Simple linear regression
About this course

M346 Linear statistical modelling uses the software package GenStat. This software is provided as part of the course, and its use is introduced in Unit 2.

GenStat Release 10.1 (PC/Windows) Copyright 2008, Lawes Agricultural Trust (Rothamsted Experimental Station). Windows is a registered trademark of Microsoft Corporation.

Cover image courtesy of Rothamsted Research. This image is an aerial photograph of an agricultural trial on mosaic disease in barley. Here the effect of different treatments (in this case, different cultivars of barley) can be seen: the stripes correspond to different plots, and the greenness of each stripe indicates how healthy the barley is in that plot. For more information see the Rothamsted Research Annual Report 2005–2006.

This publication forms part of an Open University course. Details of this and other Open University courses can be obtained from the Student Registration and Enquiry Service, The Open University, PO Box 197, Milton Keynes MK7 6BJ, United Kingdom: tel. +44 (0)845 300 6090, email general-enquiries@open.ac.uk

Alternatively, you may visit the Open University website at http://www.open.ac.uk where you can learn more about the wide range of courses and packs offered at all levels by The Open University.

To purchase a selection of Open University course materials visit http://www.ouw.co.uk, or contact Open University Worldwide, Walton Hall, Milton Keynes MK7 6AA, United Kingdom, for a brochure: tel. +44 (0)1908 858793, fax +44 (0)1908 858787, email ouw-customer-services@open.ac.uk

The Open University, Walton Hall, Milton Keynes, MK7 6AA.

First published 2009.

Copyright © 2009 The Open University

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, transmitted or utilised in any form or by any means, electronic, mechanical, photocopying, recording or otherwise, without written permission from the publisher or a licence from the Copyright Licensing Agency Ltd. Details of such licences (for reprographic reproduction) may be obtained from the Copyright Licensing Agency Ltd, Saffron House, 6–10 Kirby Street, London EC1N 8TS; website http://www.cla.co.uk.

Open University course materials may also be made available in electronic formats for use by students of the University. All rights, including copyright and related rights and database rights, in electronic course materials and their contents are owned by or licensed to The Open University, or otherwise used by The Open University as permitted by applicable law.

In using electronic course materials and their contents you agree that your use will be solely for the purposes of following an Open University course of study or otherwise as licensed by The Open University or its assigns.

Except as permitted above you undertake not to copy, store in any medium (including electronic storage or use in a website), distribute, transmit or retransmit, broadcast, modify or show in public such electronic materials in whole or in part without the prior written consent of The Open University or in accordance with the Copyright, Designs and Patents Act 1988.

Edited, designed and typeset by The Open University, using the Open University TeX System.

Printed in the United Kingdom by The Charlesworth Group, Wakefield.

ISBN 978 0 7492 2973 3

1.1
Contents

Study guide 4
Introduction 4
3.1 The simple linear regression model 5
3.2 Fitting lines and making inferences 9
3.3 Confidence intervals and prediction 15
3.4 Checking the assumptions 18
3.5 Transformations 22
3.6 Comparing slopes 27
3.7 Looking forward: extensions to simple linear regression 30
3.8 Correlation 31
Summary 33
Learning outcomes 33
Solutions to exercises 34
Index 52
Study guide

You should schedule eight study sessions for this unit. This includes time for answering the TMA question associated with this unit and for consolidating your work.

The sections vary in length. In particular Section 3.7 is much shorter than average. As you study this unit, various exercises will require the use of your computer.

One possible study pattern is as follows.

Study session 1: Section 3.1.
Study session 2: Section 3.2.
Study session 3: Section 3.3.
Study session 4: Section 3.4.
Study session 5: Section 3.5.
Study session 6: Sections 3.6 and 3.7.
Study session 7: Section 3.8.
Study session 8: TMA question on Unit 3.

Introduction

The object of this unit is to review the ideas and techniques of simple linear regression and to carry out such analyses in GenStat. You should be familiar with much (but probably not all) of this material from your previous studies, though the emphasis in this unit may be on rather different aspects from those you have concentrated on before. The terminology ‘simple linear regression’ will be used in this course because it is standard rather than because it is unambiguously descriptive.

In simple linear regression, there is only one explanatory variable, plus a response variable. The response variable is continuous. In many applications the explanatory variable is continuous as well, but this need not be the case, as we shall see. The relationship between the explanatory variable and the response variable is assumed to be linear, and the distribution of the random variables that measure the differences between the observed values of the response variable and the values predicted by the linear regression equation is taken to be normal. This sounds like rather a limited model, with rather a lot of assumptions. It is. But the model is appropriate for a surprisingly wide range of data situations, and (more importantly for this course) by generalising the model a bit at a time — adding more explanatory variables, using different distributions, and so on — we can derive all the other data models that will be introduced in the course. It is thus very important that you understand this basic linear regression model and that you can use it effectively.

Regression models are the focus of the course, but regression relationships are not the only way in which variables can be related. You should also be familiar with the idea of association or correlation, which is briefly reviewed in Section 3.8: it also has a role, albeit a comparatively minor one, to play in the course.

If you have studied M248, the only new material in this unit is some of Section 3.5, on transformations, all of Section 3.6, on comparing slopes, and most of Section 3.7. However, the emphasis in this unit is on rather different aspects than the corresponding material in M248.
3.1 **The simple linear regression model**

In this section we begin by looking briefly at an example of data that can be analysed using simple linear regression, to remind you of what is involved.

**Exercise 3.1  Plotting resistance of rubber to abrasion**

The data in Table 3.1 come from an experiment to investigate how the resistance of rubber to abrasion is affected by various factors. Each of 30 samples of rubber was tested for hardness (measured in degrees Shore: the larger the number, the harder the rubber), and was then subjected to steady abrasion for a fixed time. The weight loss due to abrasion was measured in grams per hour.

*Table 3.1  Hardness and weight loss in samples of rubber*

<table>
<thead>
<tr>
<th>Abrasion loss (g/h)</th>
<th>Hardness (degrees Shore)</th>
<th>Abrasion loss (g/h)</th>
<th>Hardness (degrees Shore)</th>
</tr>
</thead>
<tbody>
<tr>
<td>372</td>
<td>45</td>
<td>196</td>
<td>68</td>
</tr>
<tr>
<td>206</td>
<td>55</td>
<td>128</td>
<td>75</td>
</tr>
<tr>
<td>175</td>
<td>61</td>
<td>97</td>
<td>83</td>
</tr>
<tr>
<td>154</td>
<td>66</td>
<td>64</td>
<td>88</td>
</tr>
<tr>
<td>136</td>
<td>71</td>
<td>249</td>
<td>59</td>
</tr>
<tr>
<td>112</td>
<td>71</td>
<td>219</td>
<td>71</td>
</tr>
<tr>
<td>55</td>
<td>81</td>
<td>186</td>
<td>80</td>
</tr>
<tr>
<td>45</td>
<td>86</td>
<td>155</td>
<td>82</td>
</tr>
<tr>
<td>221</td>
<td>53</td>
<td>114</td>
<td>89</td>
</tr>
<tr>
<td>166</td>
<td>60</td>
<td>341</td>
<td>51</td>
</tr>
<tr>
<td>164</td>
<td>64</td>
<td>340</td>
<td>59</td>
</tr>
<tr>
<td>113</td>
<td>68</td>
<td>283</td>
<td>65</td>
</tr>
<tr>
<td>82</td>
<td>79</td>
<td>267</td>
<td>74</td>
</tr>
<tr>
<td>32</td>
<td>81</td>
<td>215</td>
<td>81</td>
</tr>
<tr>
<td>228</td>
<td>56</td>
<td>148</td>
<td>86</td>
</tr>
</tbody>
</table>

Load `rubber.gsh` into a GenStat spreadsheet, and confirm that the first two columns of data (loss and hardness) correspond to those in Table 3.1. (The third column, strength, contains another variable which we will come back to later.)

