

# ***161.320 Fitting Regression Models***

## ***1. The Simple Linear Model***

---

### **Background**

---

The paper *61.320: Fitting Regression Models* builds on material that was covered in one section of *61.220: Data Analysis*, and it requires that paper (or a similar statistical methods paper) as a prerequisite. The Study Guide for this paper however is self-contained and describes all relevant results from *Data Analysis*, though with less detail, motivation and examples than material that is new here. In particular, the first chapter and much of the second chapter should be revision of material with which you should already be familiar.

There are no mathematical prerequisites to this course, and this is reflected in the approach that is taken to the material. Very few results are proved, but some formulae are given, both for completeness and because the formulae sometimes help to explain the underlying concept. However the formulae are not derived and you are rarely expected to use them directly — a computer program can be asked to do most of the calculations required to analyse regression models.

In this course *Minitab* will be the preferred computer package. However most statistical packages will do the calculations necessary for regression and included in this study guide are some examples using SAS. If you wish to use SAS for assignment work then that is acceptable. Output from *Minitab* will be used in exams and in assignment solutions.

---

### **Computer Software, Minitab and SAS**

---

We similarly assume that you have purchased a copy of the statistical program *Minitab*, and are familiar with its basic operation, since it is used extensively in the prerequisite paper *61.220: Data Analysis*. Ideally you will be using Version 14 on a Windows computer. The latest version that has been released for Macintosh computers at the time of writing of this study guide is Version 10.5 Extra and this contains most of the features that we will use in this course.

However the logistic model that will be described in Chapter 7 of the study guide was not implemented by Minitab until Version 11, so Macintosh users will find it a little more difficult (but not impossible) to fit these models. There are also a number of problems in using the older versions of *Minitab* on a newer Macintosh machine.

WE DO NOT SUPPORT *Minitab* ON MACINTOSH COMPUTERS AND WOULD ADVISE AGAINST DOING THIS COURSE IF YOU DO NOT HAVE A WINDOWS COMPATIBLE COMPUTER AVAILABLE.

You are encouraged to use the menus in *Minitab* rather than typing commands in the ‘Session Window’. Although there are a few advanced features of *Minitab* that require typing of commands, none of them are required for this course.

I have included, in the Appendix, details of how to use Minitab to perform most of the analyses described in this course. I have used pictures of menus and dialog boxes to illustrate the procedure to be followed. These have been collected together in an Appendix at the end of the Study Guide. I have tried to reference these carefully so that they can be found when needed. Whilst Minitab is the preferred option for computing for this paper we have included SAS programs and output in some places in the Appendix. I have tried to make the computer output in the body of this text non software specific. Different packages have their own style but although the output may look different it should contain the same information and the numbers should be the same. I think that a competent statistician should be able to work in a number of software packages hence the two illustrated here.

---

## Terminology

---

This course deals with models that are used to explain how one random variable,  $Y$ , is affected by one or more other measurements,  $x_1, x_2, \dots, x_p$ .

$Y$  is called the **response** variable

$x_1, x_2, \dots, x_p$  are called the **explanatory**, or **regressor** variables<sup>1</sup>.

A statistical model specifies how the **distribution** of the response  $Y$  depends on the values of the explanatory variables; it expresses the response distribution in terms of these values and also one or more unknown parameters. Such models are called **regression models**.

In this course, we rely heavily on the normal distribution as the basis for our models (though other distributions **can** also be used — see the 700-level course *Models for Non-Normal Data*). The normal distribution is a symmetric bell-shaped distribution that is specified by two parameters, the mean  $\mu$  and the variance  $\sigma^2$ , and it is denoted by

$$Y \sim \text{normal}(\mu, \sigma^2)$$

The following might therefore be used as regression models ...

$$Y \sim \text{normal}(\beta_0 + \beta_1 x, \sigma^2)$$

$$\log Y \sim \text{normal}(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2, \sigma^2)$$

$$Y \sim \text{normal}(\beta_0 + \beta_1 e^{b_2 x}, \sigma^2)$$

These models may also be written equivalently in the form ...

---

<sup>1</sup> In some textbooks,  $Y$  is called the **dependent** variable and  $x_1, x_2, \dots, x_p$  are called **independent** variables. This terminology should be avoided since the  $\{x\}$  are not independent in the statistical sense — they are often correlated. The term **predictor** variables is also occasionally used for the explanatory variables.

$$Y = \beta_0 + \beta_1 x + \varepsilon$$

$$\log Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \varepsilon$$

$$Y = \beta_0 + \beta_1 e^{b_2 x} + \varepsilon$$

where in each case  $\varepsilon \sim \text{normal}(0, \sigma^2)$ . The random variable  $\varepsilon$  is called the model's **error** term. The error represents the extent to which the explanatory variables (through the function  $\beta_0 + \beta_1 x$ , etc.) fail to predict the response,  $Y$ , and may be due to the response being variable or some other inadequacy in the model.

The most general form for the models can be written as

$$Y \sim \text{normal}(\mu = f(x_1, x_2, \dots, x_p, \beta_0, \beta_1, \dots, \beta_q), \sigma^2) \quad , \text{ or equivalently}$$

$$Y = f(x_1, x_2, \dots, x_p, \beta_0, \beta_1, \dots, \beta_q) + \varepsilon$$

where  $f(\cdot)$  is some function involving explanatory variables and unknown parameters.

For example, we might be interested in describing how the concentration of a pollutant,  $Y$ , is affected by various characteristics of a car engine (e.g. operating temperature, and various adjustable characteristics of the fuel and ignition system).

## Uses of Regression Models

Regression models are used for various purposes.

- **Model specification.** Sometimes we are just interested in describing a physical system mathematically. We want to ensure that the model mathematically represents accurately the mechanism by which the explanatory variables affect the response.
- **Prediction.** The most important use for regression models is to allow us to **predict** the value of  $Y$  that will result from particular values of the explanatory variables. Usually we use the mean of distribution of  $Y$ 's distribution to predict it. For example, if the model is

$$Y \sim \text{normal}(\beta_0 + \beta_1 x, \sigma^2)$$

and we have obtained estimates  $b_0$  and  $b_1$  for  $\beta_0$  and  $\beta_1$ , then we would predict the  $Y$ -value at a value  $x$  to be

$$\hat{y} = b_0 + b_1 x$$

- **Parameter estimation.** The parameters of a regression model are sometimes meaningful quantities in their own right. For example, if  $Y$  = a student's mark in 61.320 and  $x$  = hours of study and we use a model

$$Y \sim \text{normal}(\beta_0 + \beta_1 x, \sigma^2)$$

then  $\beta_1$  is the extra marks per hour's study.

- **Variable screening.** In some situations we are interested in modelling some variable  $Y$ , perhaps in order to predict it, but we do not know which explanatory variables might affect it. When the explanatory variables are costly to measure, it is useful to be able to discard some explanatory variables as having little or no influence on the response.

As we do more exercises we will find that the analysis that we chose to do may depend on its eventual use. In your work you should bear this in mind, refer back to this list regularly and think what the implications may be for your analysis.

---

## Linear Models

---

When the unknown parameters in the model affect the normal distribution's mean linearly, the model is called a **linear model**. Note that it is the parameters that must appear linearly to make the model a linear model, not the variables. In the simplest linear model,  $Y$  is linearly related to  $x$ .

$$Y = \beta_0 + \beta_1 x + \varepsilon$$

However the following model is also a linear model since  $Y$  is linearly related to  $x^2$ .

$$Y = \beta_0 + \beta_1 x^2 + \varepsilon$$

The model

$$\log Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \varepsilon$$

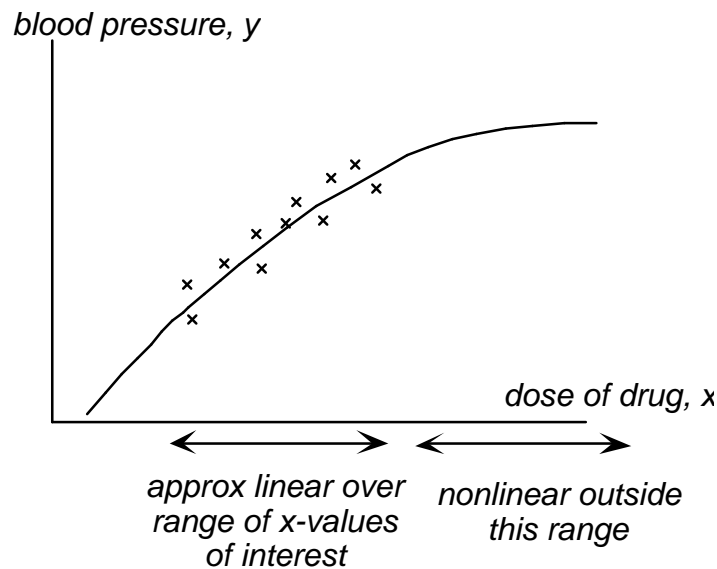
is also a linear model, whereas the model

$$Y = \beta_0 + \beta_1 e^{b_2 x} + \varepsilon$$

is not — the parameters are not involved linearly, so it is called a **nonlinear** model.

Linear models are used almost exclusively in practice. There are several reasons.

- The theory of linear models is much easier than that for nonlinear models.
- Parameter estimates can be computed easily for linear models.
- There is a general theory for inference (tests and confidence intervals) about parameters in linear models.
- Much data that arises in practice is adequately modelled by a linear model.
- Even for data sets where a simple linear model, such as  $Y = \beta_0 + \beta_1 x + \varepsilon$ , is inadequate, a more complex model that is still linear, such as  $\log Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon$ , can often be found to represent the data.
- It is often found that the response variable  $Y$  is approximately linearly related to  $x$  **over the range of  $x$ -values of interest**, even though the relationship may be curved when  $x$  is outside this range. This is illustrated in the diagram below.



A linear model may then be adequate for our needs, provided we do not try to use it outside this range of x-values.

---

## The Principle of Least Squares

---

The first two chapters of this Study Guide deal with the simplest form of linear model which involves only a single explanatory variable,

$$Y = \beta_0 + \beta_1 x + \varepsilon$$

where  $\varepsilon$  is normal  $(0, \sigma^2)$ .

Linear models always involve unknown parameters (such as  $\beta_0$ ,  $\beta_1$  and  $\sigma^2$  above) and these must be **estimated** using data that has been collected. The data are used to estimate the parameters of the model (to **calibrate** the model).<sup>2</sup> For the linear model above, the data that we use to fit the model are the y-values,  $y_1, y_2, \dots, y_n$ , corresponding to  $n$  values of the explanatory variable,  $x_1, x_2, \dots, x_n$ .

The **simple linear model** specifies that  $n$  independent observations of the response  $Y$  are made, with  $Y_i$  recorded at the value  $x_i$  of the explanatory variable.

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

where the  $\varepsilon_i$  are independent normal  $(0, \sigma^2)$ .<sup>3</sup>

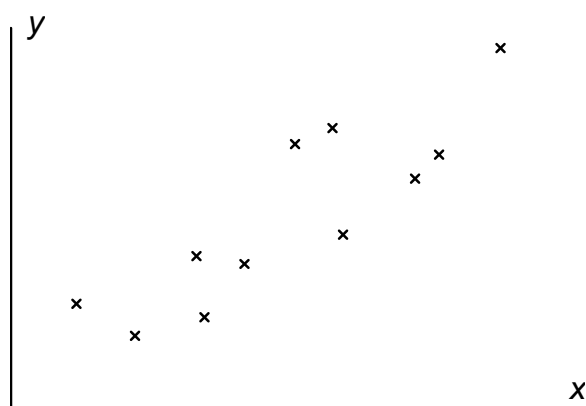
In this section, we will consider how to estimate the parameters of the simple linear model but the results are also applicable to other regression models — both linear and nonlinear.

---

<sup>2</sup>The data is also later used to assess whether the model adequately describes the situation being modelled (to **assess the goodness-of-fit** of the model) — see later in the course.

<sup>3</sup>From now on, we will dispense with the 'correct' notation of capital letters, such as  $Y$ , for random variables and small letters, such as  $y$  for observed values of these random variables. This simplification is intended to make the text more readable.

You should always start analysis of regression data with a scatterplot of the response (on the vertical axis) against the explanatory variable (on the horizontal axis). A simple scatterplot may indicate that the simple linear model relating  $y$  to  $x$  is not appropriate (see Chapter 2).



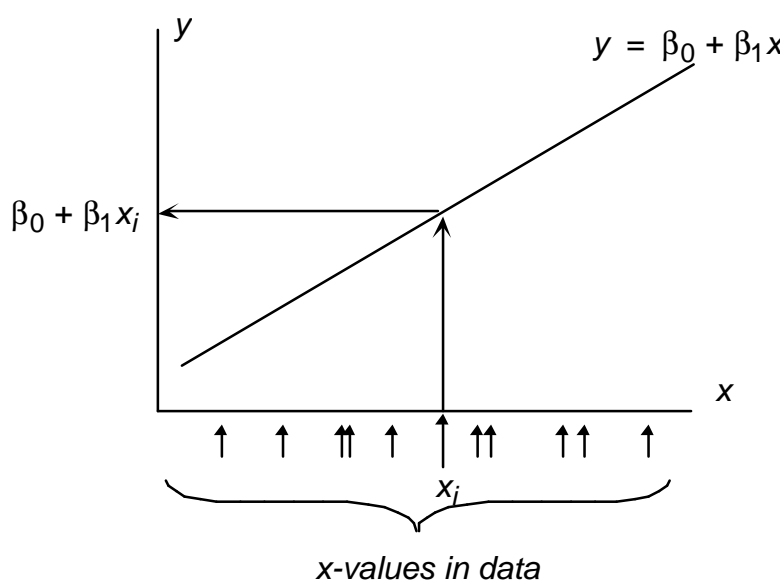
As the expected value of  $y$  is

$$E[y] = \beta_0 + \beta_1 x ,$$

if we knew the value of  $x$  (and the parameters  $\beta_0$  and  $\beta_1$ ), but did not know the  $y$ -values, then the model would predict  $y$  to be the corresponding point on the regression line,

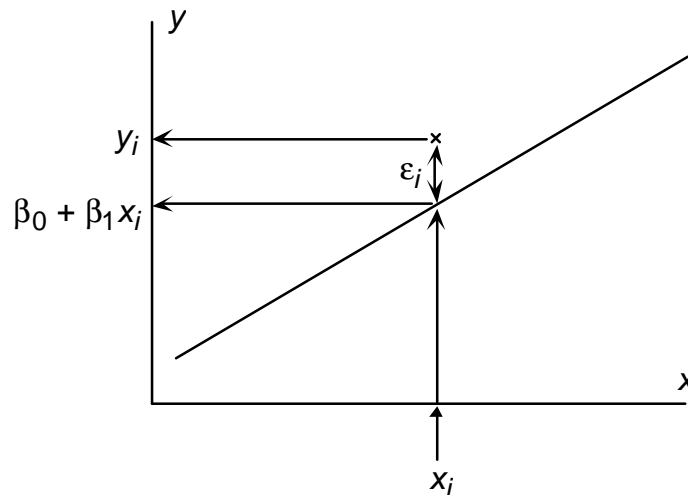
$$y = \beta_0 + \beta_1 x$$

This is shown in the diagram below for a typical observation  $x_i$

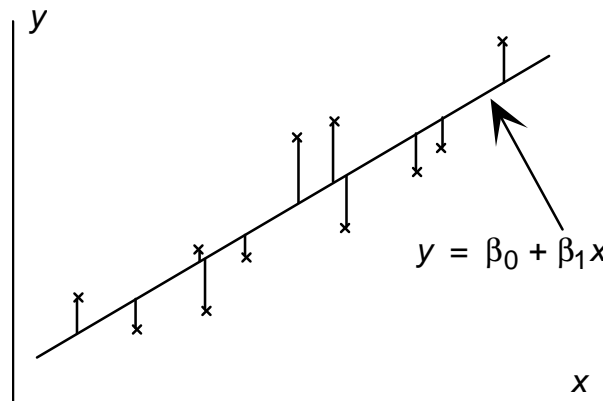


When data are collected, the actual  $y$ -values are not the same as these predictions. In fact the model states that they are normally distributed with standard deviation  $\sigma$  round the predictions,  $y_i \sim \text{normal}(\beta_0 + \beta_1 x_i, \sigma^2)$ . The errors in the predictions are

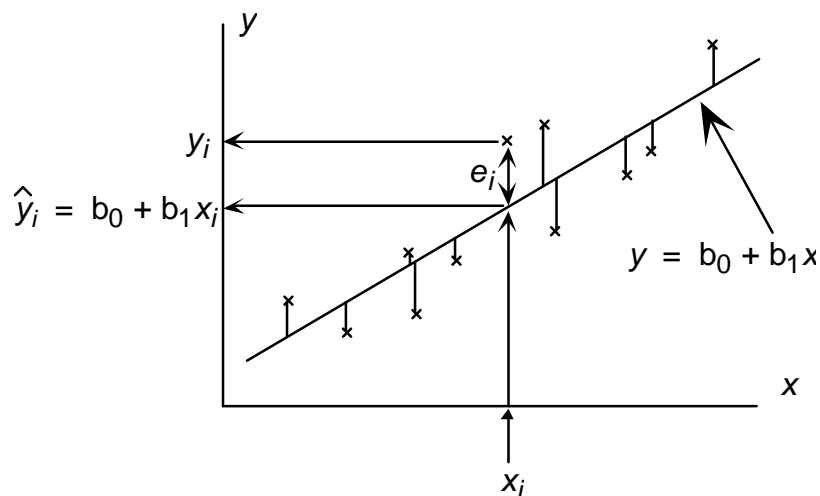
$$\varepsilon_i = y_i - (\beta_0 + \beta_1 x_i)$$



For the complete data set, the errors are shown below



In practice, since we do not know the values of  $\beta_0$  or  $\beta_1$ , the errors are also unknown, though we would expect them to be small. A reasonable way to estimate the unknown parameters  $\beta_0$  and  $\beta_1$  would be with values  $b_0$  and  $b_1$  that make the prediction errors based on these parameter estimates 'as small as possible'.



We therefore estimate  $\beta_0$  and  $\beta_1$  with the values  $b_0$  and  $b_1$  that minimise the residual sum of squares,

$$SS_{\text{Residual}} = \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

where  $\hat{y}_i = b_0 + b_1 x_i$  are called the **fitted values**. The values  $b_0$  and  $b_1$  obtained by this method are called the **least squares estimates** of the parameters<sup>4</sup>.

The differences between the actual response values and the least squares fitted values are called the **residuals** of the model and are denoted by

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

Note that the residuals,  $e_i$ , can be considered to be estimates of the unknown errors  $\varepsilon_i$ . The residuals will however be ‘on average’ smaller than the errors because we have adjusted the position of the line specifically to make them as small as possible.

We would really like to be able to partition each response value  $y_i$  into two parts — part that depends on the explanatory variable,  $\beta_0 + \beta_1 x_i$ , and an ‘unpredictable’ error,  $\varepsilon_i$ , but the best we can do in practice is to split the observation into a fitted value,  $\hat{y}_i = b_0 + b_1 x_i$ , and a residual,  $e_i$ .

