y and x is the **coefficient of determination** R^2 : it is the *proportional reduction in variation* obtained by using the least squares line instead of the naïve model. That is, the reduction in variation away from the model ($S_{yy} - RSS$) as a proportion of the total variation (S_{yy}):

$$R^2 = \frac{S_{yy} - RSS}{S_{yy}} = \frac{S_{yy} - S_{yy} + S_{xy}^2/s_{xx}}{S_{yy}} = \frac{S_{xy}^2}{s_{xx}S_{yy}}.$$

The larger the value of R^2 , the greater the reduction from S_{yy} to RSS relative to S_{yy} , and the stronger the relationship between Y and x . An estimate of R^2 is found by substituting S_{yy} and S_{xy} by the observed sums s_{yy} and s_{xy} , that is

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

Note that the square root of r^2 is exactly the estimate from Module 1 of the Pearson correlation coefficient, ρ , between x and Y when x is regarded as a random variable:

$$r = \frac{s_{xy}/(n-1)}{s_{(x)}s_{(y)}},$$

where $s_{(x)} = \sqrt{s_{xx}/(n-1)}$ and $s_{(y)} = \sqrt{s_{yy}/(n-1)}$ are the standard deviations for x and y , respectively.

The value of R^2 will always lie between 0 and 1 (or, in percentage, between 0% and 100%). It is equal to 1 if $\hat{\beta}_1 \neq 0$ and $RSS = 0$, that is, if all the data points lie precisely on the fitted straight line (*i.e.* when there is a ‘perfect’ relationship between Y and x). If the coefficient of determination is close to 1, it is an indication that the data points lie close to the least squares line. The value of R^2 is zero if $RSS = S_{yy}$, that is, the fitted straight-line model offers no more information about the value of Y than the naïve model does.

It is tempting to use R^2 as a measure of whether a model is good or not. This is *not* appropriate. Try and think of why for a moment before reading on.

The coefficient of determination is only a measure of how well a straight-line model describes the variation in the data compared to the *naïve* model—not to other models in general. Even though R^2 is close to 1 (*i.e.* a straight-line explains a large proportion of the variation), it could easily be that a non-linear model explains the data-variation much better than the linear. Methods for assessing the appropriateness of the assumption of a straight-line relationship between Y and x will be discussed in Module 4.

Example 2.2 (continued) *Age and height of children*

The relevant summary statistics for these data are

$$s_{xx} = 143, \quad s_{yy} = 58.31, \quad s_{xy} = 90.8.$$

The coefficient of determination is given by

$$r^2 = \frac{s_{xy}^2}{s_{xx}s_{yy}} = \frac{90.8^2}{143 \times 58.31} = 0.989 = 98.9\%.$$

Since the coefficient of determination is very high, the model seems to describe the variation in the data very well. ◇

2.3.3 Estimating the variance

In Subsection 2.3.1, we found that the principle of least squares can provide estimates of the regression parameters in a simple linear regression model. But, in order to fit the model we also need an estimate for the common variance σ^2 . Such an estimate is required for making statistical inferences about the true straight-line relationship between x and Y . Since σ^2 is the common variance of the residuals ε_i , $i = 1, \dots, n$, it would be natural to estimate it by the sample variance of the fitted residuals (2.6). That is, an estimate would be

$$\sum (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 / (n - 1) = RSS / (n - 1),$$

where $RSS = RSS(\hat{\beta}_0, \hat{\beta}_1)$. However, it can be shown that this is a *biased* estimate of σ^2 , that is, the corresponding estimator does not have the ‘correct’ mean value: $\mathbb{E}[RSS / (n - 1)] \neq \sigma^2$. An **unbiased estimate of the common variance**, σ^2 , is given by

$$s^2 = \frac{RSS(\hat{\beta}_0, \hat{\beta}_1)}{n - 2} = \left(s_{yy} - \frac{s_{xy}^2}{s_{xx}} \right) / (n - 2), \quad (2.9)$$

The denominator in (2.9) is the residual **degrees of freedom** (*d.f.*), that is

$$d.f. = \text{number of observations} - \text{number of estimated parameters.}$$

In particular, for simple linear regression models, we have n observations and we have estimated the two regression parameters β_0 and β_1 , so the residual *d.f.* is $n - 2$.

Example 2.2 (continued) *Age and height of children*

The relevant summary statistics for these data are

$$\begin{aligned} s_{xx} &= 143, & s_{yy} &= 58.31, \\ n &= 12, & s_{xy} &= 90.8. \end{aligned}$$

An unbiased estimate of the common variance σ^2 is given by

$$s^2 = \left(s_{yy} - \frac{s_{xy}^2}{s_{xx}} \right) / (n - 2) = \frac{0.6552}{10} = 0.0655.$$