Plot loss against hardness. To do this: select the Graphics/2D Scatter Plot menu item. A dialogue box will appear. Check that the Type of plot is given as Single XY. In the Y Data field, enter loss. This can be done either by selecting loss from the Select Y drop-down list, or by typing loss in the Select Y field and pressing <Return>. Similarly, in the X Data field, enter hardness. Click on Finish.

On the basis of the plot of these data, how are abrasion loss and hardness related?

You will be aware from your previous studies that the basic model for linear regression involves several different ideas, which will be reviewed here.

Suppose that the data consist of a set of points \((x_i, y_i), i = 1, 2, \ldots, n\). We wish to find a model that will allow us to say something useful about what the response will be for any given value of the explanatory variable. Thus the explanatory and response variables have very different roles in the model. In setting up the regression model, these different roles are taken account of by treating the values of the explanatory variable as fixed and by treating the response as a random variable. To be precise, denote by \(x_i\) the value of the explanatory variable for the \(i\)th datapoint, and denote by \(Y_i\) the random variable representing the response for the same datapoint.
Unit 3 Simple linear regression

The first assumption made is that the mean of the response variable $Y_i$ depends on the value $x_i$ of the explanatory variable in a linear fashion, where the linearity is in $x$ as well as in the parameters. This can be written as

$$E(Y_i) = \alpha + \beta x_i,$$  \hspace{1cm} (3.1)

where $\alpha$ and $\beta$ are (unknown) constants — the intercept and slope of the regression line, respectively. By way of interpretation, the intercept $\alpha$ provides the mean value of $Y$ when $x = 0$. The slope parameter $\beta$ is such that an increase of one unit in $x$ results in an increase of $\beta$ units in $E(Y)$. $\alpha$ and $\beta$ are known collectively as the regression coefficients of the regression line.

The second assumption is that the variation of the value of the response variable $Y_i$ about this mean is represented by a random variable $\epsilon_i$, which has a normal distribution. The $\epsilon_i$ are often called random ‘errors’. The use of the word ‘errors’ does not imply that random errors are incorrect in any way; the terminology is historical and derives from (actual) errors in measurement. The model becomes

$$Y_i = \alpha + \beta x_i + \epsilon_i.$$  \hspace{1cm} (3.2)

For the mean of $Y_i$ to be as given in (3.1), $E(\epsilon_i)$ has to be zero; but what about its variance? The third assumption is that the variance of $\epsilon_i$ is the same for all values of the explanatory variable. It is called the error variance (or, sometimes, the ‘residual variance’) and is usually denoted by $\sigma^2$. Therefore $\epsilon_i \sim N(0, \sigma^2)$.

There is a fourth assumption. It is not always stated explicitly, but it remains important. It is that the different $\epsilon_i$, $i = 1, 2, \ldots, n$, are independent of each other. In other words, the deviations of the response from the mean are taken to be purely random; they do not affect each other.

There is an alternative, but equivalent, way of writing model (3.2). Considering the value of each $x_i$ to be fixed, $Y_i$ is made up of a fixed quantity, $\alpha + \beta x_i$, plus a $N(0, \sigma^2)$ random variable. But from property (1.3) in Subsection 1.2.4, $aX + b \sim N(a\mu + b, a^2\sigma^2)$; this means that $Y_i$ is itself normally distributed with mean $\alpha + \beta x_i$ and variance $\sigma^2$. (To see this, set $X = Y_i$, $a = 1$ and $b = \alpha + \beta x_i$ in the above property.) The simple linear regression model can therefore be written as in Box 3.1.

**Box 3.1**

The simple linear regression model can be written

$$Y_i \sim N(\alpha + \beta x_i, \sigma^2),$$

where the $Y_i$ are independent random variables and the $x_i$ are values of the explanatory variables.

This formulation specifies the conditional distribution of each response random variable $Y_i$ given (or conditional on) the corresponding value $x_i$ of the explanatory variable. Each $Y_i$ has a normal distribution, but a different normal distribution, its parameters — in fact, just its mean — depending on $x_i$. This conditional distribution specification will prove particularly amenable to generalisation later in the course.

**Exercise 3.2  Assessing the appropriateness of simple linear regression from a scatterplot**

For the data in Exercise 3.1, do you think the simple linear regression model is appropriate, given the scatterplot in Solution 3.1?
Section 3.1 The simple linear regression model

For some data, there are reasons to use a slightly simpler model, in which the regression line is constrained to pass through the origin; that is, the (mean) response must be zero if the value of the explanatory variable is zero. In this case, the model is that

\[ Y_i = \gamma x_i + \epsilon_i, \]

where again the error terms \( \epsilon_i \) are normally distributed with mean 0 and variance \( \sigma^2 \) and are independent of each other. Equivalently,

\[ Y_i \sim N(\gamma x_i, \sigma^2), \]

with the \( Y_i \) values independent.

Exercise 3.3 Plotting the effectiveness of cavity wall insulation

The data in Table 3.2 (which you met first, briefly, in Table 2.1 in Section 2.2) come from a study of the effectiveness of cavity wall insulation. The total energy consumption (in MW h) of each of ten houses of a particular type in the Fishponds area of Bristol was recorded over one winter. Then cavity wall insulation was installed, and the total energy consumption was recorded again over the next winter.

Table 3.2 Energy consumption before and after cavity wall insulation

<table>
<thead>
<tr>
<th>Before insulation</th>
<th>After insulation</th>
</tr>
</thead>
<tbody>
<tr>
<td>12.1</td>
<td>12.0</td>
</tr>
<tr>
<td>11.0</td>
<td>10.6</td>
</tr>
<tr>
<td>14.1</td>
<td>13.4</td>
</tr>
<tr>
<td>13.8</td>
<td>11.2</td>
</tr>
<tr>
<td>15.5</td>
<td>15.3</td>
</tr>
<tr>
<td>12.2</td>
<td>13.6</td>
</tr>
<tr>
<td>12.8</td>
<td>12.6</td>
</tr>
<tr>
<td>9.9</td>
<td>8.8</td>
</tr>
<tr>
<td>10.8</td>
<td>9.6</td>
</tr>
<tr>
<td>12.7</td>
<td>12.4</td>
</tr>
</tbody>
</table>

Dataset name: bristol.
Source: data from the Electricity Council (1983).

A possible model for these data is one in which the energy use after insulation, \( y \), is (on average) a certain percentage of the energy use before insulation, \( x \), that is,

\[ E(Y) = \gamma x, \]

where \( \gamma \) is an unknown constant. If this model is appropriate, a straight line constrained to pass through the origin should be fitted to these data.

Load the dataset into GenStat and produce a scatterplot, using the method described in Exercise 3.1.

It is not easy, from GenStat’s default scatterplot, to see whether a straight line through the origin is appropriate, because the origin \((0,0)\) does not appear on the plot. To alter the scatterplot so that it is more helpful, double-click on the scatterplot to open the GenStat Graphics Editor. In this new window, select the Tools|Graph Options menu item. On the X Axis tab of the Options dialogue box, select Zero for the Y-Origin. Similarly, on the Y Axis tab, select Zero for the Y-Origin. Click on OK. This alters where the axes cross on the resulting plot. In the GenStat Graphics Editor, click on File|Save and Close to return to the GenStat Graphics Viewer.