#### Theoretical partition of response

$$y_i = \underbrace{\beta_0 + \beta_1 x_i}_{\substack{\text{systematic} \\ \text{(non-random) part}}} + \underbrace{\varepsilon_i}_{\substack{\text{error} \\ \downarrow}}$$

#### Best partition of response in practice

$$y_i = \underbrace{b_0 + b_1 x_i}_{\substack{\text{fitted value} \\ \hat{y}_i}} + \underbrace{e_i}_{\substack{\text{residual} \\ \downarrow}}$$

The concepts of

- errors* ( $\varepsilon_i$  — random variables whose values we cannot determine exactly),
- fitted values* ( $\hat{y}_i$  which predict  $y_i$  from  $x_i$  using estimates  $b_0, b_1$  of the parameters  $\beta_0, \beta_1$ ),
- residuals* ( $e_i$  which approximate the errors and are  $e_i = y_i - \hat{y}_i$ ),
- Least Squares* (choosing the parameter estimates  $b_0, b_1$  to minimise  $\sum e_i^2$ )

are extremely important, so make sure that you understand them thoroughly.<sup>5</sup>

<sup>4</sup> Other estimation methods are also possible. For example, we could minimise  $\sum_{i=1}^n |y_i - \hat{y}_i|$ , which is called least absolute deviations. However it can be proved that least squares is optimum when the distribution of Y is normal.

<sup>5</sup> Although we will not examine more complex models until later in the course, it is worth noting here that the principle of least squares does extend to all regression models with a similar motivation. In the most general regression model,

$$Y = f(x_1, x_2, \dots, x_p, \beta_0, \beta_1, \dots, \beta_q) + \varepsilon$$

where the subscripts on the  $\{x\}$  now refer to the  $p$  explanatory variables. Again we estimate the unknown parameters  $\beta_0, \beta_1, \dots, \beta_q$  to minimise the error sum of squares,

$$SS_{\text{Error}} = \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

where now  $\hat{y}_i = f(x_{i1}, x_{i2}, \dots, x_{ip}, b_0, b_1, \dots, b_q)$  and the notation  $x_{ij}$  is used to denote the value of the  $j$ th explanatory variable corresponding to  $y_i$ .



# Least Squares Estimates and their Properties

For the simple linear model, the least squares estimates of  $\beta_0$  and  $\beta_1$  are<sup>6</sup>

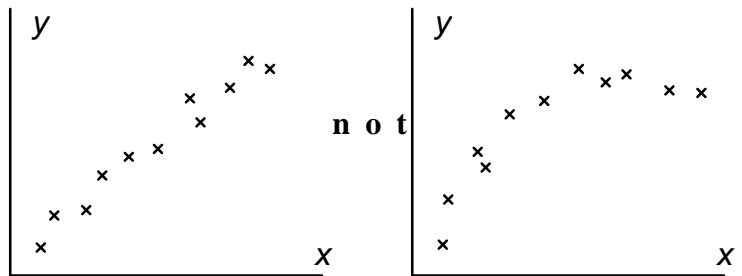
$$b_1 = \frac{S_{xy}}{S_{xx}} \quad \text{and} \quad b_0 = \bar{y} - b_1 \bar{x}$$

where

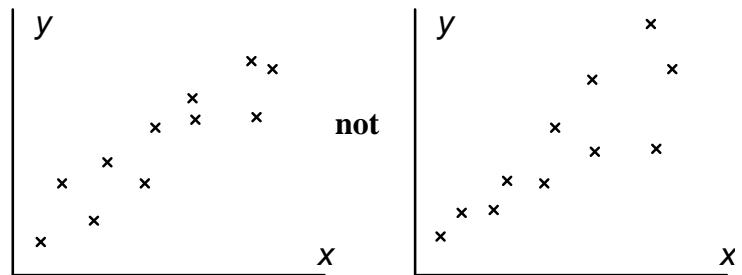
$$S_{xx} = \sum (x_i - \bar{x})^2 \quad \text{and} \quad S_{xy} = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$$

Further analysis requires that some or all of the assumptions underlying the simple linear model hold. It is therefore worthwhile to separate out the five major assumptions that are entailed by the model

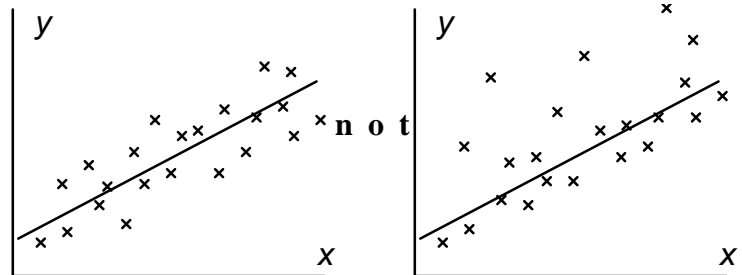
1. The  $\{x_{ij}\}$  are recorded without error.
2.  $E[y_i] = \beta_0 + \beta_1 x_i$  — the relationship between the variables is linear.



3.  $\text{Var}(y_i) = \sigma^2$  — the variance of the response is the same, whatever the value of the explanatory variables.



4. The errors  $\varepsilon_i = y_i - (\beta_0 + \beta_1 x_i)$  are uncorrelated with each other. In particular, if the measurements are made in order, successive measurements should not be correlated.
5. The distribution of the errors  $\varepsilon_i$  (and hence the  $y_i$ ) is normal.



<sup>6</sup>The proof is not particularly hard, but would not help you to understand the concept of least squares. We have therefore omitted the proof of this and most other results.

If some of these assumptions hold, it can be proved that ...

- The estimators are unbiased,

$$E[b_0] = \beta_0 \quad \text{and} \quad E[b_1] = \beta_1$$

This relies only on properties 1. and 2. above.

- The estimators have variances,

$$\text{Var}(b_0) = s^2 \left( \frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right) \quad \text{and} \quad \text{Var}(b_1) = \frac{s^2}{S_{xx}}$$

This relies on properties 1, 2, 3 and 4 only.

- The estimators have normal distributions. This requires all five properties above.

In most of what follows, we will assume that all five properties hold.

## Estimating the Error Variance

There is a third unknown parameter in the simple linear model, the error variance  $\sigma^2$ . Since the formulae above for the variances of  $b_0$  and  $b_1$  involve  $\sigma^2$ , it too must be estimated.

If we knew the values of  $\beta_0$  and  $\beta_1$ , and hence the errors  $\varepsilon_i = y_i - (\beta_0 + \beta_1 x_i)$ , we could estimate  $\sigma^2$  with the estimator

$$\hat{s}^2 = \frac{\sum \mathbf{e}_i^2}{n}$$

(Remember that each  $\varepsilon_i$  has mean 0, so  $E[\varepsilon_i^2] = \text{Var}(\varepsilon_i) = \sigma^2$ .) If the errors  $\varepsilon_i$  were known, their sample variance

$$\hat{s}^2 = \frac{\sum (\mathbf{e}_i - \bar{\mathbf{e}})^2}{n-1}$$

would be another, equally impractical (since we do **not** know the values of the errors  $\varepsilon_i$ ), estimator of  $\sigma^2$ . This estimator reduces the errors ‘on average’ by subtracting their mean from each, so we must divide by  $(n-1)$  instead of  $n$  to compensate.

We do not actually know the values of  $\varepsilon_i$ , but we have the least squares residuals,  $e_i$ , which are ‘on average’ smaller still, since **two** parameters have been adjusted to make them as small as possible. The appropriate formula for estimating  $\sigma^2$  from the residuals is

$$s^2 = \frac{\sum e_i^2}{n-2} = \frac{\text{SS}_{\text{Residual}}}{n-2}$$

This, unlike the previous two formulae, is one that can be evaluated in practice and is the estimator of  $\sigma^2$  that should be used. It can also be proved that this estimate has a chi-squared ( $\chi^2$ ) distribution

$$s^2 \sim \frac{s^2}{n-2} \times \mathbf{c}_{(n-2)df}^2$$

and that  $s^2$  is independent of  $b_0$  and  $b_1$ .

---

## Inference about $b_0$ and $b_1$

---

We can now use the results that the least squares estimates  $b_0$  and  $b_1$  are normal with

$$E[b_0] = \beta_0 \quad \text{and} \quad E[b_1] = \beta_1$$

$$\text{Var}(b_0) = s^2 \left( \frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right) \quad \text{and} \quad \text{Var}(b_1) = \frac{s^2}{S_{xx}},$$

together with the independent estimate of  $\sigma^2$ ,

$$s^2 = \frac{\sum e_i^2}{n-2}$$

which has a  $\chi^2$  distribution with  $(n-2)$  degrees of freedom, to find confidence intervals<sup>7</sup>, for  $\beta_0$  and  $\beta_1$

$$b_0 \pm t_{n-2} \times \sqrt{s^2 \left( \frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right)} \quad \text{and} \quad b_1 \pm t_{n-2} \times \sqrt{\frac{s^2}{S_{xx}}}$$

where  $t_{n-2}$  is the appropriate point from the  $t$  distribution with  $n-2$  degrees of freedom.

For hypothesis tests to test whether  $\beta_0 = k$  or  $\beta_1 = k$ , form the test statistic<sup>8</sup>,

$$t = \frac{b_0 - k}{\sqrt{s^2 \left( \frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right)}} \quad \text{or} \quad t = \frac{b_1 - k}{\sqrt{\frac{s^2}{S_{xx}}}}$$

respectively, and compare with tables of the  $t$  distribution with  $n-2$  degrees of freedom.

---

## Using the Computer to Fit the Model

---

The data below describe the results of experiments carried out by an Auckland concrete manufacturer to determine in what way, and to what extent, the hardness of a batch of concrete depends on the amount of cement used in making it. Forty batches of concrete were made up with varying amounts of cement in the mix and the hardness of each batch was measured after 7 days.<sup>9</sup>

---

<sup>7</sup> Both confidence intervals are of the form,

$$(\text{estimate}) \pm t \times (\text{estimated s.e. of estimate})$$

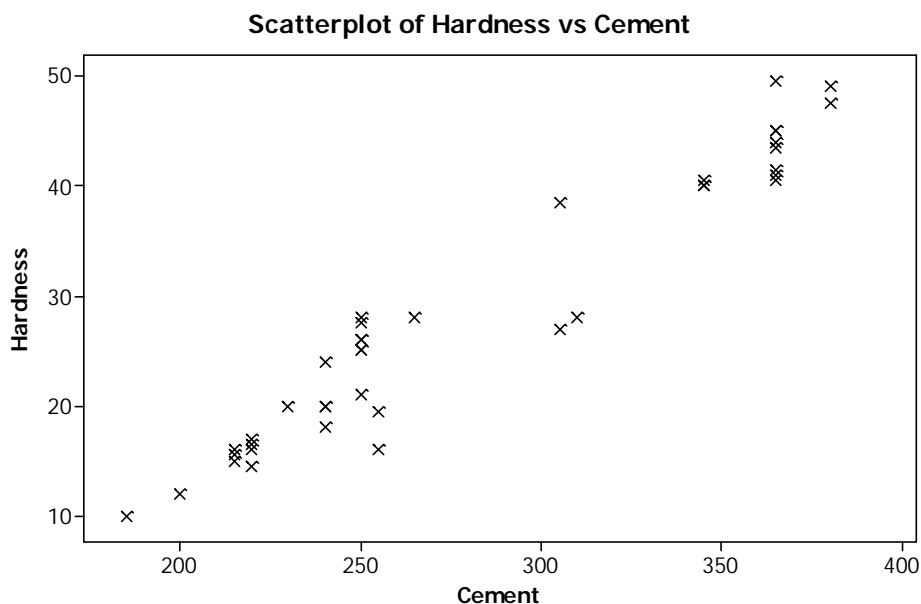
<sup>8</sup> Both are of the form,

$$t = \frac{\text{estimate} - k}{\text{estimated s.e. of estimate}}$$

<sup>9</sup> From "Introduction to Probability & Statistics" by C.J.Wild and G.A.F.Seber.

Cement	Hardness	Cement	Hardness	Cement	Hardness
365	45.0	200	12.0	365	41.0
220	17.0	185	10.0	250	27.5
240	18.0	305	27.0	265	28.0
215	16.0	305	38.5	365	45.0
255	19.5	250	26.0	345	40.0
220	16.0	345	40.5	380	47.5
250	26.0	250	25.0	365	49.5
230	20.0	365	41.5	215	15.5
240	24.0	220	16.5	310	28.0
250	28.0	365	43.5	345	40.0
240	20.0	240	20.0	215	15.0
250	21.0	365	40.5	380	49.0
220	14.5	250	25.0	255	16.0
365	44.0				

Before formally analysing the data, we should draw a scatterplot to confirm that a linear model is reasonable. Assuming that the data have been entered into two columns in Minitab with names *Cement* and *Hardness*, a scatterplot is produced. (For information on producing this graph in Minitab see the appendix at the end of this Study Guide).



As a linear relationship seems reasonable, regression is performed (Instructions in the Appendix)

Minitab responds with the following output (in the Session Window)

#### Regression Analysis: Hardness versus Cement

The regression equation is  
Hardness = - 24.1 + 0.186 Cement

Predictor	Coef	SE Coef	T	P
Constant	-24.067	2.298	-10.47	0.000
Cement	0.186471	0.007974	23.38	0.000

S = 3.10605    R-Sq = 93.5%    R-Sq(adj) = 93.3%

and in SAS

The SAS System

11:34 Monday, April 26, 2004 1

The REG Procedure  
Model: MODEL1  
Dependent Variable: Hardness Hardness

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	5275.16848	5275.16848	546.79	<.0001
Error	38	366.60652	9.64754		
Corrected Total	39	5641.77500			

Root MSE	3.10605	R-Square	0.9350
Dependent Mean	28.42500	Adj R-Sq	0.9333
Coeff Var	10.92717		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr >  t
Intercept	Intercept	1	-24.06656	2.29790	-10.47	<.0001
Cement	Cement	1	0.18647	0.00797	23.38	<.0001

This output provides least squares estimates of the parameters  $\beta_0$  and  $\beta_1$ ,

$$b_0 = -24.067 \text{ and } b_1 = 0.186471$$

It also provides an estimate of  $\sigma$ ,

$$\hat{s} = 3.106$$

The standard errors of the parameters are also given

$$se(b_0) = \sqrt{Var(b_0)} = 2.298$$

$$se(b_1) = \sqrt{Var(b_1)} = 0.007974$$

From these quantities, we can obtain 95% confidence intervals for  $\beta_0$  and  $\beta_1$

$$b_0 \pm t_{n-2} \times se(b_0) \quad \text{and} \quad b_1 \pm t_{n-2} \times se(b_1)$$

and these may be evaluated, after using t-tables to find  $t_{n-2} = t_{38} = 2.02$ , to give

$$\beta_0: -24.067 \pm 2.02 \times 2.298 = -28.71 \text{ to } -19.43$$

$$\beta_1: 0.186471 \pm 2.02 \times 0.007974 = 0.17036 \text{ to } 0.20258$$

Minitab also evaluates the t-ratios for testing whether  $\beta_0 = 0$  and for testing whether  $\beta_1 = 0$ .

These are

$$t = \frac{b_0 - 0}{\sqrt{\hat{s}^2 \left( \frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right)}} = -10.47 \quad \text{and} \quad t = \frac{b_1 - 0}{\sqrt{\frac{\hat{s}^2}{S_{xx}}}} = 23.38$$

Minitab saves you from looking these t-ratios up in tables by providing p-values for the two tests; both are reported as “p = 0.000” above, which should be interpreted as being “p < 0.0005”. We therefore conclude that there is extremely strong evidence in the data that both  $\beta_0 \neq 0$  and  $\beta_1 \neq 0$ .

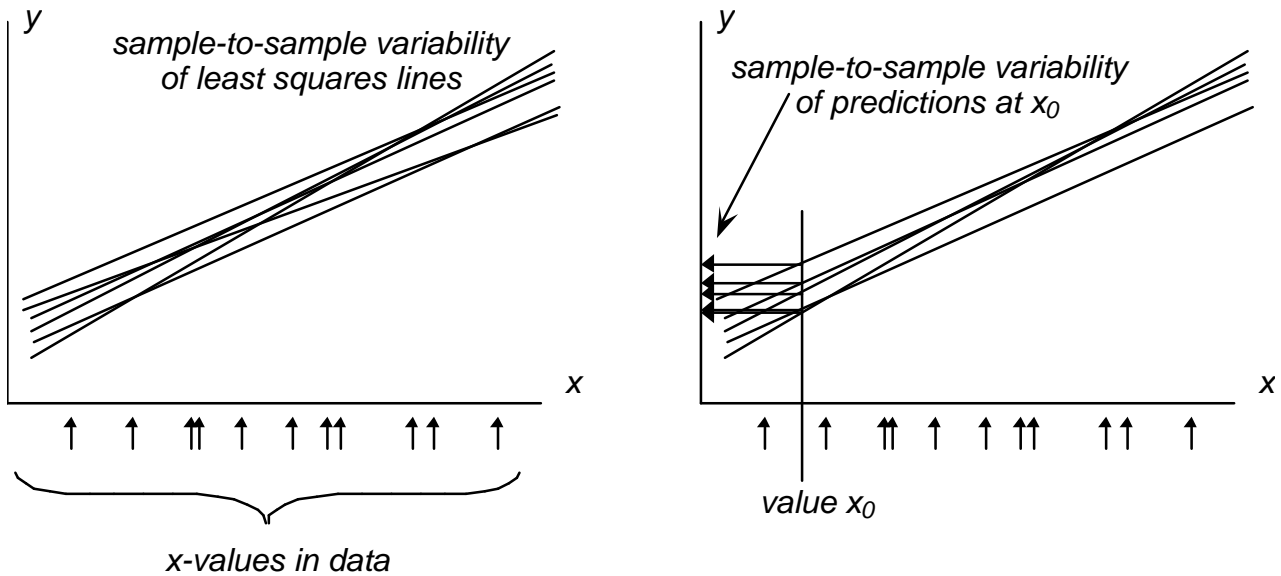
Additional information is provided in the Minitab output which will be described later in this section and in Section 2.

## Predictions

One of the most important uses of regression models is to predict the response that will be recorded at new values of the explanatory variables,  $x_0$ . The prediction is

$$\hat{y}(x_0) = b_0 + b_1 x_0$$

Since the least squares coefficients are random (i.e. vary from sample to sample), the prediction that will be made is also a random variable.