◇

2.4 Inference in simple linear regression

In Section 2.3 we produced an estimate of the straight line that describes the data-variation best. However, since the estimated line is based on the particular sample of data, x_i and y_i , $i = 1, \dots, n$, we have observed, we would almost certainly get a different line if we took a new sample of data and estimated the line on the basis of the new sample. For example, if we measured the heights and ages of children in the village neighbouring the one in Example 2.2, we would invariably get different measurements, and therefore a different least squares line. In other words: the least squares line is an observation of a *random line* which varies from one experiment to the next. Likewise, the least squares estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ of the intercept and slope, respectively, of the least squares line, are both observations of random variables. These random variables are called the **least squares estimators**. (An estimate is non-random and is an observation of an estimator, which is a random variable.) The least squares estimators are given by

$$\hat{\beta}_1 = \frac{S_{xy}}{s_{xx}} = \frac{\sum (x_i - \bar{x})(Y_i - \bar{Y})}{\sum (x_i - \bar{x})^2}, \quad (2.10)$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x}, \quad (2.11)$$

where $\bar{Y} = \sum Y_i / n$, and with all summations from $i = 1$ to n . By a similar argument we find that an unbiased estimator for the common variance σ^2 is given by

$$\begin{aligned} S^2 &= \left(S_{yy} - \frac{S_{xy}^2}{s_{xx}} \right) / (n - 2) \\ &= \sum (Y_i - \hat{Y}_i)^2 / (n - 2), \end{aligned} \quad (2.12)$$

where $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$, with $\hat{\beta}_0$ and $\hat{\beta}_1$ being the least squares estimators. Note that the randomness in the estimators is due to the response variables only, since the explanatory variables are non-random. In particular, it can be seen from (2.10) and (2.11) that $\hat{\beta}_0$ and $\hat{\beta}_1$ are linear combinations of the response variables.

It can be shown that the least squares estimators are *unbiased*, that is, that they have the ‘correct’ mean values:

$$\mathbb{E}[\hat{\beta}_0] = \beta_0 \quad \text{and} \quad \mathbb{E}[\hat{\beta}_1] = \beta_1. \quad (2.13)$$

Also, the estimator S^2 is an unbiased estimator of the common variance σ^2 , that is

$$\mathbb{E}[S^2] = \sigma^2. \quad (2.14)$$

The variances of the estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ can be found from standard results on variances (we shall not do it here). The variances are given by

$$\text{var}[\hat{\beta}_0] = \frac{\sigma^2}{n} + \frac{\bar{x}^2}{s_{xx}}\sigma^2 \quad (2.15)$$

$$\text{var}[\hat{\beta}_1] = \frac{\sigma^2}{s_{xx}}. \quad (2.16)$$

Note that both variances decrease when the sample size n increases. Also, the variances decrease if $s_{xx} = \sum(x_i - \bar{x})^2$ is increased. (That is, if the x -values are widely dispersed.) In some studies, it is possible to design the experiment such that the value of s_{xx} is high, and hence the variances of the estimators are small. It is desirable to have small variances, as it improves the precision of results drawn from the analysis.

In order to make inferences about the model, such as testing hypotheses and producing confidence intervals for the regression parameters, we need to make some assumption on the distribution of the random variables Y_i . The most common assumption—and the one we shall make here—is that the response variables Y_i are normally distributed.

Module 4 concerns various methods for checking the assumptions of regression models. In this section, we shall simply assume the following about the response variables: the Y_i s are independent normally distributed random variables with equal variances and mean values depending linearly on x_i .

2.4.1 Inference on the regression parameters

To test hypotheses and construct confidence intervals for the regression parameters β_0 and β_1 , we need the distributions of the parameter estimators $\hat{\beta}_0$ and $\hat{\beta}_1$. Recall from (2.10) and (2.11) that the least squares estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ are linear combinations of the response variables Y_i . Standard theory on the normal distribution says that a linear combination of independent, normal random variables is normally distributed. Thus, since the Y_i s are independent, normal random variables, the estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ are both normally distributed. In (2.13)–(2.16),

we found the mean values and variances of the estimators. Putting everything together, we get that

$$\begin{aligned}\hat{\beta}_0 &\sim N\left(\beta_0, \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{s_{xx}}\right)\right) \\ \hat{\beta}_1 &\sim N\left(\beta_1, \frac{\sigma^2}{s_{xx}}\right).\end{aligned}$$

It can be shown that the distribution of the estimator S^2 of the common variance σ^2 is given by

$$S^2 \sim \frac{\sigma^2 \chi_{n-2}^2}{n-2},$$

where χ_{n-2}^2 denotes a chi-square distribution with $n-2$ degrees of freedom. Moreover, it can be shown that the estimator S^2 is independent of the estimators $\hat{\beta}_0$ and $\hat{\beta}_1$. (But the estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ are not mutually independent.)