Use the resulting diagram to see whether, on the face of it, a straight line constrained to pass through the origin is appropriate for these data.

Of course, the assumptions of simple linear regression do not hold for all data consisting of pairs of \( x \) and \( y \) values.
Example 3.1  Effect of oxygen uptake on lung function

The scatterplot in Figure 3.1 is based on data collected in an experiment in kinesiology. A person performed a standard exercise task at a gradually increasing level. The two variables are oxygen uptake and expired ventilation, which is related to the rate of exchange of gases in the lungs. How does expired ventilation depend on oxygen uptake?

Dataset name: anaerob.

Figure 3.1  The effect of oxygen uptake on expired ventilation

There is clearly no point in fitting a straight line through these data! It is, however, possible to fit more complicated models, using methods described in later units. For instance, Figure 3.2 shows the results of fitting a quadratic function of oxygen uptake, i.e. one of the form \( \alpha + \beta x + \gamma x^2 \). The quadratic model is linear in its parameters but not in \( x \). Fitting such models is discussed in Unit 5.

Figure 3.2  A quadratic model of expired ventilation
Exercise 3.4  Assessing a model for extinction rates of marine animals

The data whose first five pairs of values are given in Table 3.3 were collected as part of a study of extinctions in past geological ages. Geologists divide time since the start of the Pernian period, some 290 million years ago, into a series of 48 stages (on what is known as the Harland time scale). For each of these stages, the table records when it ended (measured in millions of years before the present time, or MYBP) and also the percentage of genera of marine animals that became extinct during the period. The aim of the study was to investigate how extinction rates had changed over time, and also to identify the stages at which ‘mass extinctions’ (abnormally high extinction rates) occurred.

Table 3.3  Extinction rates of marine animals over time

<table>
<thead>
<tr>
<th>Percentage extinct</th>
<th>Time (MYBP)</th>
</tr>
</thead>
<tbody>
<tr>
<td>22</td>
<td>265</td>
</tr>
<tr>
<td>23</td>
<td>258</td>
</tr>
<tr>
<td>61</td>
<td>253</td>
</tr>
<tr>
<td>60</td>
<td>248</td>
</tr>
<tr>
<td>45</td>
<td>243</td>
</tr>
<tr>
<td>:</td>
<td>:</td>
</tr>
</tbody>
</table>

Produce a scatterplot of the data, with time (variate mybp) as the explanatory variable. Comment on the appropriateness of the simple linear regression model for these data.

Section 3.2  Fitting lines and making inferences

We have a linear regression model, but it contains several unknown parameters. In the case where the regression line is not constrained to pass through the origin, these are the intercept $\alpha$, the slope $\beta$ and the error variance $\sigma^2$. In this section, we shall discuss how to estimate these parameters and how to make inferences about their values.

The principle of estimation that is generally used in simple linear regression is that of least squares. The idea is to look at the discrepancies, or residuals, between the values of the response variable that were actually observed and the values that would be predicted by the estimated regression line. Suppose that the data are $(x_i, y_i)$, $i = 1, 2, \ldots, n$, and the parameters $\alpha$ and $\beta$ are estimated by $\hat{\alpha}$ and $\hat{\beta}$. Then the values of the response variable predicted by the regression line are

$$\hat{\alpha} + \hat{\beta}x_i, \quad i = 1, 2, \ldots, n,$$

and the residuals are given by

$$r_i = y_i - (\hat{\alpha} + \hat{\beta}x_i), \quad i = 1, 2, \ldots, n. \quad (3.3)$$

Some books use other symbols (often $\epsilon_i$ or $w_i$) for residuals. If the estimated line is to fit the data as closely as possible, then we want the residuals to be small. In this context, ‘small’ is taken to mean that we want to
choose \( \hat{\alpha} \) and \( \hat{\beta} \) to make the sum of the squared residuals, that is, the quantity
\[
\sum_{i=1}^{n} r_i^2 = \sum_{i=1}^{n} \left( y_i - (\hat{\alpha} + \hat{\beta} x_i) \right)^2,
\]
as small as possible — this explains the term ‘least squares’. Then algebra can be used to show that the least squares estimate of the slope is
\[
\hat{\beta} = \frac{n \sum x_i y_i - \sum x_i \sum y_i}{n \sum x_i^2 - (\sum x_i)^2},
\]
and the least squares estimate of the intercept is
\[
\hat{\alpha} = \bar{y} - \hat{\beta} \bar{x}, \tag{3.4}
\]
where \( \bar{x} \) and \( \bar{y} \) are the sample means of the \( x \) and \( y \) values. (The regression line thus goes through the point with coordinates \((\bar{x}, \bar{y})\).) For regression through the origin, where the model is \( Y_i = \gamma x_i + \epsilon_i \), the least squares estimate of the slope \( \gamma \) is
\[
\hat{\gamma} = \frac{\sum x_i y_i}{\sum x_i^2}.
\]
These estimates are not difficult to calculate by hand, though the process is pretty tedious except for small datasets. However, in the rest of this unit and generally in this course, you will be calculating the estimates using GenStat rather than applying the formulae directly.

In most of this course, the method of maximum likelihood is used to find estimators of unknown quantities. So why is another method being used here? In fact, with the assumptions we have made about the random errors \( \epsilon_i \) (i.e. normality, constant variance and independence), the maximum likelihood estimators of \( \alpha \) and \( \beta \) are exactly the same as the least squares estimators. If other assumptions are made about the distribution of the \( \epsilon_i \), however, the maximum likelihood estimators and the least squares estimators no longer coincide, and things can get much more complicated.

The least squares procedure for estimating \( \alpha \) and \( \beta \) does not provide an estimate of the other parameter in the model, the error variance \( \sigma^2 \). In the case of the unconstrained line (i.e. not necessarily through the origin), the estimate of \( \sigma^2 \) which is used is
\[
s^2 = \frac{\sum (y_i - \hat{y}_i)^2}{n - 2}, \tag{3.5}
\]
where \( \hat{y}_i = \hat{\alpha} + \hat{\beta} x_i \) is the (estimated) fitted value of \( Y \) at the \( i \)th datapoint. The numerator of the expression for \( s^2 \) is the value of the quantity (the sum of the squared residuals) that is minimised in finding the least squares estimates of the intercept and slope. The usual term for this is the residual sum of squares or RSS. The denominator is \( n - 2 \) essentially because two parameters, \( \alpha \) and \( \beta \), have already been estimated from the \( n \) datapoints.

For regression through the origin, the estimate of \( \sigma^2 \) which is used is
\[
s^2 = \frac{\sum (y_i - \hat{y}_i)^2}{n - 1}, \] (where \( \hat{y}_i = \hat{\gamma} x_i \) is again the fitted value of \( Y \) at the \( i \)th datapoint. The denominator is \( n - 1 \) because only one parameter has been estimated.

The residuals may appear, at first glance, to be the observed values of the normally distributed random errors \( \epsilon_i \) that appear in the simple linear regression model. In fact, they are not. The calculation \( \epsilon_i = y_i - (\alpha + \beta x_i) \) requires the true values \( \alpha \) and \( \beta \) of the intercept and slope to be known. But in practice the intercept and slope have to be estimated, and so it is \( \hat{\alpha} \) and \( \hat{\beta} \) that appear in the definition of the residuals, Equation (3.3). However, it is generally true that the residuals are reasonable estimates of the \( \epsilon_i \), so they can be used to investigate whether the modelling assumptions about \( \epsilon_i \) are satisfied. (There is more on this later.)