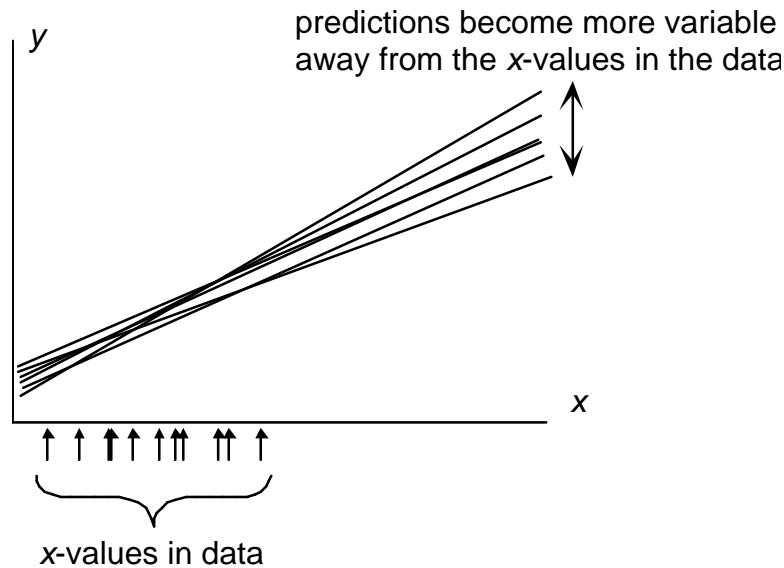


If the model assumptions hold, it can be proved that the prediction has a normal distribution with mean and standard deviation

$$E[\hat{y}(x_0)] = b_0 + b_1 x_0$$

$$se(\hat{y}(x_0)) = s \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}$$

Note that the predictions become more variable when  $x_0$  becomes further from the mean of the data,  $\bar{x}$ , as illustrated in the diagram below.



We can use this to find a confidence interval for the mean response at  $x_0$ ,  $\beta_0 + \beta_1 x_0$

$$\hat{y}(x_0) \pm t_{n-2} \times s \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}$$

where  $\sigma$  in the equation for  $se(\hat{y}(x_0))$  is replaced by its best estimate from the residuals,  $s$ , and  $t_{n-2}$  is the appropriate value from the t-tables with  $n - 2$  degrees of freedom.

Note that we have given a confidence interval for the **mean** response at  $x_0$ ,  $\beta_0 + \beta_1 x_0$ . This does not take into account the fact that a single new observation will not be at exactly  $\beta_0 + \beta_1 x_0$ , but will be at a distance  $\varepsilon$  from this,

$$y_{\text{new}} = \beta_0 + \beta_1 x_0 + \varepsilon$$

The prediction error,  $y_{\text{new}} - \hat{y}(x_0)$  will be greater. In fact it can be shown that

$$se(y_{\text{new}} - \hat{y}(x_0)) = s \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}$$

A prediction interval for predicting a **new** response at  $x_0$  is therefore

$$\hat{y}(x_0) \pm t_{n-2} \times s \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}$$

**Note carefully the distinction between estimating the mean response and predicting a single new response at  $x_0$ .**

The computer will find these confidence and prediction intervals for you, see the Appendix for details. For example, if we are interested in predicting hardness of concrete when the cement content is 250,

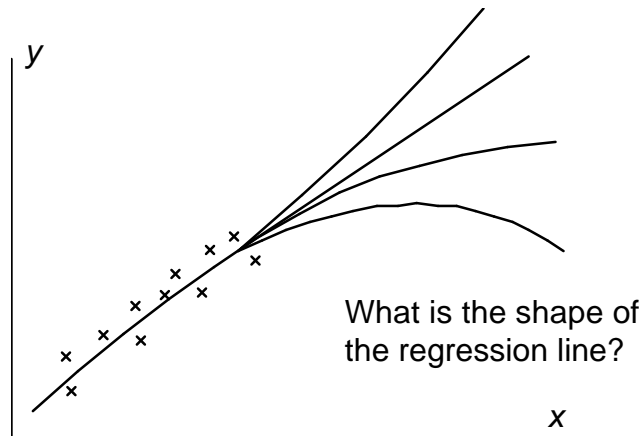
The following output would be produced:

Predicted Values for New Observations

New	Fit	SE Fit	95% CI	95% PI
Obs				
1	22.551	0.552	(21.434, 23.668)	(16.165, 28.937)

We therefore have 95% confidence that the mean strength of concrete with cement content 250 will be between 21.434 and 23.668. However if a single sample of concrete with this cement content was obtained, we would be less certain of its hardness — our 95% prediction interval for this sample would be between 16.163 and 28.939.

The equations above for  $se(\hat{y}(x_0))$  and  $se(y_{\text{new}} - \hat{y}(x_0))$  both depend on  $(x_0 - \bar{x})^2$ . This shows that predictions become less accurate the further  $x_0$  is from the mean of the data,  $\bar{x}$ . However there is a second problem with using a least squares line to predict a response at an  $x$ -value that is not close to the  $x$ -values in the data that has been collected. We have no information in the data about whether the relationship is linear in ranges of  $x$ -values where we have no data. There is also the potential for errors from assuming the relationship is linear when it is not.



**Errors that arise from model mis-specification do not show up in the confidence intervals for the predictions.**

**Since there is no way to detect or adjust for this problem, avoid using regression models to make predictions at values of the explanatory variables where there is no data.**

**AVOID EXTRAPOLATION.**

---

## Coefficient of Determination

---

The strength of a linear relationship between a response and explanatory variable may be described by the correlation coefficient,

$$R = \frac{\sum (y_i - \bar{y})(x_i - \bar{x})}{\sqrt{\sum (y_i - \bar{y})^2 \sum (x_i - \bar{x})^2}}$$

Since  $\hat{y}_i = b_0 + b_1 x_i$  is a linear function of  $x_i$ , this is also the correlation coefficient of the response and fitted values,

$$R = \frac{\sum (y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sqrt{\sum (y_i - \bar{y})^2 \sum (\hat{y}_i - \bar{\hat{y}})^2}}$$



A bit of algebraic manipulation shows that the square of the correlation coefficient can be expressed in the form

$$R^2 = \frac{\sum (\hat{y}_i - \bar{\hat{y}})^2}{\sum (y_i - \bar{y})^2} = 1 - \frac{\sum e_i^2}{\sum (y_i - \bar{y})^2}$$

$R^2$  is called the **coefficient of determination** of the linear model. Since  $\sum (y_i - \bar{y})^2$  is a measure of the total variability of the response,  $\sum (\hat{y}_i - \bar{\hat{y}})^2$  measures the variability of the fitted values, and  $\sum e_i^2$  measures the variability of the residuals,  $R^2$  can be interpreted as the proportion of the total variability that is explained by the model.

The coefficient of determination,  $R^2$ , can be extended easily to more complex models and is the most commonly used descriptive measure of the strength of a relationship. It will be discussed in more detail later in the course.

Most packages automatically provides the value of  $R^2$  when a model is fitted. For example, the output from fitting a linear model to the concrete-hardness data included the lines

In Minitab

S = 3.10605	R-Sq = 93.5%	R-Sq(adj) = 93.3%
-------------	--------------	-------------------

And in SAS

Root MSE	3.10605	R-Square	0.9350
Dependent Mean	28.42500	Adj R-Sq	0.9333
Coeff Var	10.92717		

The model therefore explains 93.5% of the total variability of concrete hardness (and the correlation coefficient of *Hardness* and *Cement* is  $\pm\sqrt{0.935}$ ).

Using  $R^2$  to assess the fit of the model is not a simple exercise. Some courses may suggest that  $R^2$  values greater than 50% are “good”.  $R^2$  needs to be assessed in the context of the problem being considered. In a Physics experiment, where variables can be carefully controlled,  $R^2$  values of 95% or more would be expected. However in social sciences and animal behaviour experiments an  $R^2$  of 35% may represent a good result.  $R^2$  is one of a number of diagnostics which can be used to build a picture of the usefulness of the model. We will look at this more in Chapter 2

---

## Random Explanatory Variable

---

In the above sections, we have implicitly assumed that the values of the explanatory variable,  $x_1, x_2, \dots, x_n$ , are fixed constants. In many applications of regression in the experimental sciences, this is true (since the values of the explanatory variable are fixed by the experimenter before the experiment is conducted), but this is not so in observational studies where there is no control over any of the measurements. For example, consider a study to determine how soil quality affects the yield of a crop. A series of plots of farmland might be marked out, with both soil quality and yield measured from each plot. Both the response, yield, and explanatory variable, soil quality, are random quantities.

Fortunately, we do not need to develop new methods to deal with this situation. We use regression to model the response **conditional** on the observed values of the explanatory variable. In other words, our model is that the conditional distribution of  $Y_i$  given that  $X_i = x_i$  is

$$Y_i \mid x_i \sim \text{normal}(\beta_0 + \beta_1 x_i, \sigma^2)$$

This particularly makes sense when there a sequential ordering involved — when the response is determined **after** the explanatory variable, as is the case with the crop yield example above — the soil quality exists before the crop is grown.

In practice, by specifying a conditional regression model of this form, we can simply ignore the randomness of the explanatory variable when the data are analysed.

### Correlation coefficient

When both the response and explanatory variables in the model are random, we might also ask questions about the population correlation coefficient between the two variables,  $\rho$ . How do we perform inference about  $\rho$ ?

As was noted in the previous section, the coefficient of determination,  $R^2$ , is the square of the correlation coefficient between the response and explanatory variables; the sample correlation coefficient,  $R$ , is a point estimate of  $\rho$ . For a confidence interval, we must use the approximation

$$\frac{1}{2} \ln \left( \frac{1+r}{1-r} \right) \sim \text{normal} \left( \mathbf{m} = \frac{1}{2} \ln \left( \frac{1+\mathbf{r}}{1-\mathbf{r}} \right), \mathbf{S}^2 = \frac{1}{n-3} \right)$$

From this, we obtain the large-sample 95% confidence interval

$$\frac{1}{2} \ln \left( \frac{1+r}{1-r} \right) - 1.96 \sqrt{\frac{1}{n-3}} < \frac{1}{2} \ln \left( \frac{1+\mathbf{r}}{1-\mathbf{r}} \right) < \frac{1}{2} \ln \left( \frac{1+r}{1-r} \right) + 1.96 \sqrt{\frac{1}{n-3}}$$

After evaluating the left and right of this inequality, it can be reexpressed as an inequality in  $\rho$  itself. For example, if  $r = 0.762$  from a data set of size  $n = 50$ ,

$$\begin{aligned} \frac{1}{2} \ln \left( \frac{1+0.762}{1-0.762} \right) - 1.96 \sqrt{\frac{1}{47}} &< \frac{1}{2} \ln \left( \frac{1+\mathbf{r}}{1-\mathbf{r}} \right) < \frac{1}{2} \ln \left( \frac{1+0.762}{1-0.762} \right) + 1.96 \sqrt{\frac{1}{47}} \\ 0.715 &< \frac{1}{2} \ln \left( \frac{1+\mathbf{r}}{1-\mathbf{r}} \right) < 1.2857 \\ \frac{e^{2 \times 0.715} - 1}{e^{2 \times 0.715} + 1} &< \mathbf{r} < \frac{e^{2 \times 1.287} - 1}{e^{2 \times 1.287} + 1} \\ 0.614 &< \mathbf{r} < 0.858 \end{aligned}$$

A hypothesis test for  $\rho$  can also be performed using this normal approximation, but the approach is not best when testing whether  $\rho = 0$ . The null hypothesis  $\rho = 0$  is equivalent to the hypothesis that  $\beta_1 = 0$  in the linear model formulation, and so it should be tested using the standard Minitab test of whether the regression slope is zero.

# **161.320 Fitting Regression Models**

## **2. Assessing and Correcting Lack of Fit**

---

### **Assumptions**

---

Use of the simple linear model and inference about its parameters are based on several assumptions. If these assumptions do not hold, use of the model may give totally misleading information about the system being modelled.

1.  $E[y_i] = \beta_0 + \beta_1 x_i$  — the relationship between the variables is linear.
2.  $\text{Var}(y_i) = \sigma^2$  — the variance of the response is the same, whatever the value of the explanatory variables.
3. The errors  $\varepsilon_i = y_i - (\beta_0 + \beta_1 x_i)$  are uncorrelated with each other. In particular, if the measurements are made in order, successive measurements should not be correlated.
4. The distribution of the errors  $\varepsilon_i$  (and hence the  $y_i$ ) is normal.

In this chapter, we will examine some graphical and numerical techniques that can be used to help assess whether or not these assumptions hold.

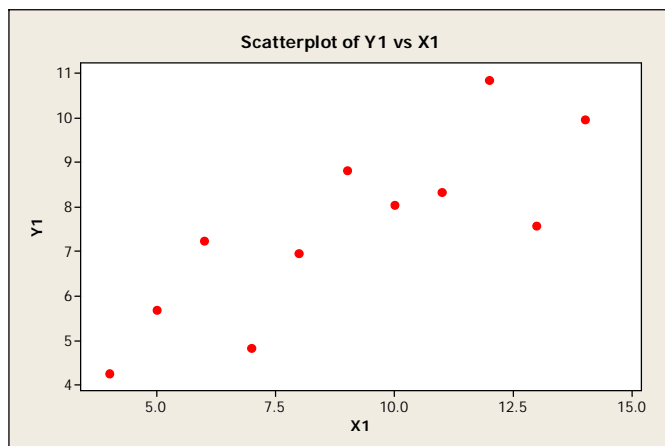
Standard summary statistics may not indicate any problems with the model. A scatterplot often reveals what  $R^2$  does not. Anscombe<sup>10</sup> uses four data sets to emphasise the importance of the scatterplot.

---

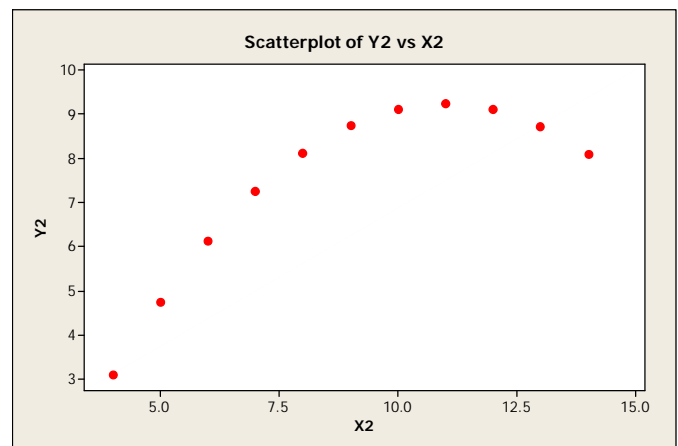
<sup>10</sup> Anscombe, FJ (1973) Graphs in Statistical Analysis, *American Statistician*, 27, 17-21.

Data Set 1		Data Set 2		Data Set 3		Data Set 4	
Y	X	Y	X	Y	X	Y	X
8.04	10.00	9.14	10.00	7.46	10.00	6.58	8.00
6.95	8.00	8.14	8.00	6.77	8.00	5.76	8.00
7.58	13.00	8.74	13.00	12.74	13.0	7.71	8.00
8.81	9.00	8.77	9.00	7.11	9.00	8.84	8.00
8.33	11.00	9.26	11.00	7.81	11.00	8.47	8.00
9.96	14.00	8.10	14.00	8.84	14.00	7.04	8.00
7.24	6.00	6.13	6.00	6.08	6.00	5.25	8.00
4.26	4.00	3.10	4.00	5.39	4.00	12.50	19.00
10.84	12.00	9.13	12.00	8.15	12.00	5.56	8.00
4.82	7.00	7.26	7.00	6.42	7.00	7.91	8.00
5.68	5.00	4.74	5.00	5.73	5.00	6.89	8.00

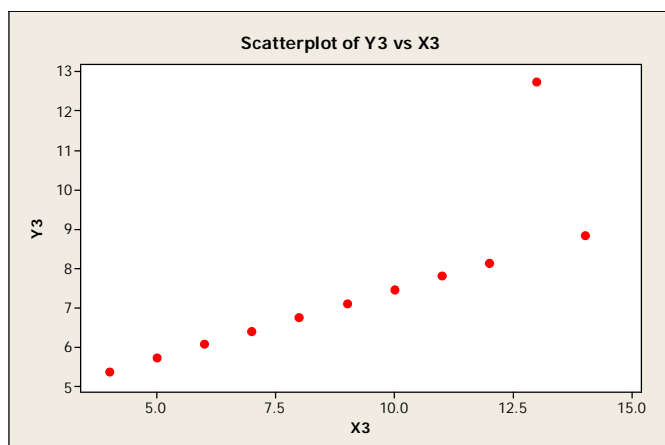
For each of these four data sets,  $\bar{x} = 9.0$ ,  $s_x = 3.317$ ,  $\bar{y} = 7.5$ ,  $s_y = 2.032$ , and  $R = 0.816$ . The 'basic' set of numerical summary statistics is therefore the same for all four data sets. That the data sets are, however, very different is clearly revealed by their scatterplots. The need to plot the data as part of the exploratory analysis of a relationship can not be over-emphasised.



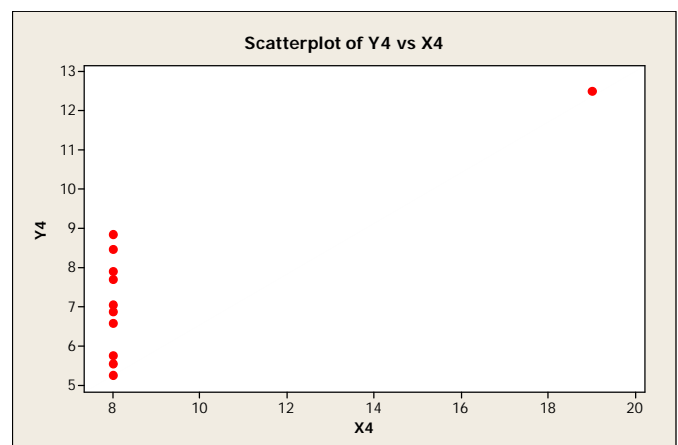
*Data Set 1: Positive linear association*



*Data Set 2: Perfect nonlinear association*



*Data Set 3: Perfect linear relationship except for one outlier*



*Data Set 4: No variability in the explanatory variable (and hence no evidence about the relationship) apart from that provided by one further x-value which is an outlier.*

Although the material in this chapter is aimed initially at the simple linear model (with a single explanatory variable), all the techniques can be easily extended to more complex models with two or more explanatory variables.

---

## Outliers

---

There are two potential problems with the assumption of linearity,

$$E[y_i] = \beta_0 + \beta_1 x_i$$

Firstly, the relationship may be curved, a problem that is addressed in the next section. Secondly, one or more ‘individuals’ about which data are recorded may have special characteristics that make their  $y$ -values different from what would be expected in the linear model — they may be outliers.

Outliers are often caused by errors in the measurement and recording process; these errors also result in wrong parameter estimates and inferences. Outliers may alternatively correspond to individuals whose characteristics are different from the bulk of the data. It is again wrong to use such ‘uncharacteristic’ individuals to make predictions for the bulk of the data.

The effect of a single outlier on the conclusions drawn from the data set can be great, depending on the value of the explanatory variable. (See the section on Leverage and Influence later in this chapter.)