We can use these distributional results to test hypotheses on the regression parameters. Since both $\hat{\beta}_0$ and $\hat{\beta}_1$ have normal distributions with variances depending on the unknown quantity σ^2 , we can apply standard results for normal random variables with unknown variances. Thus, in order to test β_i equal to some value β_i^* , $i = 0, 1$, that is, to test hypotheses of the form $H_0 : \beta_i = \beta_i^*$, for $i = 0, 1$, we can use the t -test statistic, given by

$$t_{\hat{\beta}_i}(y) = \frac{\hat{\beta}_i - \beta_i^*}{\text{se}[\hat{\beta}_i]}, \quad i = 0, 1, \quad (2.17)$$

where $\text{se}[\hat{\beta}_i]$ denotes the estimated standard error of the estimator $\hat{\beta}_i$. That is

$$\text{se}[\hat{\beta}_0] = \sqrt{\text{var}[\hat{\beta}_0]} = \sqrt{s^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{s_{xx}}\right)}$$

and

$$\text{se}[\hat{\beta}_1] = \sqrt{\text{var}[\hat{\beta}_1]} = \sqrt{s^2/s_{xx}}.$$

It can be shown that both test statistics $t_{\hat{\beta}_0}(y)$ and $t_{\hat{\beta}_1}(y)$ have t -distributions with $n-2$ degrees of freedom.

The test statistics in (2.17) can be used for testing the parameter β_i ($i = 0, 1$) equal to any value β_i^* . However, for the slope parameter β_1 , one value is particularly important: if we can test β_1 equal to zero, the simple linear regression model simplifies to

$$Y_i = \beta_0 + \varepsilon_i, \quad i = 1, \dots, n.$$

That is, the value of Y_i does not depend on the value of x_i . In other words: the response variable and the explanatory variable are unrelated!

It is common—for instance in computer output—to present the estimates and standard errors of the least squares estimators in a table like the following.

Parameter	Estimate	Standard error	t -statistic	p -value
β_0	$\hat{\beta}_0$	$\text{se}[\hat{\beta}_0]$		
β_1	$\hat{\beta}_1$	$\text{se}[\hat{\beta}_1]$		

The column ‘ t -statistic’ contains the t -test statistic (2.17) for testing the hypotheses $H_0 : \beta_0 = 0$ and $H_0 : \beta_1 = 0$, respectively. (Should you wish to test a parameter equal to a different value, it is easy to produce the appropriate test statistic (2.17) from the table.) The column ‘ p -value’ contains the p -values corresponding to the t -test statistic in the same row.

Example 2.2 (continued) *Age and height of children*

For the data on age and height of Egyptian children, the table is given by

Parameter	Estimate	Standard error	t -statistic	p -value
β_0	64.9283	0.5084	127.7085	0.0000
β_1	0.6350	0.0214	29.6647	0.0000

Not surprisingly, neither parameter can be tested equal to zero. If, for some reason, we wished to test whether the slope parameter was equal to 0.58, say, the test statistic would be

$$t_{\hat{\beta}_1}(y) = \frac{\hat{\beta}_1 - 0.58}{\text{se}[\hat{\beta}_1]} = \frac{0.635 - 0.58}{0.0214} = 2.570.$$

Since $n = 12$ in this example, the test statistic has a $t(10)$ -distribution. The p -value for this test is 0.0279, thus, on the basis of these data we reject the hypothesis that the slope parameter is 0.58, at the 5% significance level. \diamond

A second practical use of the table is to provide confidence intervals for the regression parameters. The $1 - \alpha$ confidence interval for β_0 and β_1 are given by, respectively,

$$\hat{\beta}_0 \pm t_{1-\alpha/2}(n-2)\text{se}[\hat{\beta}_0],$$

and

$$\hat{\beta}_1 \pm t_{1-\alpha/2}(n-2)\text{se}[\hat{\beta}_1].$$

In order to construct the confidence intervals, all that is needed is the table and $t_{1-\alpha/2}(n-2)$: the $(1 - \alpha/2)$ -quantile of a $t(n-2)$ -distribution.

Example 2.2 (continued) *Age and height of children*

For the data on age and height of Egyptian children, the 95% confidence intervals for the regression parameters can be obtained from the table for these data and the 0.975-quantile of a $t(10)$ -distribution: $t_{0.975}(10) = 2.2281$. The confidence intervals for β_0 and β_1 are, respectively,

$$\beta_0 : (63.80, 66.06)$$

and

$$\beta_1 : (0.587, 0.683).$$

\diamond

2.5 Summary

In this module, the simple linear regression model has been discussed. We have described a method, based on the principle of least squares, for fitting simple linear regression models to data. The principle of least squares says to estimate the regression line by the line which minimises the sum of the squared deviations of the observed data away from the line. The intercept and slope of the fitted line are estimates of the regression parameters β_0 and β_1 , respectively. Further, an unbiased estimate of the common variance has been given. Under the assumption of normality of the response variables, we have tested hypotheses and constructed confidence intervals for the regression parameters.

Keywords: simple linear regression model, regression parameters, regression line, linear predictor, observed residuals, residuals, principle of least squares, residual sum of squares, least squares estimates, least squares line, fitted regression line, fitted values, predicted values, fitted residuals, R^2 , coefficient of determination, bias corrected estimate of common variance, degrees of freedom, least squares estimators, distributions of least squares estimators, hypotheses on regression parameters, confidence intervals for regression parameters.