**Exercise 3.5 Fitting a simple linear regression line**

In this exercise you will use GenStat to fit a straight line to the data on abrasion loss and hardness in Table 3.1, using the method of least squares.
Load `rubber.gsh` into a GenStat spreadsheet again if necessary. Choose the Stats/Regression Analysis/Linear Models menu item. A dialogue box will appear, the Linear Regression dialogue box. The Regression field contains Simple Linear Regression, which is what we want. Fill in the Response Variate field as `loss` (either by double-clicking on `loss` in the Available Data area, or by entering the name directly); fill in the Explanatory Variate field as `hardness`; click on Run.

(a) Look in the Output window at the output produced by fitting the line. There will probably be parts that you do not recognise, but some parts should be familiar. Which part of the output do you recognise and what does it tell you?

(b) Now produce a scatterplot of the data with the fitted line plotted on it. To do this, go back to the Linear Regression dialogue box, which should still be open. (Double-click on Linear Regression in the Window panel to bring it back to the front — if it is not open, you will have to go through the process of fitting the line again before you can produce the plot.) Click on the Further Output button. Another dialogue box will appear; click on the Fitted Model button to produce the plot. Comment on whether the line seems to fit the data well.

**Exercise 3.6  Modelling contamination in peanuts**

The data in `peanuts` comprise, for 34 batches of peanuts, the average level of the fungal contaminant aflatoxin in a sample of 120 pounds of peanuts, and the percentage of non-contaminated peanuts in the whole batch. The data, five values of which are given in Table 3.4, were collected with the aim of being able to predict the percentage of non-contaminated peanuts (variates `percent`) from the aflatoxin level (variates `toxin`) in a sample.

| Table 3.4  Contamination and aflatoxin level in samples of peanuts |
|-----------------|-----------------|
| Percentage non-contaminated | Aflatoxin level (parts per billion) |
| 99.971 | 3.0 |
| 99.979 | 4.7 |
| 99.982 | 8.3 |
| 99.971 | 9.3 |
| 99.957 | 9.9 |
| | |

(a) Make a scatterplot of the data in `peanuts`. Describe the relationship between the two variables. Would you say that the percentage of non-contaminated peanuts in a batch could be predicted accurately from the level of aflatoxin in a sample?

(b) Use GenStat to fit a straight line to these data.

**Exercise 3.7  Fitting a straight line through the origin**

For the data in Table 3.2 on house insulation, you saw that a straight line through the origin (treating the ‘before’ data as the explanatory variable) looked reasonable, on the basis of a scatterplot at least. To fit such a line using GenStat, load the data (`bristol.gsh`), and obtain the Linear Regression dialogue box as in Exercise 3.5. Enter after as the Response Variate and before as the Explanatory Variate. Click on Options. In the resulting dialogue box, click on Estimate Constant Term to remove its check mark. (This causes GenStat to leave out the constant term from the regression equation, so the line is forced to go through the origin.) Click on OK, and then on Run in the Linear Regression dialogue box. What is the equation of the fitted line?

Plot the line on a scatterplot of the data, using the same method as in Exercise 3.5. Comment on the fit.
Exercise 3.8  Modelling finger tapping

An experiment was carried out to investigate the effect of the stimulant caffeine on performance of a simple physical task. Thirty male college students were trained in finger tapping. They were then divided at random into three groups of ten, and the groups received different doses of caffeine (0, 100 and 200 mg). Two hours after this treatment, each man was required to do finger tapping and the number of taps per minute was recorded. The data are given in Table 3.5.

Table 3.5  Finger taps in response to caffeine

<table>
<thead>
<tr>
<th>Caffeine (mg)</th>
<th>0mg caffeine</th>
<th>100mg caffeine</th>
<th>200mg caffeine</th>
</tr>
</thead>
<tbody>
<tr>
<td>Taps</td>
<td>242 245 244</td>
<td>248 247 248</td>
<td>242 244 246</td>
</tr>
<tr>
<td></td>
<td>242</td>
<td>248 246 245</td>
<td>248 250 247</td>
</tr>
<tr>
<td></td>
<td></td>
<td>248 252 248</td>
<td>250 246 244</td>
</tr>
</tbody>
</table>

The aim was to investigate whether caffeine affected performance on this task and, if it did, to describe how the effect was related to the dose.

At first, these data do not look like typical regression data. Table 3.5 does not consist of a list of paired measurements. However, the question of interest is about the relationship between an explanatory variable $x$ (dose of caffeine, variate dose) and a response variable $y$ (number of taps, variate taps); and one can think of the data in the form of $(x, y)$ pairs: $(0, 242), (0, 245),\ldots, (200, 248), (100, 246),\ldots, (100, 244), (200, 246), (200, 248),\ldots, (200, 250)$.

(a) Produce a scatterplot of the data in Table 3.5. Do you think it is appropriate to fit a straight line?

(b) Fit a straight line to the data. If you have not restarted GenStat since you fitted the line through the origin in Exercise 3.7, GenStat will again fit a line through the origin, which is not appropriate here. To avoid this, click on the Options button and re-check the Estimate Constant Term option before clicking OK and then Run.

The least squares method provides estimates of the slope and intercept of the regression line; but they are only estimates, and — like all estimates from sample data — are subject to error. In order to use statistical methods to make inferences from the sample estimates of the parameters of the regression line to the population from which the sample was drawn, we must use some results on the sampling distributions of the estimators involved. The key results are as follows.

For the unconstrained regression model, assuming that the random errors $e_i$ are normally distributed, the intercept and slope estimators $\hat{\alpha}$ and $\hat{\beta}$ are both normally distributed:

$$\hat{\alpha} \sim N\left(\alpha, \frac{\sigma^2}{n} \left(1 + \frac{n\bar{x}^2}{s_{xx}}\right)\right), \quad \hat{\beta} \sim N\left(\beta, \frac{\sigma^2}{s_{xx}}\right),$$

where

$$s_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2.$$

However, $\hat{\alpha}$ and $\hat{\beta}$ are not independent (except when $\bar{x}$ happens to be zero). In practice these results are not very useful for making inferences because the variances of both estimators involve the error variance $\sigma^2$, which is unknown. The estimator $s^2$ of the error variance, as defined in Equation (3.5), has the distribution

$$\frac{(n-2)s^2}{\sigma^2} \sim \chi^2(n-2).$$

This can be used to obtain the following very useful result:

$$\frac{\hat{\beta} - \beta}{s/\sqrt{s_{xx}}} \sim t(n-2). \quad (3.6)$$

The $\chi^2$-distribution and Student’s $t$-distribution were reviewed in Unit 1.
This result can be used, for instance, to test the null hypothesis that the true value $\beta$ of the slope of a regression line is zero. With $\beta = 0$, the left-hand side of relation (3.6) can be calculated from the data as the $t$-statistic for this $t$-test, and then compared with the $t$-distribution with $n - 2$ degrees of freedom to calculate a significance probability. You will see how to use the same result to calculate a confidence interval for $\beta$ in Section 3.3.

Closely related results hold for fitting a line constrained to go through the origin. The details are not given here.

**Exercise 3.9 Interpreting the Estimates of parameters output**

For the rubber data of Table 3.1, carry out the regression analysis again with loss as the response variable and hardness as the explanatory variable, as in Exercise 3.5. Look at the output and see if you can work out how the information in the Estimates of parameters section relates to (3.6).

What of the rest of the output from the regression of abrasion loss on hardness? It is, sensibly enough, headed Regression analysis. You are then reminded that you took loss to be what GenStat calls the Response variate, and you are also reminded of the explanatory variable used, hardness, in what GenStat calls Fitted terms. In fact, GenStat gives the useful reminder that you are fitting both a constant (i.e. a non-zero intercept) and one explanatory variable (hardness).