If we could evaluate the model errors,

$$\varepsilon_i = y_i - (\beta_0 + \beta_1 x_i)$$

then they would provide a good indication of outliers. Since the errors are normally distributed, only 5% of them should be outside the range  $\pm 2\sigma$ , and virtually none should be outside the range  $\pm 3\sigma$ . Unusually large or small residuals could therefore be classified as outliers. Of course, since we do not know the values of  $\beta_0$  or  $\beta_1$ , the errors are unknown in practice.

Since the least squares residuals,

$$e_i = y_i - (b_0 + b_1 x_i)$$

are estimates of the errors  $\varepsilon_i$ , it would initially appear that they could be used in the same way, but this is not so. There are two problems.

Firstly, although the unknown errors,  $\varepsilon_i$  all have variance  $\sigma^2$ , the residuals  $e_i$  do not all have the same variance. In fact,<sup>11</sup>

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

where  $h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}}$ . From this formula, clearly  $h_{ii} = 0$ , and since  $\text{Var}(e_i) = 0$ ,  $h_{ii} = 1$ . The residuals therefore have variances that are at least as small as the errors (i.e.  $\text{Var}(e_i) = \sigma^2$ ). Further, since

$$\begin{aligned} \sum_{i=1}^n \text{Var}(e_i) &= \sum_{i=1}^n \sigma^2 (1 - h_{ii}) \\ &= \sigma^2 \left( n - \sum_{i=1}^n \left( \frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}} \right) \right) \\ &= \sigma^2 \left( n - 1 - \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{S_{xx}} \right) = (n - 2)\sigma^2 \end{aligned}$$

the average variance of the residuals is less than  $\sigma^2$ .

Even more important in practice is the fact that the residual variance depends on  $(x_i - \bar{x})^2$ , so residuals corresponding to  $x_i$  that are far from  $\bar{x}$  have smaller variance than those that are close to  $\bar{x}$  — the regression line is pulled closer to these data values than to data values near  $\bar{x}$ .

To assess which residuals are unexpectedly large, we should therefore first standardise them by dividing by their standard deviation, giving the **studentised residuals**,

$$r_i = \frac{e_i}{\hat{\sigma} \sqrt{1 - h_{ii}}}$$

Studentised residuals should only be outside the range  $\pm 2$  with probability 5%, and virtually never outside the range  $\pm 3$ .

There is however a second problem with using the model's residuals to detect outliers. The residuals are based on the least squares line and its position may be greatly influenced by the outlier.

We'd hope for these residuals...

but what we get are these

---

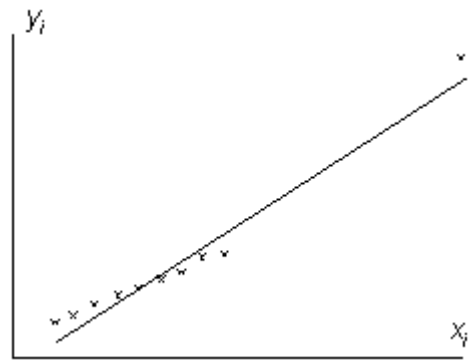
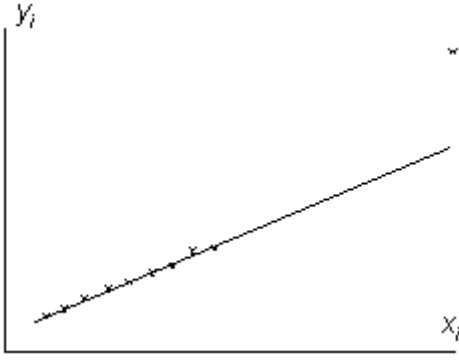
<sup>11</sup>The formula for  $h_{ii}$  may seem familiar. If you look back to the section of Chapter 1 dealing with predictions, you will find that

$$\text{Var}(\hat{y}_i) = \left( \frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}} \right) \sigma^2 = h_{ii} \sigma^2$$

Therefore

$$\text{Var}(y_i) = \sigma^2 = \text{Var}(\hat{y}_i) + \text{Var}(e_i)$$

where, of course,  $y_i = \hat{y}_i + e_i$ .



One solution to this problem is to find the fitted values corresponding to each  $x_i$  using all the data except for that data point. In other words, we delete  $x_i$  and  $y_i$  from the data set and fit a line by least squares to the remaining  $n - 1$  data points; using this least squares line, we find the prediction corresponding to  $x_i$ ,  $\hat{y}_{i,-i}$ , and the **deleted residual**, or **PRESS residual**,

$$e_{i,-i} = y_i - \hat{y}_{i,-i}$$

This procedure is repeated  $n$  times, deleting each data point in turn, and giving a deleted residual for each data point. It may initially seem computationally tedious to calculate the deleted residuals, but it can actually be proved that

$$e_{i,-i} = \frac{e_i}{1 - h_{ii}}$$

so the deleted residuals can actually be found as easily as the studentised ones.

Note that the deleted residuals do not all have the same variance. Whereas

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

it can be seen that

$$\text{Var}(e_{i,-i}) = \frac{\sigma^2}{1 - h_{ii}}$$

The deleted residuals now have a higher variance when  $x_i$  is far from  $\bar{x}$ . Standardising the deleted residuals (dividing by their standard deviation) results in the same studentised residuals that we defined earlier!

$$r_i = e_{i,-i} \times \frac{\sqrt{1 - h_{ii}}}{\hat{\sigma}}$$

An alternative way to standardise the deleted residuals is more popular; the deleted residuals may be standardised by using an estimate of  $\sigma$  that is also based on the data without the  $i$ 'th point,  $s_{-i}$  instead of the estimate  $\hat{\sigma}$  based on all the data. This is called the **externally studentised residual**, and is sometimes called **R-student**,

$$t_i = e_{i,-i} \times \frac{\sqrt{1 - h_{ii}}}{s_{-i}}$$

Like the ordinary studentised residual,  $t_i$  may be informally compared with  $\pm 2$  or  $\pm 3$  to assess how unusually large it is. For small sample sizes, it can be proved that the 95% and 99% percentage points of the t-distribution with  $(n - 3)$  degrees of freedom are more appropriate to compare  $t_i$  against than the corresponding points from the normal distribution (1.96 and 2.58),

but in moderate or large data sets they could be assessed in the same way as the ordinary studentised residuals.

If an outlier is detected (or suspected), the initial reaction should be to carefully check the observation. Was there an error in transcribing the data point or typing it into the computer for analysis? Was there anything unusual about the 'individual' from which the measurements were made?

If there is auxiliary information to confirm that the outlier is indeed different from the remaining values, or if the outlier is 'extreme', then it should be deleted from the data set before continuing with the analysis.

Minitab automatically flags any observations with standardised residuals that are outside the range  $\pm 2$  in the output from any regression analysis. See the Appendix for the parallel SAS analysis. For example, in the output from fitting a model explaining the variability of the hardness of concrete against the amount of cement used in its manufacture (data that was analysed in Chapter 1), the Minitab output is

```
MTB > Regress 'Hardness' 1 'Cement'.
```

```
The regression equation is
```

```
Hardness = - 24.1 + 0.186 Cement
```

```
...
```

```
Unusual Observations
```

Obs.	Cement	Hardness	Fit	Stdev.Fit	Residual	St.Resid
40	255	16.000	23.484	0.535	-7.484	-2.45R

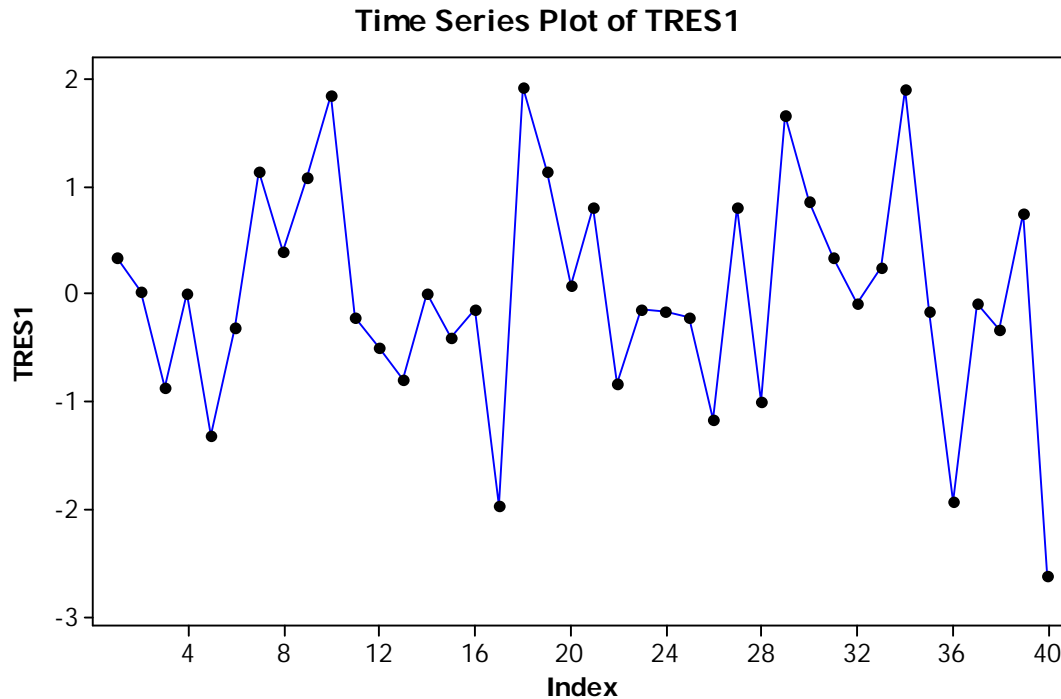
```
R denotes an obs. with a large st. resid.
```

Observation 40 has a standardised residual of -2.45. Comparing this value with normal tables, we find a probability of 0.014 of obtaining a residual as far from zero as this. However since there are 40 residuals, it is perhaps not surprising that one is this small. We should (if possible) check our data to make sure that there were no transcription errors and that there was nothing special about this run of the experiment (it was the final run), but if there was no supporting evidence, the size of the studentised residual is not large enough on its own to cause concern.



For further analysis of the residuals, we must ask for them to be evaluated for us when the model is fitted. For detailed description in Minitab and SAS see the appendix.

Although the columns of residuals may be scanned, it is usually better to plot them against either the fitted values, an explanatory variable or simply time order in order to assess whether any observations stand out and should be classified as outliers. For example, plotting the deleted residuals against their time order gives the following graph..



The final observation's deleted t residual does not stand out from the other residuals in this plot.

---

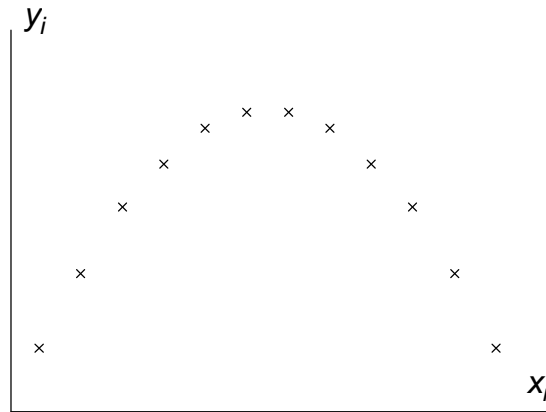
## Nonlinearity

---

We next address the problem of curvature in the relationship between  $y$  and  $x$ ,

$$E[y_i] = f(x_i) \quad ? \quad \beta_0 + \beta_1 x_i$$

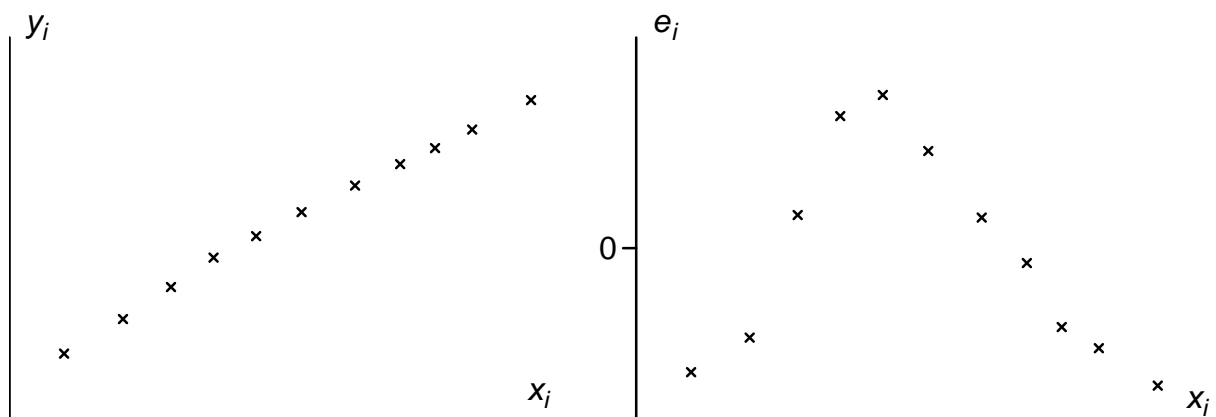
If the correct form of the relationship is not specified, the resulting parameter estimates do not describe the relationship. For example, in the following data set, the least squares slope is zero and the correlation coefficient is zero, even though the curved relationship is strong.



Even though such an extreme case is unusual, the correlation coefficient usually underestimates the strength of a curved relationship and gives poor predictions, especially outside the range of values of the explanatory variable in the data.

Before any analysis is done of regression data, it should be plotted in a scatterplot of  $y_i$  against  $x$ . Most occurrences of nonlinearity in data with a single explanatory variable are apparent in such a plot.

For data where the error variance  $\sigma^2$  is small however, nonlinearity is more effectively displayed in a scatterplot of the least squares residuals,  $e_i$ , against  $x_i$ .



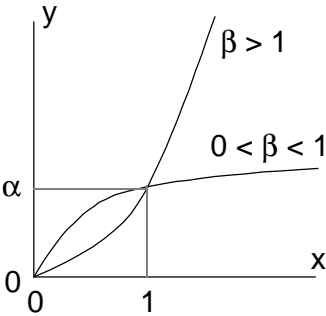
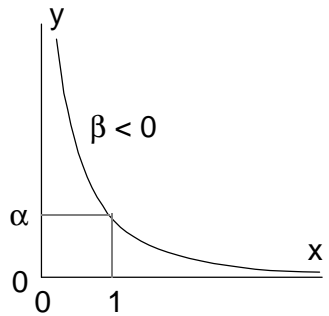
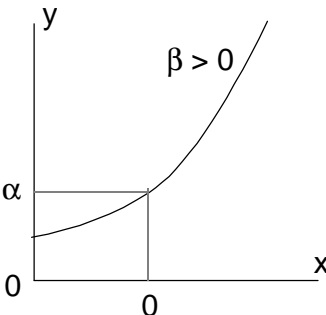
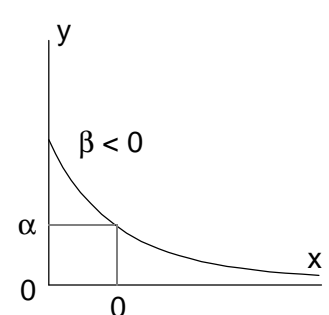
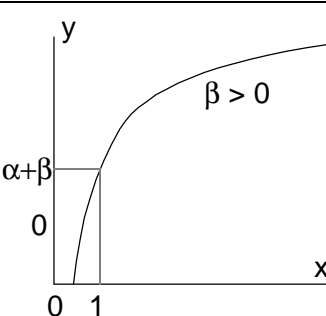
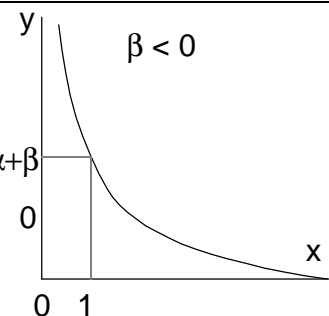
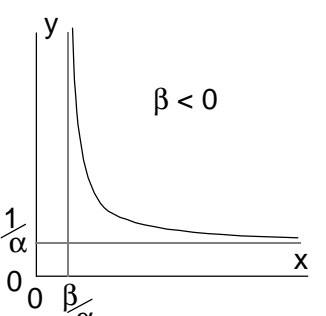
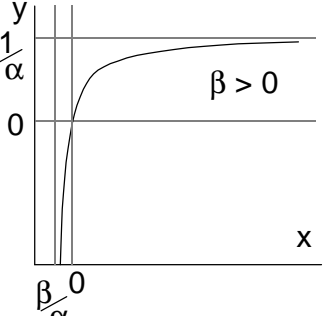
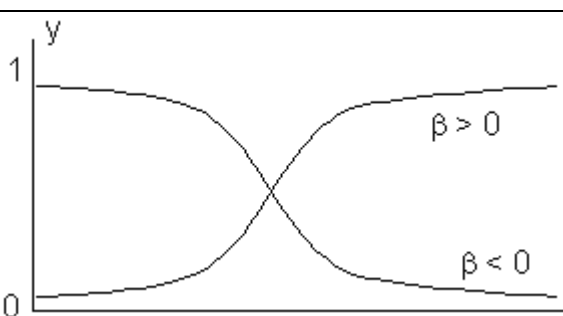
If we detect problems with the fit of a model, what can be done to improve the fit? There are two main solutions to the problem of nonlinearity. One is to add extra terms to the model involving higher powers of  $x$ ,

$$E[y_i] = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 \dots$$

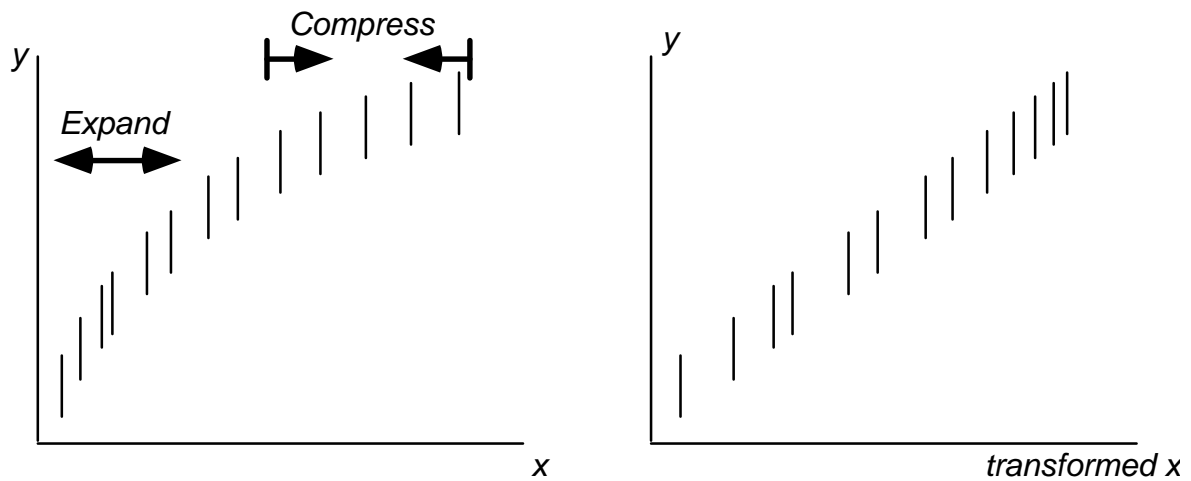
Polynomial models of this form are still linear models (since they are linear in the unknown parameters) but their analysis must wait until the theory of linear models has been extended to cover more than one explanatory variable.