Next comes a table headed Summary of analysis. You will not have seen such a table before, but in ensuing units tables like these will become a familiar sight. Later on, they will often be called ‘analysis of variance’ tables, for reasons that can remain obscure for now. Here is the one associated with the analysis of Exercise 3.9.

<table>
<thead>
<tr>
<th>Source</th>
<th>d.f.</th>
<th>s.s.</th>
<th>m.s.</th>
<th>v.r.</th>
<th>F pr.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>1</td>
<td>122455</td>
<td>122455</td>
<td>33.43</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Residual</td>
<td>28</td>
<td>102556</td>
<td>3663.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>29</td>
<td>225011</td>
<td>7759.</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The table consists of three rows and five columns, not counting row and column labels (and not all cells have an entry in them).

The rather abstract explanation that follows will become clearer in the specific contexts of later units, so don’t worry if you do not grasp every detail of it at this stage.

Explanation of the table is facilitated by starting with the meaning of the column label s.s.. This stands for ‘sum of squares’. As you already know, the sum of squares of numbers $z_1, z_2, \ldots, z_N$ is simply

$$z_1^2 + z_2^2 + \cdots + z_N^2 = \sum_{i=1}^{N} z_i^2.$$

Let us look down the s.s. column and identify the corresponding row labels. All that is done to the data to obtain the values in the row labelled Total is that the responses $y_1, y_2, \ldots, y_n$ have their overall mean $\bar{y}$ subtracted off. Thus s.s.(Total) is just the sum of squares of what is left after this operation:

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

The word Total refers to this being a measure of the total variability (about the mean) in the data.

The s.s.(Residual) entry is the residual sum of squares. Recall that each residual is the difference between $y_i$ and its value $\hat{y}_i = \hat{\alpha} + \hat{\beta}x_i$ predicted by the regression model. So s.s.(Residual) is the amount of variability remaining in the data once the regression model is fitted.
Unit 3  Simple linear regression

The first two entries in the s.s. column add up to the last; by definition,
\[
\text{s.s.}(\text{Regression}) + \text{s.s.}(\text{Residual}) = \text{s.s.}(\text{Total}),
\]
or equivalently,
\[
\text{s.s.}(\text{Regression}) = \text{s.s.}(\text{Total}) - \text{s.s.}(\text{Residual}).
\]
We can therefore say that \text{s.s.}(\text{Regression}) is the amount of variation in the (response) data that is taken account of by the linear relationship between the mean response and the explanatory variable.

The d.f. heading stands for ‘degrees of freedom’. The d.f.(Total) entry is \(n - 1\) where \(n\) is the sample size. \(30 - 1 = 29\) in the abrasion loss case. If there were no explanatory variable, and we had only the set of response data, we would estimate the variance of the responses by dividing s.s.(Total) by precisely this quantity \(n - 1\). This, in fact, is what is done in the m.s. column:
\[
\text{m.s.}(\text{Total}) = \frac{\text{s.s.}(\text{Total})}{\text{d.f.}(\text{Total})}
\]
The abbreviation m.s. stands for ‘mean square’. The \(n - 1\) is the degrees of freedom of the \(\chi^2\)-distribution associated with the variance estimator m.s.(Total) under normality (Subsection 1.5.1).

Also, d.f.(Residual) = \(n - 2\) (which is 28 in this case). This is precisely the divisor of s.s.(Residual) used in Equation (3.5) to form the estimate \(s^2\) of the error variance (or residual variance) \(\sigma^2\). In fact,
\[
\text{m.s.}(\text{Residual}) = \frac{\text{s.s.}(\text{Residual})}{\text{d.f.}(\text{Residual})} = s^2.
\]
So, the estimate of the error variance in this case can be read from the table to be 3663. Again, the d.f. is the degrees of freedom of the associated \(\chi^2\)-distribution.

As sums of squares of regression and residual add up to that of the total, so degrees of freedom add up in the same way. (Note that mean squares do not add up in this way.) Thus,
\[
\text{d.f.}(\text{Regression}) = \text{d.f.}(\text{Total}) - \text{d.f.}(\text{Residual})
= (n - 1) - (n - 2) = 1.
\]
In fact, the 1 corresponds to the one extra parameter that has been fitted by the regression, the slope parameter \(\beta\). The intercept, \(\alpha\), is not an extra parameter in this sense, because in the Total row of the table the sample mean has already been subtracted, which essentially corresponds to fitting the model \(E(Y) = \alpha\). And
\[
\text{m.s.}(\text{Regression}) = \frac{\text{s.s.}(\text{Regression})}{1} = \text{s.s.}(\text{Regression}).
\]
The table directly addresses one particular null hypothesis, and that is whether \(\beta = 0\). The test statistic it uses to test this hypothesis is given under v.r., which stands for ‘variance ratio’. In fact,
\[
\text{v.r.} = \frac{\text{m.s.}(\text{Regression})}{\text{m.s.}(\text{Residual})}.
\]
Now, under the null hypothesis that \(\beta = 0\), it turns out that m.s.(Regression) and m.s.(Residual) are both estimates of \(\sigma^2\) (hence the term ‘variance ratio’). It also turns out (and you are not expected to see why) that the two estimates are independent, that
\[
\frac{\text{m.s.}(\text{Residual})}{\sigma^2} \sim \chi^2(\text{d.f.}(\text{Residual}))
\]
and that, under \(H_0\),
\[
\frac{\text{m.s.}(\text{Regression})}{\sigma^2} \sim \chi^2(\text{d.f.}(\text{Regression})).
\]
This is just the setting of Subsection 1.5.2, implying that the ratio v.r. has, under the null hypothesis, an \(F\)-distribution on d.f.(Regression) and d.f.(Residual) degrees of freedom. GenStat works out the \(p\) value for this test for you and gives
Section 3.3 Confidence intervals and prediction

It under F pr., which stands for ‘F probability’: explicitly, in this case, it is \( P(F \geq 33.43) \) where \( F \sim F(1, 28) \). The important thing to grasp from this paragraph is that, under \( H_0 : \beta = 0 \), v.r. \( \sim F(\text{d.f.}(\text{Regression}), \text{d.f.}(\text{Residual})) \).

**Exercise 3.10 Testing \( \beta = 0 \)**

The GenStat output from Exercise 3.9 has given you two tests of \( H_0 : \beta = 0 \). What are they? Are the two \( p \) values the same? Is there some very close connection between the two tests, particularly the two test statistics?

Beneath the Summary of analysis table, GenStat prints out some further information before coming to the Estimates of parameters. The first line of this is to give you the **Percentage variance accounted for**. This statistic is interpreted as a measure of the strength of the linear association of the response variable with the explanatory variable. (Its formula is not important just now.) A perfect linear relationship would produce 100%; a value close to 100 is generally interpreted as ‘strong’ linear relationship, whereas a value close to zero is interpreted as ‘weak’. However, precisely what is meant by ‘strong’ or ‘weak’ in this context is difficult to enunciate. The abrasion loss’s 52.8% variance accounted for is in a grey area, which to many people is not very strong, but to some is.

The value in the **Standard error of observations** line is simply the estimated error standard deviation (or residual standard deviation), i.e. the square root, \( s \), of the \( \text{m.s.}(\text{Residual}) \). In this case, \( s = \sqrt{3663} = 60.5 \).

Underneath these lines is the space GenStat reserves for any warning **Message**s. In this case, it mentions a ‘unit’ which seems to have ‘high leverage’. This is a concept to do with model checking that you should ignore for now, and indeed not worry about until it is considered properly in Unit 10. Often, but not this time, GenStat also uses this area to warn of particularly large residuals.

**Exercise 3.11 Modelling finger tapping, continued**

For the data in Table 3.5 (dataset taps), investigate whether caffeine does indeed have an effect on performance of the tapping task by testing the hypothesis that the true slope of the regression line is zero. Report your conclusions.

### 3.3 Confidence intervals and prediction

This section investigates other statistical inferences that can be made on the basis of regression models, namely confidence intervals for regression parameters and inferences to do with prediction.

First, let us look at confidence intervals. From relation (3.6), a 100(1 − \( \delta \))% confidence interval for \( \beta \) can be found as

\[
\hat{\beta} \pm t_{1-\delta/2}(n-2) \frac{s}{\sqrt{\hat{\Sigma}_{xx}}}
\]

The notation \( \delta \), rather than the previously used \( \alpha \), is introduced here to avoid confusion with the parameter \( \alpha \) in the equation of the regression line. The quantity \( t_{1-\delta/2}(n-2) \) is the 1 − \( \delta \)/2 quantile of the \( t(n-2) \) distribution. Just like some of the confidence intervals in Unit 1, this interval is of the form \( \hat{\beta} \pm \) multiple of \( \text{s.e.}(\hat{\beta}) \), where \( \text{s.e.}(\hat{\beta}) \) is the estimated standard error of \( \hat{\beta} \). The interval for \( \alpha \) is also of the form \( \hat{\alpha} \pm \) multiple of \( \text{s.e.}(\hat{\alpha}) \), and in fact the multiplier is exactly the same \( t \) percentage point as for the confidence interval for \( \beta \). But, to calculate the confidence intervals in GenStat, the detailed formulae are not

**Hint:** For the last question, you might like to compare the logs of the (absolute values of the) test statistics.
Unit 3  Simple linear regression

required. GenStat gives us the estimates $\hat{\alpha}$ and $\hat{\beta}$, together with their estimated standard errors. Thus GenStat can be used to get all the information necessary to easily work out the confidence intervals using a calculator. In the next exercise we shall see how this works out for the abrasion loss data.

Exercise 3.12  Calculating confidence intervals for regression coefficients

Load rubber.gsh into GenStat again, if necessary, and fit a straight line to these data, using loss as the response variable and hardness as the explanatory variable.

(a) Write down the values of $\hat{\alpha}$, $\hat{\beta}$, s.e.$(\hat{\alpha})$ and s.e.$(\hat{\beta})$.

(b) We also need the relevant quantile of the $t(28)$ distribution. The degrees of freedom, 28 in this case, is given in the Summary of analysis table. Using the Data|Probability Calculations menu item, obtain the value of $t_{0.025}(28)$.

(c) Using your calculator, work out the 95% confidence intervals for $\alpha$ and $\beta$.

Exercise 3.13  Modelling temperature difference and gas consumption

The data whose first five values are in Table 3.6 were collected by the Open University’s Energy Research Group in the early 1980s. For each of 15 houses of similar design in the Neath Hill district of Milton Keynes, they recorded over a period of time the average temperature difference (in °C) between the inside and the outside of the house, and the average daily gas consumption (in kWh). The aim was to investigate how these quantities were related in real lived-in houses.

<table>
<thead>
<tr>
<th>Daily gas consumption (kWh)</th>
<th>Temperature difference (°C)</th>
</tr>
</thead>
<tbody>
<tr>
<td>69.81</td>
<td>10.3</td>
</tr>
<tr>
<td>82.75</td>
<td>11.4</td>
</tr>
<tr>
<td>81.75</td>
<td>11.5</td>
</tr>
<tr>
<td>80.38</td>
<td>12.5</td>
</tr>
<tr>
<td>85.89</td>
<td>13.1</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

According to physical theory, the heat energy needed to sustain a particular temperature difference is proportional to the temperature difference. Thus, if the only energy input into the houses came from the gas, the theory predicts that gas consumption ($y$, variate gascons) against temperature difference ($x$, variate tempdiff) should be well fitted by a regression line through the origin. Having temperature difference as the explanatory variable is justified by the fact that the occupants had (partial) control over temperature by setting the thermostat; interest centres on how this control affects gas consumption.

The data were collected in winter, and the houses had gas central heating, so gas is certainly the predominant energy input. However, gas is not the only energy input; for instance, there was energy from electricity and from the body heat generated by the inhabitants. Also, the data give the energy content of the gas supplied to the house, and not all this energy is converted into heat within the house, mainly because the efficiency of the central heating boiler in turning chemical energy in gas to heat energy is less than 100% and because the gas also provides hot water.

(a) Produce a scatterplot of the data. Does it indicate that a regression line constrained to pass through the origin would provide a reasonable model?

(b) Fit an unconstrained straight line to the data. Calculate a 95% confidence interval for the intercept of this line. Comment on the appropriateness of calculating such a confidence interval for these data. What does the 95% confidence interval tell you about the wisdom of trying to fit a straight line through the origin to these data?

There are also of ways of doing the whole calculation in GenStat, but we will not cover them in M346.

Using GenStat to find quantiles of the $t$-distribution was done in Exercise 2.4.

Dataset name: temperat.
Section 3.3  
Confidence intervals and prediction

Very often the main aim of fitting a regression line to data is to produce an equation that can be used to predict new responses from given values of the explanatory variable. If all that is required is a point estimate of the response, things are very easy; one merely substitutes the given value of the explanatory variable into the fitted regression equation. That is, if \(x_0\) is the value of the explanatory variable for an individual whose response \(Y_0\) is not known, the obvious predictor of \(Y_0\) is

\[
\hat{Y}_0 = \hat{\alpha} + \hat{\beta}x_0.
\]

But this does not take account of two further issues. First, the true regression line is not known; we merely have an estimate for it. Second, even if the true line were indeed somehow known, then all we would have is a value for the mean of the response variable, as in (3.1). We do not have a value for the random quantity that defines how far the actual value of \(Y\) will be from its mean, so it would still not be possible to predict the response exactly.

One way of taking into account these uncertainties is to calculate interval estimates. The first interval estimate we shall consider is a confidence interval for the mean of the response variable \(\hat{Y}_0\) at a new value \(x_0\) of the explanatory variable. Such a \(100(1 - \delta)\%\) interval is given by

\[
\hat{\alpha} + \hat{\beta}x_0 \pm [t_{1-\delta/2}(n-2)] s \sqrt{\frac{(x_0 - \bar{x})^2}{s_{xx}} + \frac{1}{n}}.
\]

(3.7)

This is in fact again of the form estimate \(\pm\) multiple of s.e.(estimate), and the multiplier is the same \(t\) percentage point as in the confidence intervals for \(\alpha\) and \(\beta\). Note that \(\bar{x}\) and \(s_{xx}\) depend on \(x_1, x_2, \ldots, x_n\) only, and not on \(x_0\). Also, for values of \(x_0\) close to the mean \(\bar{x}\), the value of \((x_0 - \bar{x})^2\) is small and the interval will be relatively narrow. For \(x_0\) a long way from \(\bar{x}\), the interval will be relatively wide.

The other type of interval we shall consider is a prediction interval for the response \(Y_0\) at a new value \(x_0\) of the explanatory variable. The confidence interval for the mean of \(Y_0\) takes into account the sampling variability inherent in the estimation of the regression line, so that it gives a range of plausible values for the mean of \(Y_0\), but does not take into account the fact that \(Y_0\) varies about its mean value. A prediction interval takes into account both types of variability, so that it provides a range of plausible values for the value of \(Y_0\) that will actually be observed. The prediction interval thus has to be wider than the confidence interval for the mean of \(Y_0\). The expression for the prediction interval is

\[
\hat{\alpha} + \hat{\beta}x_0 \pm [t_{1-\delta/2}(n-2)] s \sqrt{\frac{(x_0 - \bar{x})^2}{s_{xx}} + \frac{1}{n} + 1}.
\]

(3.8)

Again, prediction intervals are (relatively) narrower for values of \(x_0\) near the mean \(\bar{x}\).