The second solution to nonlinearity is to write the model as a simple linear model whose 'response' and 'explanatory' variables are transformations of the original measurements.

The following table gives a few types of relationship that are linear in terms of transformed response or explanatory variables. The relationship is then called ***intrinsically linear***.

Function	Transforms	Linear form	Graphs	
$y = \alpha x^\beta$	$y^* = \ln y$ $x^* = \ln x$	$y^* = \ln \alpha + \beta x^*$		
$y = \alpha e^{\beta x}$	$y^* = \ln y$	$y^* = \ln \alpha + \beta x$		
$y = a + b \ln x$	$x^* = \ln x$	$y = \alpha + \beta x^*$		
$y = \frac{x}{\alpha x - \beta}$	$y^* = y^{-1}$ $x^* = x^{-1}$	$y^* = \alpha - \beta x^*$		
$y = \frac{e^{\alpha + \beta x}}{1 + e^{\alpha + \beta x}}$	$y^* = \ln \left( \frac{y}{1 - y} \right)$	$y^* = \alpha + \beta x$		

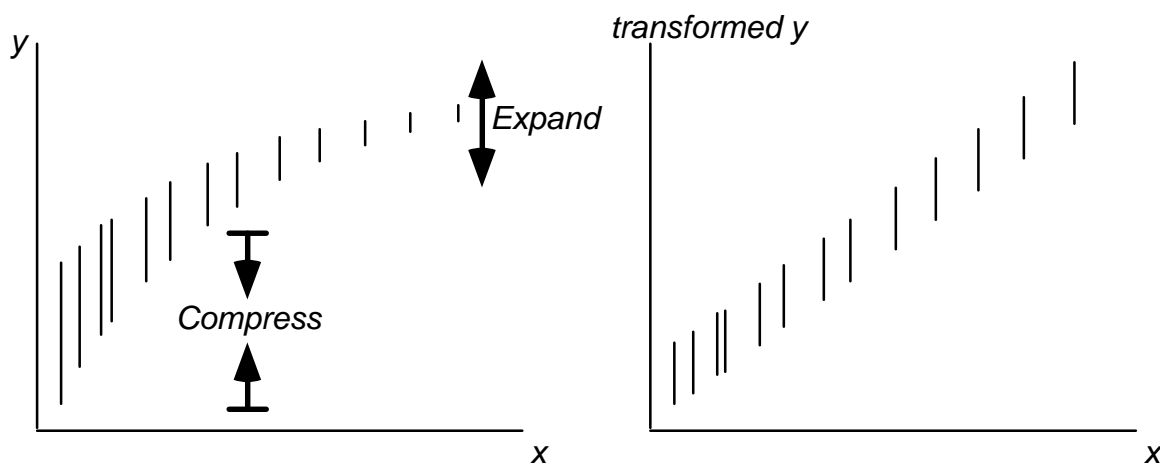
If a model can be linearised with a transformation of  $x$  only, then the assumptions of normality and constant variance for  $y$  are equivalent in the original and transformed models. The diagrams below show the 'typical' range of  $y$ -values for several  $x$ -values in a regression model before transformation, and the corresponding range after transforming  $x$  to straighten the relationship.



However if a nonlinear transformation of  $y$  is used, then the error structure will be different for  $y$  and the transformed response,  $y^*$ .

- If  $y$  is normal with the same variance for each  $x$ , then  $y^*$  will not have a symmetric distribution and its variance will be different for different  $x$ .
- If  $y^*$  is normal with the same variance for each  $x$ , then  $y$  will not have a symmetric distribution and its variance will be different for different  $x$ .

This is illustrated in the diagram below

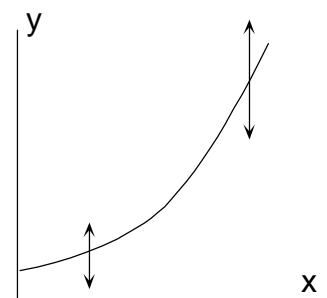


Our aim is for the **transformed** (linearised) model to satisfy the standard linear model assumptions. If this is to be so, we should expect a greater variability in  $y$  at regions of  $x$  where the slope of the  $y$  vs  $x$  curve is greatest.

Also, we would expect that the distribution of  $y$  would be skew with a longer upper tail if its curvature is concave up (as in the diagram on the right), and a longer lower tail if its curvature is convex up.

Luckily, it **usually** happens that when a relationship between  $y$  and  $x$  is curved, skewness and non-constant variance also occur and the transformation that linearises the relationship also helps cure the other problems.

However you **must** check the assumptions with residual plots for the transformed variables.

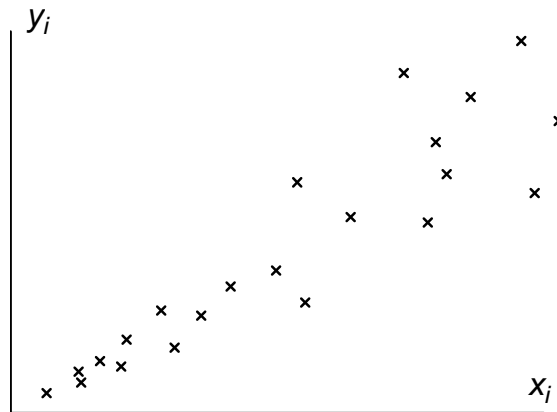


---

## Non-constant Variance

---

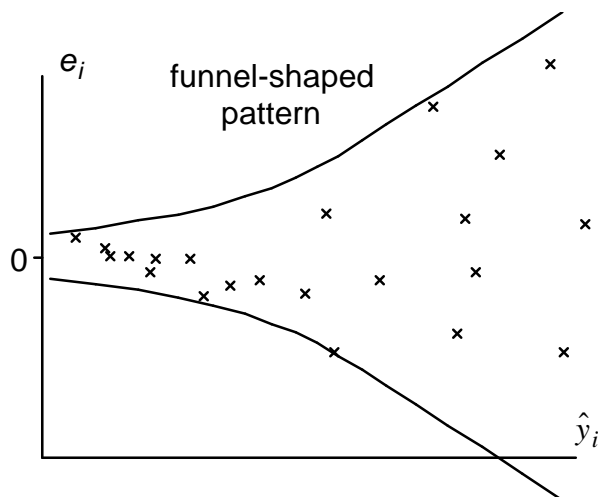
Another problem that may arise is that the response variance may not be the same for all observations. Although this may occur in different ways, it is most common for the response variance to be an increasing function of the response mean,



This most often occurs when the response is a non-negative 'quantity', such as a concentration of chemical or a count of insects.

The fitted line should be closer to the data (on average) where the response variance is smallest. Least squares does not enforce this and therefore gives less accurate predictions than other fitting methods.

Although this may be detected in a scatterplot of  $y_i$  against  $x_i$  in the case of simple linear regression, it is best detected from a scatterplot of the residuals,  $e_i$ , against the fitted values,  $\hat{y}_i$ .



This plot can also be used for models with more than one explanatory variable.

Non-constant variance (also called **heteroscedasticity**) is often associated with non-linearity and with non-normal error distributions. Transformation of the response to give linearity often also helps to remedy non-constant variance, as described in the previous section.

In most situations, a suitable transformation of the response is obtained empirically to correct for non-constant variance (e.g. a power transformation), some results are available when the response has a known non-normal distribution.

The table below gives a few transformations called **variance-stabilising** transformations that can be used when the distribution of the response is known, but is not a normal distribution. The transformed response has approximately constant variance and is more symmetrical than the original response. The transformed response is however not normally distributed and the use of variance-stabilising transformations provides only an approximate analysis.<sup>12</sup>

Distribution of $y$	Variance( $y$ ) in terms of $\mu$	Transformation	Variance	Alternative transform
Poisson	$\mu$	$\sqrt{y}$	0.25	$(\sqrt{y} + \sqrt{y+1})$ or $\sqrt{y+0.25}$
Binomial	$\frac{\mu(1-\mu)}{n}$	$\sin^{-1} \sqrt{y/n}$ (degrees)	$\frac{821}{n}$	$\sin^{-1} \sqrt{y/n}$ (radians)
Negative binomial <sup>13</sup>	$\mu + \lambda^2 m^2$	$\lambda^{-1} \sinh^{-1}(\lambda \sqrt{y})$	0.25	$\lambda^{-1} \sinh^{-1}(\lambda \sqrt{y} + 0.5)$

Transformations of the response also affect the linearity (or nonlinearity) of the relationship and the shape of the distribution of the response. If the original relationship is linear with symmetrically distributed errors but the error variance increases with  $x$ , transformation of the response may correct the non-constant variance, but will introduce nonlinearity and a skew error distribution. A different solution is required.

We can perform an analysis using a simple linear regression model in one special case. If we assume that the error standard deviation is proportional to  $x_i$

$$E[y_i] = \beta_0 + \beta_1 x_i$$

$$\text{Var}(y_i) = x_i^2 \sigma^2$$

then we can rewrite the model in terms of a 'response'  $y_i^* = \frac{y_i}{x_i}$

$$E[y_i^*] = \frac{\beta_0 + \beta_1 x_i}{x_i} = \beta_1 + \beta_0 \frac{1}{x_i} = \beta_0^* + \beta_1^* x_i^*$$

$$\text{Var}(y_i^*) = \text{Var}\left(\frac{y_i}{x_i}\right) = \frac{x_i^2 \sigma^2}{x_i^2} = \sigma^2$$

Therefore in terms of 'response'  $y_i^* = \frac{y_i}{x_i}$  and explanatory variable  $x_i^* = \frac{1}{x_i}$ , the model is linear with constant variance. The regression parameters are those in the original formulation of the model, but note that the transformed variables themselves are usually difficult to interpret.

<sup>12</sup>An exact analysis of Poisson and binomial data is possible, but is beyond the scope of this course. The paper *Models for Non-Normal Data* extends the topics of linear models and regression to non-normal distributions.

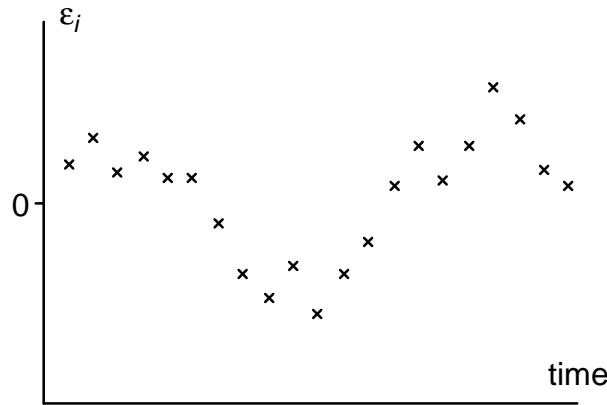
<sup>13</sup>Students who have taken 161.200 (or 161.230) will recall the negative binomial distribution; it will probably be new to other students. You will not be required to use the distribution here.

---

## Correlated Errors

---

Inference about linear models requires the assumption that the model errors,  $\varepsilon_i$ , are uncorrelated with each other. This is most often violated when the observations are made in time order; there may be some environmental influences on the response that are not explained in the model and that vary slowly over time, resulting in patterns where errors of similar magnitude follow each other.



When successive observations of the response are positively correlated, the relationship between the variables often appears stronger than it actually is, so confidence intervals are too narrow and the evidence for the slope being non-zero is exaggerated.

To detect this type of **serial correlation**, we therefore plot the residuals,  $e_i$ , against their time order and look for ‘unusual’ runs of positive or negative residuals<sup>14</sup>.

Checking the residuals graphically for serial correlation is often accompanied by a hypothesis test called the **Durbin-Watson test**. This is based on the test statistic,

$$d = \frac{\sum_{t=2}^n (e_t - e_{t-1})^2}{\sum_{t=1}^n e_t^2}$$

If successive residuals are positively correlated with each other, the test statistic will be low. The hypothesis of uncorrelated errors is therefore rejected for small values of this test statistic. A table of critical values for the test is provided at the end of this Chapter. For any sample size, number of explanatory variables and significance level, two values are presented in the table,  $d_L$  and  $d_U$ . If the test statistic  $d$  is below  $d_L$ , then the hypothesis of independence of errors is rejected — we conclude that the errors are autocorrelated. Values greater than  $d_U$  allow us to

---

<sup>14</sup>It is also possible for the residuals to be negatively correlated — most residuals being the opposite sign of the previous one — but this is far less common than positively correlated residuals. An example of negatively correlated errors might occur when modelling the annual price of a crop such as potatoes. High prices in one year might result in growers planting more in the following year and therefore unusually low prices; similarly, low prices in one year might be followed by growers planting less and therefore higher prices in the following year.

accept the hypothesis of independence, whereas values between  $d_L$  and  $d_U$  are interpreted as being 'inconclusive'.

Large values of the test statistic indicate negative serial correlation, but percentage points are not available.

A detailed example of the Durbin-Watson test follows; remedies for serial correlation will be considered later in the course.

The computer can provide you with the value of the Durbin-Watson test statistic see the Appendix for details.

The output then includes the value of the test statistic (but not its p-value)

Durbin-Watson statistic = 1.91

As there are  $n = 40$  observations and  $k = 1$  explanatory variables, for a 5% test we obtain  $d_L = 1.44$  and  $d_U = 1.54$ . Since the test statistic is greater than  $d_U$ , we conclude that there is no evidence of autocorrelation.

---

## Non-normal Errors

---

The errors  $\varepsilon_i$  in a regression model should be normally distributed for the standard confidence intervals and hypothesis tests about regression coefficients and predictions to be accurate.

Unless the errors are badly non-normal, violation of this assumption does not usually have much effect on the conclusions.

Since the studentised residuals,

$$r_i = \frac{e_i}{s\sqrt{1 - h_{ii}}}$$

are all approximately<sup>15</sup> normally distributed with mean 0 and variance 1, we can examine the distribution of the  $r_i$  to assess normality.

Various hypothesis tests have been proposed to test for normality. In particular, the  $\chi^2$  goodness-of-fit test can be used<sup>16</sup>. However we will concentrate on graphical displays that highlight non-normality.

A histogram of the  $r_i$  could be examined for symmetry. However a better assessment is based on a **normal probability plot** (also called a **quantile-quantile plot** or a **q-q plot**).

---

<sup>15</sup>They would be exactly normal (0, 1) if  $\sigma$  was known. When  $\sigma$  is replaced by an estimate, the distribution is still approximately normal.

<sup>16</sup>Group the studentised residuals into classes (e.g. -8 to -1.5, -1.5 to -1.0, etc.). Use normal tables to evaluate the expected number of residuals in each class, then group classes if necessary to get expected counts of at least 5 in each class. Evaluate the statistic  $\chi^2 = \sum \frac{(o_i - e_i)^2}{e_i}$  and compare with the  $\chi^2$  distribution with  $(g - 2)$  degrees of freedom, where  $g$  is the number of classes.



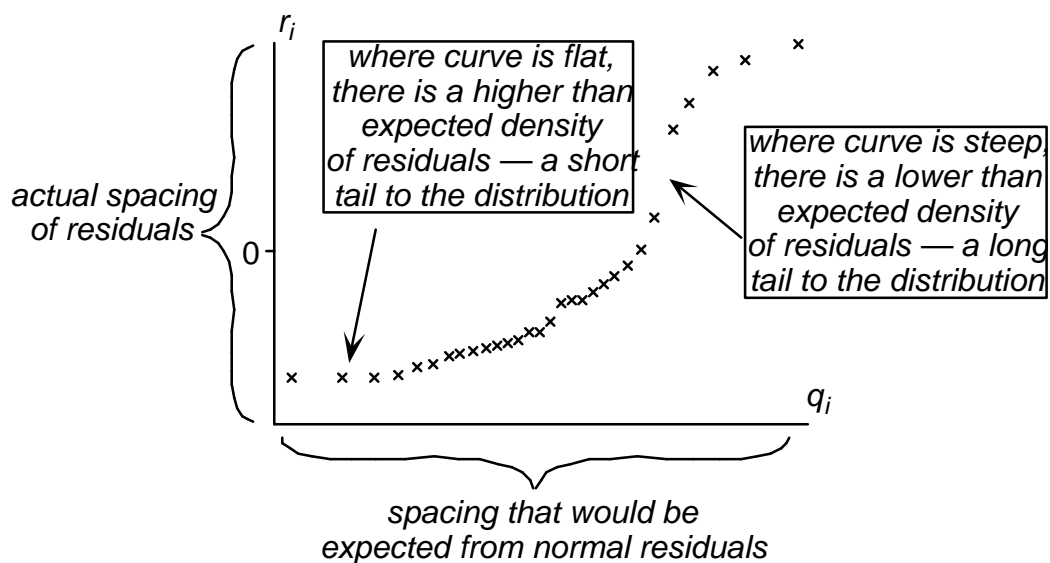
For a probability plot, we find the expected value for the lowest residual,  $q_1$ , the expected value for the second smallest residual,  $q_2$ , ..., and the expected value of the largest residual,  $q_n$ , assuming that they come from a normal distribution. These values are often called **normal scores**. Tables of these values are available for use when  $n$  is small, but when  $n$  is 15 or more, it is good enough to use the approximation,

$$q_i = z\left(\frac{i - 0.375}{n + 0.25}\right)$$

where the function  $z(\cdot)$  refers to looking up normal tables.

Minitab and SAS contain functions that can be used to plot normal probability plots and perform tests of normality. Details in the Appendix on page \*\*\*

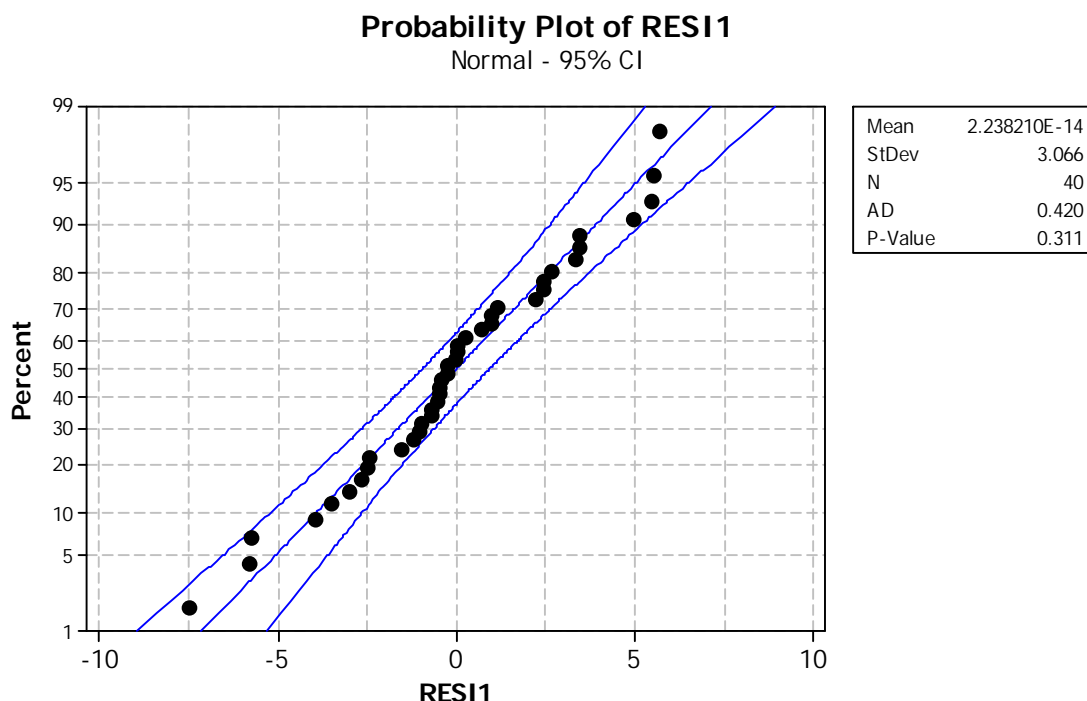
The points on the probability plot should lie near a straight line. If the probability plot is noticeably curved, this indicates that the distribution is not normal.



Sample sizes of  $n = 30$  or more are usually needed to detect non-normality since sample-to-sample variability in the shape of probability plots can be large with small sample sizes.

In Minitab the resulting probability plot unfortunately draws the residuals on the **horizontal** axis and the quantiles on the **vertical** axis — the **opposite way from our recommendation above**. This should be borne in mind when interpreting the shape of the plot. Also, the labels on the axis for the quantiles refer to the probabilities,  $\frac{i - 0.375}{n + 0.25}$ , rather than the actual normal scores,

$$q_i = z\left(\frac{i - 0.375}{n + 0.25}\right) \text{ themselves.}$$



Despite the disadvantage of Minitab’s probability plot having its axes swapped, it does have an important advantage — it automatically provides a p-value for a test of whether the nonlinearity in the probability plot is significant. In the above plot, the p-value is “> 0.1000”, so we would conclude that there is no evidence in the data against the assumption of normality. The SAS output is voluminous as usual but the same Anderson-Darling test is included (highlighted below).

The SAS System

13: 51 Thursday, March 4, 2004 19

The UNIVARIATE Procedure  
Variable: resid (Residual)

#### Moments

N	40	Sum Weights	40
Mean	0	Sum Observations	0
Std Deviation	3.0659692	Variance	9.40016715
Skewness	-0.1682736	Kurtosis	0.13286408
Uncorrected SS	366.606519	Corrected SS	366.606519
Coeff Variation	.	Std Error Mean	0.4847723

#### Basic Statistical Measures

Location		Variability	
Mean	0.00000	Std Deviation	3.06597

Median	-0.26590	Variance	9.40017
Mode	-0.68646	Range	13.17646
		Interquartile Range	3.71762

NOTE: The mode displayed is the smallest of 5 modes with a count of 2.

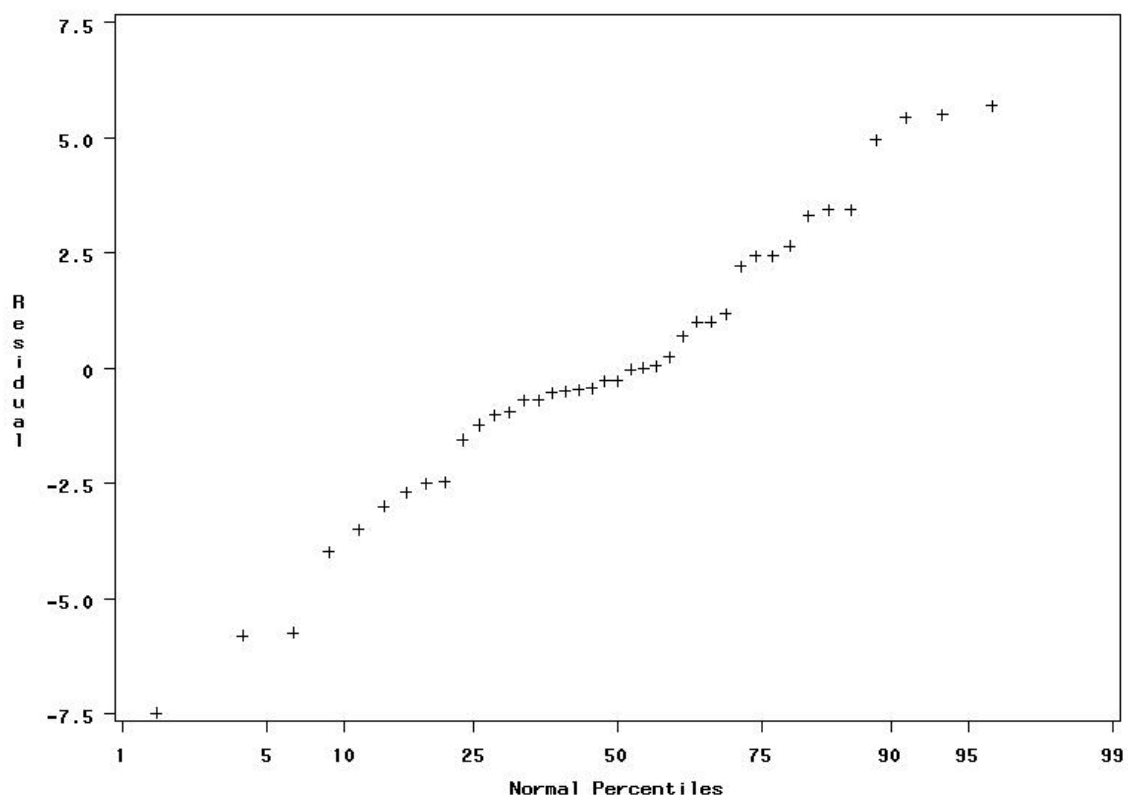
#### Tests for Location: $\mu_0=0$

Test	-Statistic-	-----p Value-----
Student's t	t      0	Pr >  t       1.0000
Sign	M      -2	Pr >=  M     0.6358
Signed Rank	S      -16	Pr >=  S     0.8328

#### Tests for Normality

Test	--Statistic--	-----p Value-----
Shapiro-Wilk	W      0.972417	Pr < W      0.4278
Kolmogorov-Smirnov	D      0.094569	Pr > D      >0.1500
Cramer-von Mises	W-Sq   0.075297	Pr > W-Sq   0.2368
Anderson-Darling	A-Sq   0.419705	Pr > A-Sq   >0.2500

The SAS graph follows the normal convention of plotting the data on the y axis.



---

## Leverage and Influence

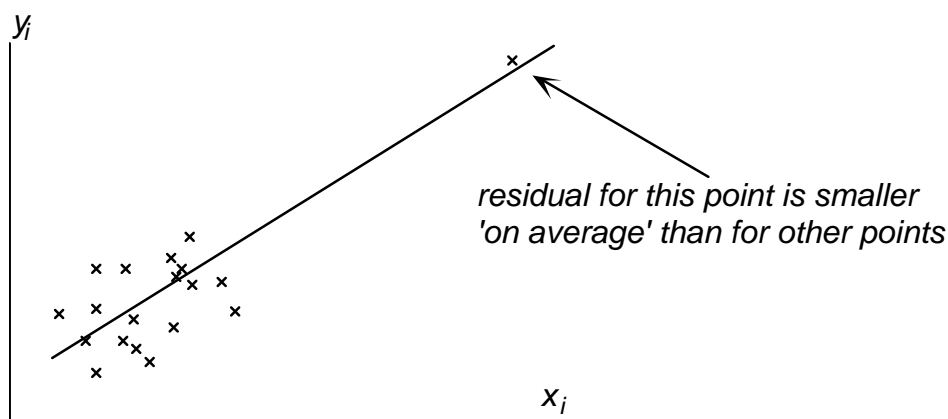
---

As explained earlier, the residual corresponding to an  $x$ -value far from  $\bar{x}$  has a smaller variance than a residual corresponding to an  $x$ -value near  $\bar{x}$ .

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

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

This means that  $x$ -values far from  $\bar{x}$  pull the regression line closer to the corresponding  $y$ -values — they have higher **leverage**.



Leverage is often measured by

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

Leverage always lies between zero and one and the average leverage over all  $n$  data points is  $2/n$  in a linear model with a single explanatory variable.

- $0 = h_{ii} = 1$
- $\frac{\sum h_{ii}}{n} = \frac{2}{n}$  when fitting a simple linear model (with 1 explanatory variable)

The higher the leverage of an  $x$ -value, the greater its potential influence on the regression results. Values of  $h_{ii}$  greater than  $4/n$  are generally regarded as having high leverage.

Notice that the leverage depends only on the  $x$ -values — it is not affected by the recorded  $y$ -values. Leverage measures the **potential** for a data point to affect the regression results, not its actual influence on these results.

High leverage points should always be carefully examined. Minitab automatically flags any observation that it regards as having high leverage when a regression model is fitted, in the format shown below.

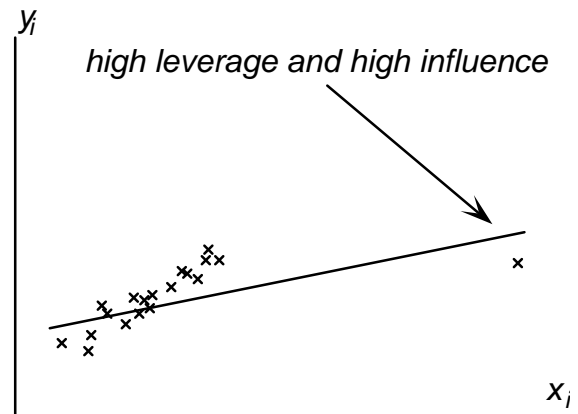
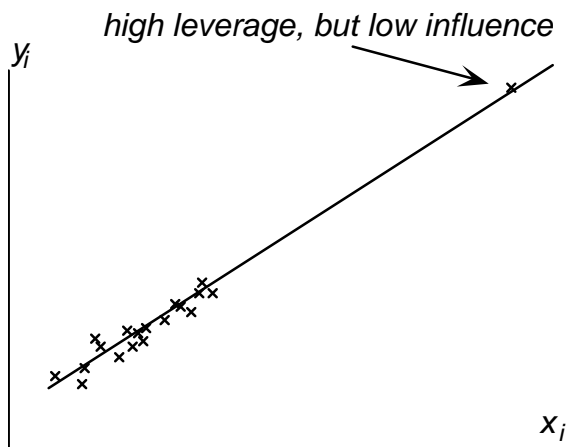
Unusual Observations						
Obs.	x	y	Fit	Stdev.Fit	Residual	St.Resid
1	0.770	4.5400	4.5062	0.1813	0.0338	0.11 X
13	0.770	4.2800	4.5062	0.1813	-0.2262	-0.73 X

X denotes an obs. whose X value gives it large influence.

It should be noted that Minitab uses the word ‘influence’ when ‘leverage’ should really be used. We have already seen how SAS can also produce these results.

In a simple linear model, a scatterplot of  $y$  against  $x$  will show how far the high leverage point’s  $x$ -value is from the bulk of the data, and a plot of the leverages (such as a time series plot) will allow assessment of how much higher the observation’s leverage is than that of the other observations in the data set.

High leverage on its own does not indicate that the corresponding observation should be omitted from a regression analysis. A high leverage point has the **potential** to badly affect the results of an analysis, but does it?



The actual influence of a data point depends not only on its leverage, but also on how closely it follows the pattern of the rest of the data — measured by its residual. We end this chapter with an examination of the influence of each data point on the regression results.

Firstly we examine the data point’s influence on the fitted value at  $x_i$  by considering the difference between the prediction at  $x_i$  from the whole data set and the data set without the  $i$ ’th observation,  $(\hat{y}_i - \hat{y}_{i,-i})$ . Since

$$\text{Var}(\hat{y}_i) = h_{ii} \sigma^2$$

we can use the ‘external’ estimate  $s_{-i}$  of  $\sigma$  that was defined earlier, to obtain the influence measure,

$$(DFITS)_i = \frac{\hat{y}_i - \hat{y}_{i,-i}}{s_{-i} \sqrt{h_{ii}}}$$

Since it can also be shown that

$$(DFITS)_i = t_i \sqrt{\frac{h_{ii}}{1 - h_{ii}}}$$

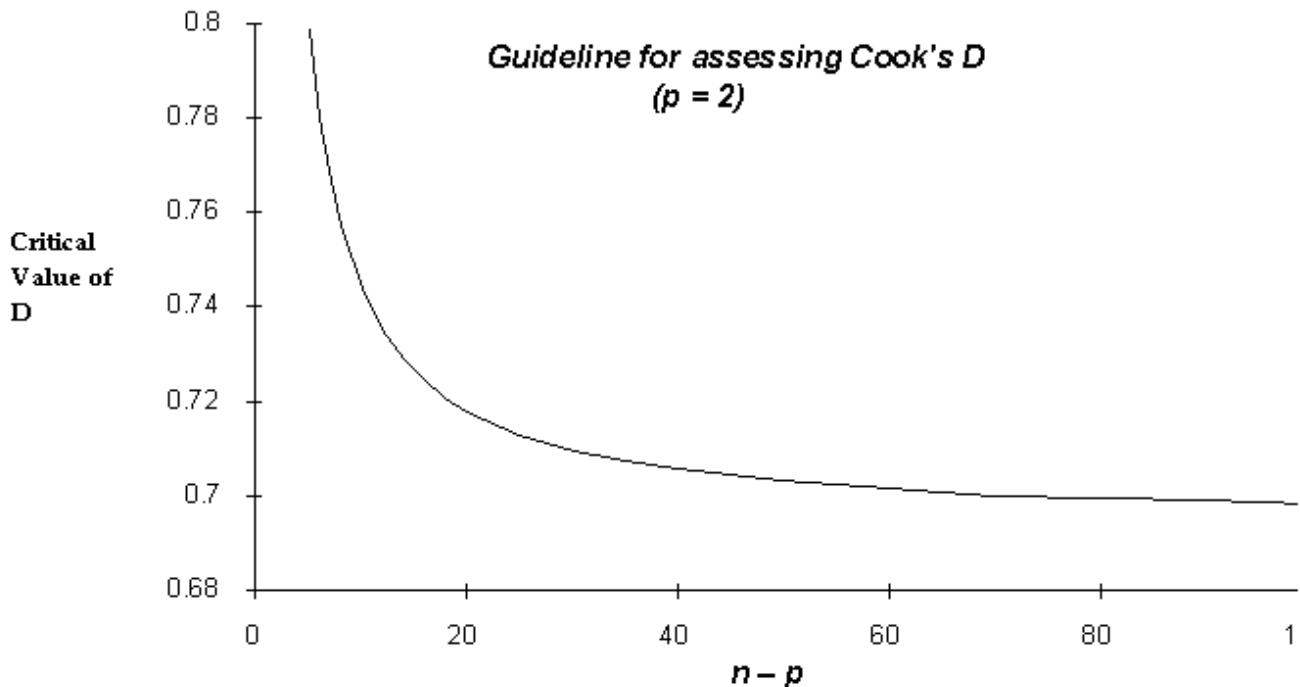
where  $t_i$  is the  $i$ 'th externally studentised deleted residual, it can be seen that  $(DFITS)_i$  depends both on the size of the residual, and also on the leverage of the point. Note that standardising  $(DFITS)_i$  (in order to compare it against  $\pm 2$  and  $\pm 3$ ) just gives the externally studentised deleted residual,  $t_i$ . Observations with  $(DFITS)_i$  greater than  $2\sqrt{\frac{2}{n-2}}$  are often classified as 'influential'.

Plots of  $(DFITS)_i$  against either fitted values, the explanatory variable or in time order are often useful to assess the degree of influence.

A final measure of the influence of the  $i$ 'th observation on the regression coefficients is 'Cook's distance'. A detailed explanation of the statistic cannot be given at this stage, but it is a commonly used combined measure of the difference between the regression coefficients with and without the  $i$ 'th data value. It can also be written in the form

$$D_i = \left( \frac{r_i^2}{p} \right) \left( \frac{h_{ii}}{1 - h_{ii}} \right)$$

where  $r_i$  is the  $i$ 'th studentised residual, and  $p$  is the number of regression coefficients,  $p = 2$  for the simple linear model. Again, this depends only on the  $i$ 'th residual and the  $i$ 'th leverage. Assessing the value of  $D_i$  is more difficult than the earlier diagnostics. A yard-stick that has been suggested is to compare  $D_i$  with the median of the F-distribution with  $p$  and  $n - p$  degrees of freedom; values larger than this may be regarded as having unusually high influence.



As a rule of thumb, for  $n > 15$ , we can treat points with  $D_i > 0.7$  as being influential. However, rather than formally applying this rule, you are encouraged to examine plots of  $D_i$  against an explanatory variable, fitted values or time order for a visual identification of any observation with markedly higher influence than the other observations in the data.

Both computer packages can calculate these statistics and plots of them can be drawn as before.