There are corresponding expressions for the confidence interval for the mean and for the prediction interval in the case of regression lines constrained to pass through the origin, but they are not given here.

Note that the intercept is the mean value of the response variable when the explanatory variable takes the value 0. Thus (3.7), with \(x_0 = 0\), gives the confidence interval for \(\hat{\alpha}\) that you calculated in Exercises 3.12 and 3.13.

Formulae (3.7) and (3.8) are intuitively sensible, reflecting the facts that estimation and prediction close to the bulk of the data are reasonable things to do, but estimation and prediction beyond the range of the data are much less reliable. In fact, the latter type of estimation and prediction, commonly referred to as extrapolation, additionally puts great trust in the continued suitability of the model away from the data. Although these formulae look rather complicated, in practice GenStat will calculate the worst parts of them for you.
Exercise 3.14  Calculating confidence intervals for the mean and prediction intervals

(a) In this part of the exercise you will use the regression line for the data on contaminated peanuts in Table 3.4 (Exercise 3.6) to predict the mean percentage of non-contaminated peanuts in batches for which the aflatoxin level in a sample is 13.2 parts per billion, in the form of a point estimate and also a 90% confidence interval.

Forming a confidence interval for the mean response is very similar to forming confidence intervals for the slope and intercept. We have to obtain the appropriate t value, the estimated means and their standard errors, and then use a calculator as in Exercise 3.12.

First, if necessary, load peanuts.gsh into GenStat again. Fit a straight line to these data, using percent as the response variable and toxin as the explanatory variable.

Return to the Linear Regression dialogue box. Click on Predict. In the Predictions - Simple Linear Regression dialogue box, enter 13.2, the value of our explanatory variable toxin for which we want the point estimate and confidence interval, in the Predict Values at field. Check that the Predictions and Standard Errors options are selected, but that Include variance of future observation is not selected, before clicking on Run.

In the Output window, the point estimate for the mean response is given as the value of Prediction, 99.96 in this case. The corresponding standard error (s.e.) for this point estimate is given as 0.008681.

Now work out the 90% confidence interval for the mean response, using the same basic method as in Exercise 3.12.

(b) A different standard error is required to calculate a 90% prediction interval for the percentage of non-contaminated peanuts in batches for which the aflatoxin level in a sample is 13.2 parts per billion. Return to the Predictions - Simple Linear Regression dialogue box. Again enter 13.2 in the Predict Values at field and ensure that the Predictions and Standard Errors options are selected. However, this time also select the Include variance of future observation option. This will ensure that the standard error produced by GenStat reflects the sampling variability both of the regression line and of \( Y_0 \) about its mean value. Click on Run.

Notice that the point estimate (the value Prediction in the Output window) has not changed. However, the standard error of this point estimate has increased to 0.04027.

So what is the 90% prediction interval that you were asked for?

---

### 3.4 Checking the assumptions

This section is concerned with only one class of methods for checking regression assumptions, but it is a very useful class. You will probably be familiar from your previous studies with the idea of using residuals to investigate how well a regression model fits. Recall that residuals are the (observed) differences between the observed values of the response variable, \( y_i \), and the corresponding fitted values, \( \hat{\alpha} + \hat{\beta} x_i \), predicted by the regression line, that is,

\[
    r_i = y_i - (\hat{\alpha} + \hat{\beta} x_i), \quad i = 1, 2, \ldots, n.
\]  

(3.9)
Section 3.4  Checking the assumptions

(This is the basic idea, which will be developed somewhat later in the section.) As mentioned in Section 3.2, the residuals can be thought of as estimates of the random errors $\epsilon_i$ in the regression model. Plotting the residuals in various ways can provide useful checks on the assumptions about the random errors. Two types of residual plot that are especially useful in most simple linear regression situations are the following.

First, plotting the residuals against the explanatory variable or against the fitted values can indicate whether there is some aspect of the relationship between the explanatory variable and the response variable that has not been taken account of by the straight-line model. This sort of plot can also show up problems with the assumption of constant variance. Since the fitted values are linear functions of the values of the explanatory variable, a plot of residuals against fitted values looks the same as a plot of residuals against the explanatory variable, apart from a change of scale on the horizontal axis. (The horizontal scale will be reversed if $\hat{\beta}$ is negative.) In this course we will plot residuals against fitted values rather than against the explanatory variable because this sort of plot is easier to extend to situations where there are several explanatory variables.

Second, producing a normal probability plot of the residuals can give a guide as to whether the assumption of normality of the $\epsilon_i$ is appropriate.

Each of these plots can also highlight abnormally large residuals, which can draw attention to potential outliers in the data.

Other types of residual plot can also be useful in some contexts. A third such plot, useful when the data have been collected in time order, is to plot the residuals in order of time. Such a plot can provide evidence of trends in time in the residuals. Such trends can be problematic, because they indicate that the $\epsilon_i$ are not independent.

A fourth type of residual plot is to plot the residuals against the values of some other potential explanatory variable apart from the one you have used in fitting the line. This can indicate that you need a more complicated model that takes into account the new explanatory variable as well.

You should make it a habit to look routinely at the first two types of residual plot whenever you fit a regression line. The third and fourth types are not always appropriate, but can be useful when they are. In later parts of the course, notably Unit 10, you will meet other ways of checking the assumptions behind regression data, some of them also based on residuals.

Exercise 3.15  Producing individual residual plots

Load the rubber.gsh data into GenStat again, if necessary, and once more carry out a simple linear regression with loss as the response variate and hardness as the explanatory variate. Make sure the Linear Regression dialogue box is active.

(a)  To produce a plot of residuals against fitted values, click on the Further Output button. A dialogue box will appear: click on the Model Checking button. Another dialogue box, Model Checking, will appear. Change Type of Graph from Composite to Fitted Values; change Type of Residual from Deviance to Simple; click on the Run button. ‘Simple’ residuals are the type defined in Equation (3.9). In the resulting plot, the straight red line across the middle shows where zero comes on the vertical axis, and the curved blue line gives an indication of the general pattern of the residuals. What does the plot tell you about the appropriateness of the linear regression model?

(b)  Repeat the procedure that gets you to the Model Checking dialogue box and select simple residuals again. Now change Type of Graph from Composite to Normal; click on the Run button. What sort of diagram is produced, and what does it tell you about the appropriateness of the model?

If possible, leave the the Linear Regression dialogue box open for the next exercise.

The default Type of Residual in the Model Checking dialogue box is standardised deviance residuals (Deviance). More details about standardised deviance residuals will be given in Unit 10. However, for now the important thing is the fact that
these deviance residuals are standardised. For simple linear regression and all the
other linear models based on the normal distribution covered in Units 4 to 7, the
standardised deviance residuals are equivalent to dividing simple residuals by
their estimated standard error.

The idea with standardising residuals is to ‘calibrate’ the vertical scale on residual
plots so that the plotted residuals, although not independent, individually have
approximately a standard normal distribution. It may seem counter-intuitive that
the residuals (standardised or simple) are not independent of each other, since
they are estimates of the random terms \( \epsilon_i \) in model (3.2) and the \( \epsilon_i \) are
independent. But the residuals have to sum to zero (as can be shown for simple
residuals from the formulae for the parameter estimates, Equations (3.3)
and (3.4)). Therefore, if you know all but one of them, you can work out the
missing one. Hence they cannot be independent. In future, we shall generally look
at standardised deviance residuals, and you should assume from here onwards
that all linear regression residuals are standardised deviance residuals, unless
stated otherwise.