---

**Table – Distribution of Durbin-Watson  $d$ :**


---

**5% Significance Points of  $d_L$  and  $d_U$** 

$n$	$k = 1$		$k = 2$		$k = 3$		$k = 4$		$k = 5$	
	$d_L$	$d_U$	$d_L$	$d_U$	$d_L$	$d_U$	$d_L$	$d_U$	$d_L$	$d_U$
15	1.08	1.36	0.95	1.54	0.82	1.75	0.69	1.97	0.56	2.21
16	1.10	1.37	0.98	1.54	0.86	1.73	0.74	1.93	0.62	2.15
17	1.13	1.38	1.02	1.54	0.90	1.71	0.78	1.90	0.67	2.10
18	1.16	1.39	1.05	1.53	0.93	1.69	0.82	1.87	0.71	2.06
19	1.18	1.40	1.08	1.53	0.97	1.68	0.86	1.85	0.75	2.02
20	1.20	1.41	1.10	1.54	1.00	1.68	0.90	1.83	0.79	1.99
21	1.22	1.42	1.13	1.54	1.03	1.67	0.93	1.81	0.83	1.96
22	1.24	1.43	1.15	1.54	1.05	1.66	0.96	1.80	0.86	1.94
23	1.26	1.44	1.17	1.54	1.08	1.66	0.99	1.79	0.90	1.92
24	1.27	1.45	1.19	1.55	1.10	1.66	1.01	1.78	0.93	1.90
25	1.29	1.45	1.21	1.55	1.12	1.66	1.04	1.77	0.95	1.89
26	1.30	1.46	1.22	1.55	1.14	1.65	1.06	1.76	0.98	1.88
27	1.32	1.47	1.24	1.56	1.16	1.65	1.08	1.76	1.01	1.86
28	1.33	1.48	1.26	1.56	1.18	1.65	1.10	1.75	1.03	1.85
29	1.34	1.48	1.27	1.56	1.20	1.65	1.12	1.74	1.05	1.84
30	1.35	1.49	1.28	1.57	1.21	1.65	1.14	1.74	1.07	1.83
31	1.36	1.50	1.30	1.57	1.23	1.65	1.16	1.74	1.09	1.83
32	1.37	1.50	1.31	1.57	1.24	1.65	1.18	1.73	1.11	1.82
33	1.38	1.51	1.32	1.58	1.26	1.65	1.19	1.73	1.13	1.81
34	1.39	1.51	1.33	1.58	1.27	1.65	1.21	1.73	1.15	1.81
35	1.40	1.52	1.34	1.58	1.28	1.65	1.22	1.73	1.16	1.80
36	1.41	1.52	1.35	1.59	1.29	1.65	1.24	1.73	1.18	1.80
37	1.42	1.53	1.36	1.59	1.31	1.66	1.25	1.72	1.19	1.80
38	1.43	1.54	1.37	1.59	1.32	1.66	1.26	1.72	1.21	1.79
39	1.43	1.54	1.38	1.60	1.33	1.66	1.27	1.72	1.22	1.79
40	1.44	1.54	1.39	1.60	1.34	1.66	1.29	1.72	1.23	1.79
45	1.48	1.57	1.43	1.62	1.38	1.67	1.34	1.72	1.29	1.78
50	1.50	1.59	1.46	1.63	1.42	1.67	1.38	1.72	1.34	1.77
55	1.53	1.60	1.49	1.64	1.45	1.68	1.41	1.72	1.38	1.77
60	1.55	1.62	1.51	1.65	1.48	1.69	1.44	1.73	1.41	1.77
65	1.57	1.63	1.54	1.66	1.50	1.70	1.47	1.73	1.44	1.77
70	1.58	1.64	1.55	1.67	1.52	1.70	1.49	1.73	1.46	1.77
75	1.60	1.65	1.57	1.68	1.54	1.71	1.51	1.74	1.49	1.77
80	1.61	1.66	1.59	1.69	1.56	1.72	1.53	1.74	1.51	1.77
85	1.62	1.67	1.60	1.70	1.57	1.72	1.55	1.75	1.52	1.77
90	1.63	1.68	1.61	1.70	1.59	1.73	1.57	1.75	1.54	1.78
95	1.64	1.69	1.62	1.71	1.60	1.73	1.58	1.75	1.56	1.78
100	1.65	1.69	1.63	1.72	1.61	1.74	1.59	1.76	1.57	1.78

Source: J. Durbin and G.S. Watson, *Biometrika*, **38** (1951)



**Table – Distribution of Durbin–Watson  $d$ : (Continued)**  
**1% Significance Points of  $d_L$  and  $d_U$**

	$k = 1$		$k = 2$		$k = 3$		$k = 4$		$k = 5$	
$n$	$d_L$	$d_U$	$d_L$	$d_U$	$d_L$	$d_U$	$d_L$	$d_U$	$d_L$	$d_U$
15	0.81	1.07	0.70	1.25	0.59	1.46	0.46	1.70	0.39	1.96
16	0.84	1.09	0.74	1.25	0.63	1.44	0.53	1.66	0.44	1.90
17	0.87	1.10	0.77	1.25	0.67	1.43	0.57	1.63	0.48	1.85
18	0.90	1.12	0.80	1.26	0.71	1.42	0.61	1.60	0.52	1.80
19	0.93	1.13	0.83	1.26	0.74	1.41	0.65	1.58	0.56	1.77
20	0.95	1.15	0.86	1.27	0.77	1.41	0.68	1.57	0.60	1.74
21	0.97	1.16	0.89	1.27	0.80	1.41	0.72	1.55	0.63	1.71
22	1.00	1.17	0.91	1.28	0.83	1.40	0.75	1.54	0.66	1.69
23	1.02	1.19	0.94	1.29	0.86	1.40	0.77	1.53	0.70	1.67
24	1.04	1.20	0.96	1.30	0.88	1.41	0.80	1.53	0.72	1.66
25	1.05	1.21	0.98	1.30	0.90	1.41	0.83	1.52	0.75	1.65
26	1.07	1.22	1.00	1.31	0.93	1.41	0.85	1.52	0.78	1.64
27	1.09	1.23	1.02	1.32	0.95	1.41	0.88	1.51	0.81	1.63
28	1.10	1.24	1.04	1.32	0.97	1.41	0.90	1.51	0.83	1.62
29	1.12	1.25	1.05	1.33	0.99	1.42	0.92	1.51	0.85	1.61
30	1.13	1.26	1.07	1.34	1.01	1.42	0.94	1.51	0.88	1.61
31	1.15	1.27	1.08	1.34	1.02	1.42	0.96	1.51	0.90	1.60
32	1.16	1.28	1.10	1.35	1.04	1.43	0.98	1.51	0.92	1.60
33	1.17	1.29	1.11	1.36	1.05	1.43	1.00	1.51	0.94	1.59
34	1.18	1.30	1.13	1.36	1.07	1.43	1.01	1.51	0.95	1.59
35	1.19	1.31	1.14	1.37	1.08	1.44	1.03	1.51	0.97	1.59
36	1.21	1.32	1.15	1.38	1.10	1.44	1.04	1.51	0.99	1.59
37	1.22	1.32	1.16	1.38	1.11	1.45	1.06	1.51	1.00	1.59
38	1.23	1.33	1.18	1.39	1.12	1.45	1.07	1.52	1.02	1.58
39	1.24	1.34	1.19	1.39	1.14	1.45	1.09	1.52	1.03	1.58
40	1.25	1.34	1.20	1.40	1.15	1.46	1.10	1.52	1.05	1.58
45	1.29	1.38	1.24	1.42	1.20	1.48	1.16	1.53	1.11	1.58
50	1.32	1.40	1.28	1.45	1.24	1.49	1.20	1.54	1.16	1.59
55	1.36	1.43	1.32	1.47	1.28	1.51	1.25	1.55	1.21	1.59
60	1.38	1.45	1.35	1.48	1.32	1.52	1.28	1.56	1.25	1.60
65	1.41	1.47	1.38	1.50	1.35	1.53	1.31	1.57	1.28	1.61
70	1.43	1.49	1.40	1.52	1.37	1.55	1.34	1.58	1.31	1.61
75	1.45	1.50	1.42	1.53	1.39	1.56	1.37	1.59	1.34	1.62
80	1.47	1.52	1.44	1.54	1.42	1.57	1.39	1.60	1.36	1.62
85	1.48	1.53	1.46	1.55	1.43	1.58	1.41	1.60	1.39	1.63
90	1.50	1.54	1.47	1.56	1.45	1.59	1.43	1.61	1.41	1.64
95	1.51	1.55	1.49	1.57	1.47	1.60	1.45	1.62	1.42	1.64
100	1.52	1.56	1.50	1.58	1.48	1.60	1.46	1.63	1.44	1.65

Source: J. Durbin and G.S. Watson, *Biometrika*, **38** (1951)

# ***161.320 Fitting Regression Models***

## ***3. The General Linear Model***

---

### **Models with more than One Explanatory Variable**

---

In Chapter 3, we will extend the methods and results from Chapters 1 and 2 from the simple linear model (with a single explanatory variable) to linear models with an arbitrary number of explanatory variables.

For example, information was recorded on each of several consecutive months from a steam plant at a large industrial concern. The variables recorded are:

<i>steam:</i>	pounds of steam used monthly
<i>storage</i>	pounds of real fatty acid in storage per month
<i>glycerin</i>	pounds of crude glycerin made
<i>wind</i>	average wind velocity (in mph)
<i>calDays</i>	calendar days per month
<i>opDays</i>	operating days per month
<i>coldDays</i>	days below 32°F
<i>temp</i>	average atmospheric temperature (°F)
<i>startups</i>	number of startups

and the data are:

<b>steam</b>	<b>storage</b>	<b>glycerin</b>	<b>wind</b>	<b>calDays</b>	<b>opDays</b>	<b>coldDays</b>	<b>temp</b>	<b>startups</b>
10.98	5.20	.61	7.4	31	20	22	35.3	4
11.13	5.12	.64	8.0	29	20	25	29.7	5
12.51	6.19	.78	7.4	31	23	17	30.8	4
8.40	3.89	.49	7.5	30	20	22	58.8	4
9.27	6.28	.84	5.5	31	21	0	61.4	5
8.73	5.76	.74	8.9	30	22	0	71.3	4
6.36	3.45	.42	4.1	31	11	0	74.4	2
8.50	6.57	.87	4.1	31	23	0	76.7	5
7.82	5.69	.75	4.1	30	21	0	70.7	4
9.14	6.14	.76	4.5	31	20	0	57.5	5
8.24	4.84	.65	10.3	30	20	11	46.4	4
12.19	4.88	.62	6.9	31	21	12	28.9	4
11.88	6.03	.79	6.6	31	21	25	28.1	5
9.57	4.55	.60	7.3	28	19	18	39.1	5
10.94	5.71	.70	8.1	31	23	5	46.8	4
9.58	5.67	.74	8.4	30	20	7	48.5	4
10.09	6.72	.85	6.1	31	22	0	59.3	6
8.11	4.95	.67	4.9	30	22	0	70.0	4
6.83	4.62	.45	4.6	31	11	0	70.0	3
8.88	6.60	.95	3.7	31	23	0	74.5	4
7.68	5.01	.64	4.7	30	20	0	72.1	4
8.47	5.68	.75	5.3	31	21	1	58.1	6
8.86	5.28	.70	6.2	30	20	14	44.6	4
10.36	5.36	.67	6.8	31	20	22	33.4	4
11.08	5.87	.70	7.5	31	22	28	28.6	5

In this example, we are interested in modelling how the quantity of steam used (the response variable) depends on the eight explanatory variables. This chapter will describe models for data sets of this form and will describe how to fit such models. It should be noted that trying to fit eight explanatory to a model with only 25 data points would not be advisable. As a rule of thumb eight data points per explanatory variable would be a minimum.

---

## Matrix Notation for the Simple Linear Model

---

The key to describing linear models concisely and writing results and formulae relating to them is matrix notation. In the first section, we show how the simple linear model can be expressed concisely in matrix notation.<sup>17</sup> This notation can be extended in a fairly straightforward way to models with more explanatory variables.

We have written the simple linear model in the form,

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

In matrix notation, this can be written as

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{e}$$

where

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} b_0 \\ b_1 \end{bmatrix}, \text{ and } \mathbf{e} = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}$$

If the vector  $\mathbf{b}$  is estimated by  $\hat{\mathbf{b}}$ , the model's fitted values and residuals are then

$$\hat{\mathbf{y}} = \begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_n \end{bmatrix} \text{ and } \mathbf{e} = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix} = \begin{bmatrix} y_1 - \hat{y}_1 \\ y_2 - \hat{y}_2 \\ \vdots \\ y_n - \hat{y}_n \end{bmatrix} = \mathbf{y} - \mathbf{X}\hat{\mathbf{b}}$$

The method of least squares chooses the parameters  $\mathbf{b}$  to minimise the residual sum of squares,

$$SS_{\text{Resid}} = \sum_{i=1}^n e_i^2 = \mathbf{e}^T \mathbf{e} = (\mathbf{y} - \mathbf{X}\mathbf{b})^T (\mathbf{y} - \mathbf{X}\mathbf{b})$$

---

<sup>17</sup>We will only assume knowledge of matrix addition, subtraction and multiplication. Matrix inverses are also used in some of the formulae that we will present, but understanding of them is less important.

---

## The General Linear Model

---

In this section, we extend the simple model to allow more than one explanatory variable.

$$\begin{aligned} y_1 &= \beta_0 + \beta_1 x_1 + \beta_2 z_1 + \dots + \varepsilon_1 \\ y_2 &= \beta_0 + \beta_1 x_2 + \beta_2 z_2 + \dots + \varepsilon_2 \\ &\vdots \\ y_n &= \beta_0 + \beta_1 x_n + \beta_2 z_n + \dots + \varepsilon_n \end{aligned}$$

An alternative notation for the explanatory variables uses two subscripts — the first subscript referring to the ‘variable’ and the second referring to the case.

$$\begin{aligned} y_1 &= \beta_0 + \beta_1 x_{11} + \beta_2 x_{21} + \dots + \beta_p x_{p1} + \varepsilon_1 \\ y_2 &= \beta_0 + \beta_1 x_{12} + \beta_2 x_{22} + \dots + \beta_p x_{p2} + \varepsilon_2 \\ &\vdots \\ y_n &= \beta_0 + \beta_1 x_{1n} + \beta_2 x_{2n} + \dots + \beta_p x_{pn} + \varepsilon_n \end{aligned}$$

In matrix notation, the model can again be written as

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{e}$$

where

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad \mathbf{X} = \begin{array}{c} \overbrace{\begin{bmatrix} 1 & x_1 & z_1 & \dots \\ 1 & x_2 & z_2 & \dots \\ \vdots & \vdots & \vdots & \ddots \\ 1 & x_n & z_n & \dots \end{bmatrix}}^{p \text{ columns}}, \quad \mathbf{b} = \begin{bmatrix} b_0 \\ b_1 \\ b_2 \\ \vdots \end{bmatrix} \left\} \begin{array}{l} p \text{ parameters, and} \end{array} \right. \quad \mathbf{e} = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}$$

We will also assume that the errors  $\varepsilon_i$  are independent of each other and are normally distributed with mean 0 and variance  $\sigma^2$ .

If we estimate  $\mathbf{b}$  with  $\hat{\mathbf{b}}$ , the model’s fitted values and residuals can still be written as

$$\hat{\mathbf{y}} = \begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_n \end{bmatrix} = \mathbf{X}\hat{\mathbf{b}} \quad \text{and} \quad \mathbf{e} = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix} = \begin{bmatrix} y_1 - \hat{y}_1 \\ y_2 - \hat{y}_2 \\ \vdots \\ y_n - \hat{y}_n \end{bmatrix} = \mathbf{y} - \mathbf{X}\hat{\mathbf{b}}$$

The method of least squares chooses the parameters  $\hat{\mathbf{b}}$  to minimise the residual sum of squares,

$$SS_{\text{Resid}} = \sum_{i=1}^n e_i^2 = \mathbf{e}^T \mathbf{e} = (\mathbf{y} - \mathbf{X}\mathbf{b})^T (\mathbf{y} - \mathbf{X}\mathbf{b})$$

Matrix theory can be used to show that the solution to this minimisation problem is

$$\mathbf{b} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$$

and this vector of parameter estimates is called the **least squares** estimate of  $\beta$ . Note that matrix representation of the problem allows us to use a single simple formula for the least squares estimates of **all** parameters at once — we don't need separate formulae for the different parameters!

It can also be shown that each element of  $\mathbf{b}$  is an unbiased estimate of the corresponding element of  $\beta$ ,

$$E[\mathbf{b}] = \mathbf{b}$$

Theory can also show that the variances of the least squares estimates are given by the diagonal elements<sup>18</sup> of the matrix,  $(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{s}^2$ ,

$$(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{s}^2 = \begin{bmatrix} \text{Var}(b_0) & ? & \cdots & ? \\ ? & \text{Var}(b_1) & \cdots & ? \\ \vdots & \vdots & \ddots & \vdots \\ ? & ? & \cdots & \text{Var}(b_{p-1}) \end{bmatrix}$$

Finally, the residual sum of squares,  $SS_{\text{Resid}}$  has  $(n - p)$  degrees of freedom and the mean residual sum of squares,

$$\hat{s}^2 = MSS_{\text{Resid}} = \frac{SS_{\text{Resid}}}{n - p} = \frac{\mathbf{e}^T \mathbf{e}}{n - p} = \frac{(\mathbf{y} - \mathbf{X}\mathbf{b})^T (\mathbf{y} - \mathbf{X}\mathbf{b})}{n - p}$$

is the best estimate of the error variance,  $\sigma^2$ .

---

## Illustrations of the General Theory

---

To illustrate these results, we will first apply them to two extremely simple models.

### Model with no explanatory variables

Firstly consider the model,

---

<sup>18</sup>The off-diagonal elements of the matrix are the covariances between the parameter estimates.

$$y_i = \beta_0 + \varepsilon_i \quad \text{for } i = 1, \dots, n$$

where the  $\varepsilon_i$  are independent normal random variables,  $\varepsilon_i \sim \text{normal}(0, \sigma^2)$ . This model simply states that each  $y_i$  is normal with mean  $\beta_0$  and standard deviation  $\sigma$ . This model can be written as a general linear model with

$$\mathbf{X} = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}, \text{ and } \mathbf{b} = [b_0],$$

Our general theory therefore gives the least squares parameter estimate

$$\begin{aligned} \mathbf{b} &= (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y} \\ &= (n)^{-1} \left( \sum y_i \right) \\ &= \bar{y} \end{aligned}$$

It also states that the variance of this estimate is

$$\text{Var}(b_0) = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{s}^2 = \frac{\mathbf{s}^2}{n}$$

and that the best estimate of  $\sigma^2$  is

$$\hat{\mathbf{s}}^2 = \frac{SS_{\text{Resid}}}{n-p} = \frac{\begin{bmatrix} y_1 - \bar{y} \\ y_2 - \bar{y} \\ \vdots \end{bmatrix}^T \begin{bmatrix} y_1 - \bar{y} \\ y_2 - \bar{y} \\ \vdots \end{bmatrix}}{n-1} = \frac{\sum (y_i - \bar{y})^2}{n-1}$$

These are the standard results for estimating the mean of a normal distribution.

### **Model with 1 explanatory variables and no intercept**

For a second simple illustration, consider the simple linear model with no intercept term,

$$y_i = \beta_1 x_i + \varepsilon_i \quad \text{for } i = 1, \dots, n$$

This model may be appropriate if the mean response is expected to be zero whenever the explanatory variable is zero. For example, in a chemical reaction,  $x$  may be the concentration of a catalyst and  $y$  may be the concentration of the compound being produced; with no catalyst, the reaction may not take place. For this model,

$$\mathbf{X} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}, \text{ and } \mathbf{b} = [b_1],$$

Our general theory therefore gives parameter estimate

$$\begin{aligned}\mathbf{b} &= (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y} \\ &= \left( \sum x_i^2 \right)^{-1} \left( \sum x_i y_i \right) \\ &= \frac{\sum x_i y_i}{\sum x_i^2}\end{aligned}$$

It also states that the variance of this estimate is

$$\text{Var}(b_1) = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{s}^2 = \frac{\mathbf{s}^2}{\sum x_i^2}$$

and that the best estimate of  $\sigma^2$  is

$$\hat{\mathbf{s}}^2 = \frac{\text{SS}_{\text{Resid}}}{n-p} = \frac{\begin{bmatrix} y_1 - b_1 x_1 \\ y_2 - b_1 x_2 \\ \vdots \end{bmatrix}^T \begin{bmatrix} y_1 - b_1 x_1 \\ y_2 - b_1 x_2 \\ \vdots \end{bmatrix}}{n-1} = \frac{\sum (y_i - b_1 x_i)^2}{n-1}$$

### Simple linear model (with intercept and single explanatory variable)

Showing that the matrix representation of the simple linear model results in the same parameter estimates that were presented earlier involves considerably more algebraic manipulation which we do not want to stress in this course. A brief sketch is given below for the more mathematically inclined student, but it is not an examinable part of the course.

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} b_0 \\ b_1 \end{bmatrix}, \text{ and } \mathbf{e} = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}$$

Matrix multiplication gives

$$\mathbf{X}^T \mathbf{X} = \begin{bmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{bmatrix}, \text{ and } \mathbf{X}^T \mathbf{y} = \begin{bmatrix} \sum y_i \\ \sum x_i y_i \end{bmatrix}$$

The inverse of the matrix  $\mathbf{X}^T \mathbf{X}$  is given by<sup>19</sup>

$$(\mathbf{X}^T \mathbf{X})^{-1} = \frac{1}{n \sum x_i^2 - (\sum x_i)^2} \begin{bmatrix} \sum x_i^2 & -\sum x_i \\ -\sum x_i & n \end{bmatrix} = \frac{1}{n \sum (x_i - \bar{x})^2} \begin{bmatrix} \sum x_i^2 & -\sum x_i \\ -\sum x_i & n \end{bmatrix}$$

---

<sup>19</sup>In general,  $\begin{bmatrix} a & b \\ c & d \end{bmatrix}^{-1} = \frac{1}{ad-bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$



The least squares estimates are therefore

$$\mathbf{b} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y} = \frac{1}{n \sum (x_i - \bar{x})^2} \begin{bmatrix} \sum x_i^2 \cdot \sum y_i - \sum x_i \cdot \sum x_i y_i \\ - \sum x_i \cdot \sum y_i + n \sum x_i y_i \end{bmatrix} = \begin{bmatrix} \bar{y} - \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} \bar{x} \\ \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} \end{bmatrix}$$

Finally, the variances of the least squares estimates are given by the diagonal elements of

$$(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{s}^2 = \frac{\mathbf{s}^2}{n \sum (x_i - \bar{x})^2} \begin{bmatrix} \sum x_i^2 & -\sum x_i \\ -\sum x_i & n \end{bmatrix}$$

and the best estimate of  $\sigma^2$  is given by the mean error sum of squares,

$$\hat{\mathbf{s}}^2 = \text{MSS}_{\text{Resid}} = \frac{\text{SS}_{\text{Resid}}}{n-2} = \frac{\sum (y_i - b_0 - b_1 x_i)^2}{n-2}$$

You should look back in your notes to verify that these results are indeed the same as those given earlier for the simple linear model.

## Inference about Single Parameters

From our general results, we can obtain confidence intervals for individual parameters in the general linear model. A confidence interval for  $\beta_i$  would be of the form

$$b_i \pm t_{n-p} \times \text{se}(b_i)$$

where  $t_{n-p}$  is the appropriate value from t-tables with  $(n - p)$  degrees of freedom. Using the results given earlier,

$$\text{se}(b_i) = \hat{\mathbf{s}} \times \sqrt{i\text{th diagonal element of } (\mathbf{X}^T \mathbf{X})^{-1}}$$

Similarly, a hypothesis test for whether  $\beta_i = k$  can be based on the test statistic,

$$t = \frac{b_i - k}{\text{se}(b_i)}$$

This would again be compared with the t-distribution with  $(n - p)$  degrees of freedom to assess significance.

## Fitting the Models

We will fit the following linear model to the *steam* data (see appendix page 254 for details of computing).

$$\begin{aligned} \text{steam}_i = & \beta_0 + \beta_1 \text{storage}_i + \beta_2 \text{glycerin}_i + \beta_3 \text{wind}_i + \beta_4 \text{calDays}_i + \beta_5 \text{opDays}_i + \\ & \beta_6 \text{coldDays}_i \\ & + \beta_7 \text{temp}_i + \\ & \beta_8 \text{startups}_i + \varepsilon_i \end{aligned}$$

The model is specified in a similar way to the simple linear regression models in Chapter 1.

The computer output is shown below.

### Regression Analysis: steam versus storage, glycerin, ...

The regression equation is

$$\begin{aligned} \text{steam} = & 6.29 + 0.936 \text{ storage} - 4.64 \text{ glycerin} - 0.087 \text{ wind} + 0.105 \text{ calDays} \\ & + 0.221 \text{ opDays} - 0.0181 \text{ coldDays} - 0.0874 \text{ temp} - 0.255 \text{ startups} \end{aligned}$$

Predictor	Coef	SE Coef	T	P
Constant	6.291	6.810	0.92	0.369
storage	0.9365	0.5760	1.63	0.124
glycerin	-4.639	3.998	-1.16	0.263
wind	-0.0868	0.1035	-0.84	0.414
calDays	0.1053	0.2152	0.49	0.631
opDays	0.22061	0.08104	2.72	0.015
coldDays	-0.01806	0.02580	-0.70	0.494
temp	-0.08737	0.01627	-5.37	0.000
startups	-0.2548	0.2142	-1.19	0.252

S = 0.599682    R-Sq = 91.0%    R-Sq(adj) = 86.5%

The least squares estimates of the parameters are given in the column 'Coef' with their standard deviations in the following column.

The t-ratios and associated p-values for testing whether the different parameters are zero are given in the final two columns. For example, the p-value for testing whether the coefficient of *glycerin* is zero (i.e. for testing whether *glycerin* affects *steam* in a model with all 8 explanatory variables) is 0.263, so we would conclude that there is no evidence that *glycerin* is necessary in this model. Conversely, the p-value for testing whether *opDays* is necessary is 0.015, so there is quite strong evidence that this variable should be retained in the model.

A 95% confidence interval for the coefficient of *storage* is

$$b_1 \pm t_{25-9} \times \text{se}(b_1) = 0.9365 \pm 2.120 \times 0.5760$$

---

## Predictions

---

Predictions for the general linear model are similar to those for the simple linear model. At explanatory variable values  $x_{0,1}, x_{0,2}, \dots, x_{0,p-1}$ , the predicted value for  $y$  would be

$$\hat{y} = b_0 + b_1 x_{0,1} + b_2 x_{0,2} + \dots + b_{p-1} x_{0,p-1}$$

This is both an estimate of the mean response level when the explanatory variables take these values and is also a prediction of a new single observation at these values. In matrix notation, this can be written as

$$\hat{y} = \mathbf{x}_0^T \mathbf{b}$$

where

$$\mathbf{x}_0 = \begin{bmatrix} 1 \\ x_{0,1} \\ x_{0,2} \\ \vdots \\ x_{0,p-1} \end{bmatrix} \quad \text{and} \quad \mathbf{b} = \begin{bmatrix} b_0 \\ b_1 \\ b_2 \\ \vdots \\ b_{p-1} \end{bmatrix}$$

Although we give no proofs of results involving matrices for the general linear model, we state here the formula for the variance of a prediction of this form.

$$\text{Var}(\hat{y}) = \mathbf{x}_0^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_0 \mathbf{s}^2$$

We can therefore find a confidence interval for the **mean** response at  $\mathbf{x}_0$ , which is of the form

$$\hat{y} \pm t \times \hat{\mathbf{S}} \times \sqrt{\mathbf{x}_0^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_0}$$

where  $t$  is the appropriate value from t-tables with  $n - p$  degrees of freedom and  $\hat{\mathbf{S}}$  is the mean residual sum of squares.

As in the simple linear model, the errors in prediction of an individual's  $y$ -value are greater than the errors in predicting the mean response,

$$\text{Var}(y_0 - \hat{y}) = \left(1 + \mathbf{x}_0^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_0\right) \mathbf{s}^2$$

A prediction interval for a single new response value at  $\mathbf{x}_0$  is wider, and is of the form

$$\hat{y} \pm t \times \hat{\mathbf{S}} \times \sqrt{1 + \mathbf{x}_0^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_0}$$

In Minitab, predictions are obtained by clicking the Options button in the regression dialog box and specifying the values for all the explanatory variables

for which a prediction is required. The output is similar to that for the simple linear model.

Predicted Values for New Observations								
New								
Obs	Fit	SE Fit	95% CI	95% PI				
1	8.712	0.265	(8.150, 9.273)	(7.322, 10.101)				
Values of Predictors for New Observations								
New								
Obs	storage	glycerin	wind	calDays	opDays	coldDays	temp	startups
1	5.00	0.600	7.00	30.0	20.0	10.0	60.0	4.00

The 95% confidence interval for the mean response at these x-values is therefore (8.150 to 9.273), and a 95% prediction interval for a single new response is the wider interval (7.322 to 10.102).

---

## Regression Diagnostics

---

The multiple linear regression model makes various assumptions about the data that should be checked. Conclusions drawn from simply fitting a multiple linear regression model may be invalid if these assumptions do not hold. The assumptions and the methods we use to detect violations are similar to those for the simple linear model.

Because the diagnostic plots and statistics are generalisations of those described in Chapter 2, we only briefly describe how they are generalised here. The worked examples following Chapter 4 illustrate their use.

### ***Nonlinearity***

For the multiple linear model, plots of  $y_i$  against each of the explanatory variables in turn may show up curvature and should always precede any analysis.

However problems with the model are usually more obvious from plots of the residuals,  $e_i$ , against each of the explanatory variables in turn. The plots are interpreted in a similar way to the corresponding plots for the simple linear model.

### ***Outliers***

As in the simple linear model, the residuals,

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

are estimates of the errors  $\varepsilon_i$ , and are the basis of a quest for outliers. The residuals do not have constant variance; their variance is

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

where  $h_{ii}$  is the  $i$ 'th diagonal element of the  $n \times n$  matrix  $\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ . Luckily the computer will find the  $h_{ii}$  for us!

We can therefore find the **studentised residuals**,

$$r_i = \frac{e_i}{\hat{S}\sqrt{1-h_{ii}}}$$

which can be assessed in the same way as in the simple linear model.

We can also find **deleted residuals**, or **PRESS residuals**,

$$e_{i,-i} = y_i - \hat{y}_{i,-i} = \frac{e_i}{1-h_{ii}}$$

to highlight residuals corresponding to high leverage points.

Again

$$\text{Var}(e_{i,-i}) = \frac{\sigma^2}{1-h_{ii}}$$

so the deleted residuals do not have constant variance. Standardising them and using the same estimate of  $\sigma$ ,  $\hat{S}$ , for each residual, again results in the studentised residuals,  $r_i$ .

Alternatively, the deleted residuals may be standardised by using an estimate of  $\sigma$  that is also based on the data without the  $i$ 'th point,  $s_{-i}$  leading to the **externally studentised residual**,

$$t_i = e_{i,-i} \times \frac{\sqrt{1-h_{ii}}}{s_{-i}}$$

### **Non-constant variance**

This is usually detected from a funnel-shaped scatter of points in a scatterplot of  $e_i$  against the fitted values,  $\hat{y}_i$ .

### **Correlated errors**

Serial correlation is usually detected with the **Durbin-Watson test**. The test statistic,

$$d = \frac{\sum_{t=2}^n (e_t - e_{t-1})^2}{\sum_{t=1}^n e_t^2}$$

is compared with special tables.

### Non-normal errors

A normal probability plot of the studentised residuals,

$$r_i = \frac{e_i}{\hat{S} \sqrt{1 - h_{ii}}}$$

can be used to detect non-normality. Curvature is interpreted in the same way as for the simple linear model.

### Leverage and influence

The values  $h_{ii}$  are often used as measures of the **leverage** of an observation's explanatory variables. Leverage is higher (closer to 1.0) when the explanatory variables are far from the body of explanatory variables in the rest of the data. Leverage again measures the **potential** for a data point to affect the regression results, not its actual influence on these results. The average value of  $h_{ii}$  is  $p/n$ , where  $p$  is the number of parameters in the model (usually the number of explanatory variables plus one). Values greater than double this should be examined carefully.

The effect of a data point on its fitted value, externally standardised, is

$$(DFITS)_i = \frac{\hat{y}_i - \hat{y}_{i,-i}}{s_{-i} \sqrt{h_{ii}}} = t_i \sqrt{\frac{h_{ii}}{1 - h_{ii}}}$$

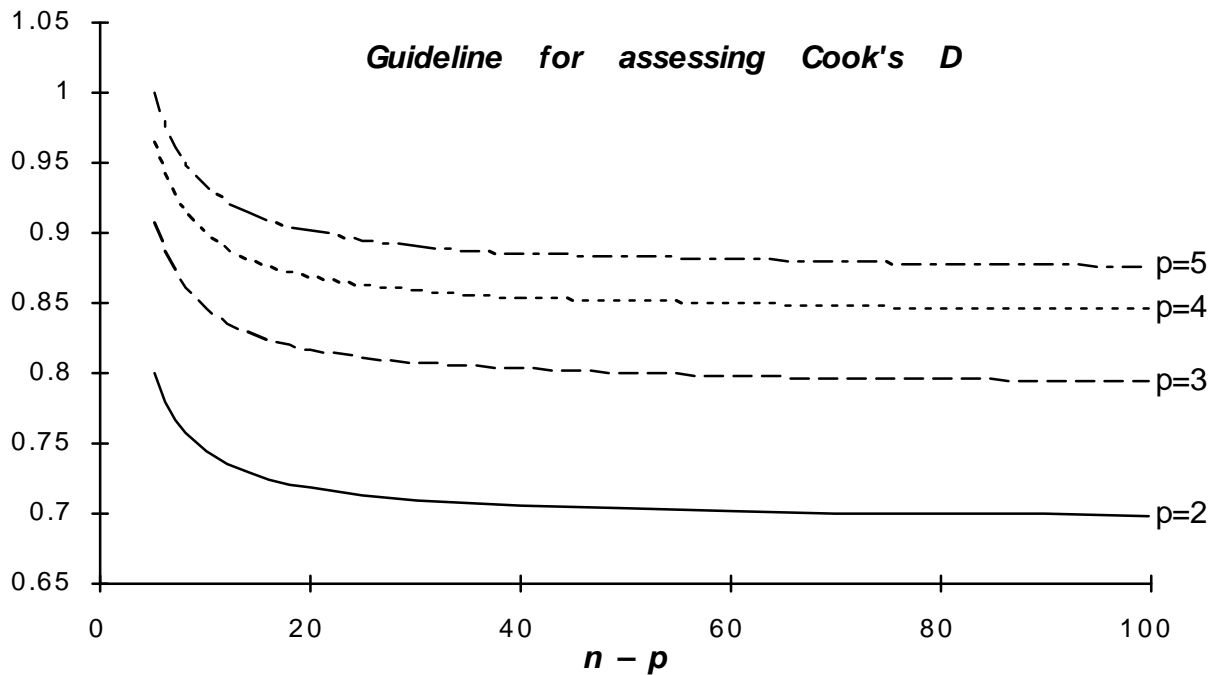
Observations with large  $(DFITS)_i$  have a great influence on the fitted value at that combination of  $x$ -values. A cut-off that is sometimes used to identify 'large'  $(DFITS)_i$  is

$$2 \sqrt{\frac{p}{n - p}}$$

Finally, **Cook's Distance** provides a measure of how much the  $i$ 'th data point influences the regression estimates,

$$D_i = \frac{(\mathbf{b} - \mathbf{b}_{-i})^T (\mathbf{X}^T \mathbf{X}) (\mathbf{b} - \mathbf{b}_{-i})}{p \hat{S}^2} = \left( \frac{r_i^2}{p} \right) \left( \frac{h_{ii}}{1 - h_{ii}} \right)$$

where  $\mathbf{b}$  and  $\mathbf{b}_{-i}$  are the least squares estimates with and without the  $i$ 'th observation respectively. It can be interpreted as a sort of sum of squares of these differences. A yard-stick is to compare with the median of the F-distribution with  $p$  and  $n-p$  degrees of freedom. The graph below shows these values for  $p = 2, 3, 4$  and 5.



As a rule of thumb for  $n > 15$ , we can consider points with  $D_i > 0.7$  as influential for  $p = 2$ . Similarly use  $D_i > 0.8$  for  $p = 3$ , and  $D_i > 0.85$  for  $p > 3$ . Greater accuracy is rarely helpful since the values are only a rough guide, though the computer can evaluate the median of the  $F(n, n - p)$  distribution for any values of  $n$  and  $p$ .

In addition to this guideline, *Cook's  $D$  should be plotted*, either against time order, fitted values or an explanatory variable.

### ***On the Computer***

All of the above diagnostic statistics and plots of them may be obtained in exactly the same ways as described in Chapter 2 for simple linear regression.

---

## **Other Diagnostic Plots**

---

Two further plots are described in this section that help to show graphically the contribution of a single explanatory variable to the fit of a regression model. Confusingly, both plots are given the name 'partial residual plot' by some authors, so it is better to avoid that term.

In this section, to simplify the notation, the  $i$ 'th of the  $p$  explanatory variable will be denoted by  $x_i$ . The response will be similarly denoted by  $y$  and the residuals from the full model ( $y$  against  $x_1$  to  $x_p$ ) will be denoted by  $e$ .

### Added variable plots

An added-variable plot shows graphically the effect of a single explanatory variable on the response. The added-variable plot for variable  $x_i$  plots two residuals against each other.

- residuals from regressing  $y$  against all of  $x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_p$
- residuals from regressing  $x_i$  against all of  $x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_p$

These two residuals can be thought of as adjusting  $y$  and  $x_i$  to remove the effect of the other explanatory variables.

The most important feature of this plot is that the slope of the least squares line fitted to the plot is equal to the regression coefficient of  $x_i$  in the full regression of  $y$  against all  $p$  explanatory variables.

Further, the coefficient of determination for a linear model fitted to the plot is the proportion of the remaining unexplained variation after using  $x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_p$ , that is explained by adding the variable  $x_i$  to the model,

$$\frac{SS_{\text{Regn}}(b_i | b_0, \dots, b_{i-1}, b_{i+1}, \dots, b_p)}{SS_{\text{Resid}}(b_0, \dots, b_{i-1}, b_{i+1}, \dots, b_p)}$$

The plot therefore indicates visually the importance of adding the variable  $x_i$ .

This plot also indicates any points of influence in the regression.

### Component-plus-residual plots

A *component-plus-residual* plot, is intended to indicate whether there is curvature in the effect of a single variable  $x_i$  on the response. It is obtained from the fit of the full model with all explanatory variables. The quantities plotted are

- $e + b_i x_i$  ( $= y - b_0 - b_1 x_1 - \dots - b_{i-1} x_{i-1} - b_{i+1} x_{i+1} \dots - b_p x_p$ )
- $x_i$

Again, the slope of a least squares line fitted to this plot is  $b_i$ .

The plot also indicates whether there is still a nonlinear effect of  $x_i$  on the response,  $y$ , that remains unexplained by the model, and may suggest a nonlinear function of  $x_i$  to use instead of the linear term. It is more effective for this purpose than the added-variable plot since  $x_i$  is on the horizontal axis.

The plot does not however give any indication of the significance of the variable (unlike the added-variable plot).