The default Type of Graph in the Model Checking dialogue box is Composite. This
produces a window with four different graphs in it: the two types you produced in
Exercise 3.15, a histogram of the residuals, and a half-normal plot (which will be
used later in the course). A half-normal plot just consists of the negative half of
the normal probability plot superimposed on the positive half; it thus assesses
absolute values of residuals rather than residuals themselves.

**Exercise 3.16 Producing a composite residual plot**

If necessary, load rubber.gsh into GenStat and refit the simple linear regression
with loss as the response variable and hardness as the explanatory variable.

Obtain the Model Checking dialogue box again, and this time leave all the items
at their default values to produce a composite graph of standardised deviance
residuals. Does it tell you anything about the model you did not already know?

Note that if you fit two or more different regression models during a GenStat
session, then the residual plots will correspond to the most recent regression. If
the residuals you want to plot are not from the most recent regression, then you
have to repeat the appropriate regression.

**Exercise 3.17 Checking the appropriateness of a simple linear regression**

Produce a composite plot of the (standardised deviance) residuals for the data on
peanut contamination in Table 3.4. Comment on the appropriateness of the
model.

**Exercise 3.18 Modelling the iron content of slag**

The iron content of crushed blast-furnace slag can be determined by an accurate
chemical test at a laboratory or estimated by a cheaper, quicker, magnetic test.
Data were collected to investigate the extent to which the results of the chemical
test of iron content could be predicted from the magnetic test. The observations,
the first five of which are shown in Table 3.7, are given in order of the time they
were made.

**Table 3.7 Chemical and magnetic determinations of iron content**

<table>
<thead>
<tr>
<th>Chemical</th>
<th>Magnetic</th>
</tr>
</thead>
<tbody>
<tr>
<td>24</td>
<td>25</td>
</tr>
<tr>
<td>16</td>
<td>22</td>
</tr>
<tr>
<td>24</td>
<td>17</td>
</tr>
<tr>
<td>18</td>
<td>21</td>
</tr>
<tr>
<td>18</td>
<td>20</td>
</tr>
</tbody>
</table>

Dataset name: iron.
Scientific Press/McGraw-Hill.
### Section 3.4 Checking the assumptions

(a) Produce a scatterplot of these data, treating the magnetic test value as the 
explanatory variable and the chemical test value as the response variable. 
Describe how the two variables appear to be related. Fit a regression line. 
Produce the ‘default’ composite residual plot.

(b) Investigate whether there is a change in the relationship between the two 
measurements over time by plotting the residuals in time order. This can be 
done in GenStat by obtaining the Model Checking dialogue box and choosing 
Index as the Type of Graph. The resulting index plot plots each residual 
against the corresponding row number in the dataset. It is thus a plot of \( r_i \) 
(or to be precise the deviance version of \( r_i \)) against \( i \). In this case, since the 
data are given in time order, this plots the residuals against time.

(c) Comment on the appropriateness of the simple linear regression model for 
these data.

#### Exercise 3.19 Modelling ice cream consumption

Ice cream consumption (in pints per capita, variate pints) was measured over 30 
successive four-week periods from 18 March 1951 to 11 July 1953. One of the 
variables thought to influence consumption was the mean temperature 
(°F, variate temp). (This dataset contains values for other variables, which you 
should ignore.) The first five of the data values, which are given in time order, are 
shown in Table 3.8.

**Table 3.8** Ice cream consumption and temperature for a succession of four-week periods

<table>
<thead>
<tr>
<th>Consumption (pints per capita)</th>
<th>Temperature (°F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.386</td>
<td>41</td>
</tr>
<tr>
<td>0.374</td>
<td>56</td>
</tr>
<tr>
<td>0.393</td>
<td>63</td>
</tr>
<tr>
<td>0.425</td>
<td>68</td>
</tr>
<tr>
<td>0.406</td>
<td>69</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Produce a scatterplot of these data and fit a regression line. Produce the default 
composite residual plot, and produce a plot of the residuals in time order. 
Comment on the appropriateness of the simple linear regression model for these 
data.

In the above exercise, GenStat flagged a residual as corresponding to a possible 
outlier. It uses a rather complicated procedure, based on the standardised 
deviance residuals, to decide which points to flag.

Now, each of these residuals follows, approximately, a standard normal 
distribution. So a threshold based on the standard normal distribution is used to 
decide whether a residual is sufficiently unusual to be considered a possible 
outlier. When \( \text{d.f.}(\text{Residual}) \leq 20 \), a threshold of 2 is used. That is, residuals that 
are either smaller than \(-2\) or greater than \(2\) are flagged. This is based on the fact that 
the central 95% of the probability mass of the standard normal distribution 
lies in \((-1.96, 1.96)\), and 1.96 is approximately 2. For larger values of 
\( \text{d.f.}(\text{Residual}) \), higher thresholds are used. This increase in the threshold is done to 
limit the number of residuals flagged as possible outliers in situations where there 
is in fact no problem with the fit of the model.

#### Exercise 3.20 Checking residuals against additional variables

In the experiment on abrasion loss in rubber, more data were collected than were 
shown in Table 3.1. The tensile strength of the rubber (in kg/cm²) was also 
recorded, and the overall aim was to investigate how abrasion loss was determined 
by both hardness and tensile strength. The first five values of the full rubber 
dataset are shown in Table 3.9.
Table 3.9 Abrasion loss, hardness and tensile strength of samples of rubber

<table>
<thead>
<tr>
<th>Abrasion loss (g/h)</th>
<th>Hardness (degrees Shore)</th>
<th>Tensile strength (kg/cm²)</th>
</tr>
</thead>
<tbody>
<tr>
<td>372</td>
<td>45</td>
<td>162</td>
</tr>
<tr>
<td>206</td>
<td>55</td>
<td>233</td>
</tr>
<tr>
<td>175</td>
<td>61</td>
<td>232</td>
</tr>
<tr>
<td>154</td>
<td>66</td>
<td>231</td>
</tr>
<tr>
<td>136</td>
<td>71</td>
<td>231</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Fit (yet again) the regression line of abrasion loss against hardness. Obtain the Model Checking dialogue box. Select Index as Type of graph. Enter strength in the Index Variate field. By doing this the (deviance) residuals will be plotted against tensile strength (variate strength). What do you find?

### 3.5 Transformations

In the previous section you saw several examples of data for which the simple linear regression model is not appropriate. This raises the question of how else one might model the data. Answers to this question will be covered in several places in the rest of this course. This section discusses an approach that you have already met — transforming the data. The idea here is that, if the simple linear regression model is not appropriate to model the relationship between two variables $y$ and $x$, it might be possible to find a transformation of one or both of $y$ and $x$ such that the relationship between the transformed variables is well fitted by the simple linear regression model.

To summarise very crudely, there are broadly two approaches to finding appropriate transformations. One approach is to use insights from known scientific (or other) theory or from background knowledge that describe how the quantities involved might be expected to be related. Often such theory or knowledge can suggest appropriate transformations that can lead to a linear relationship. Another approach is simply to look at plots of the data and choose transformations that appear to lead to straight-line relationships with the right properties. This empirical approach often involves a certain amount of trial and error before an appropriate transformation is found.

In this unit we confine ourselves to transformations on the ladder of powers (including the logarithmic transformation; see Subsection 1.2.3). (Of course, these are not the only transformations that can be used.) One consequence of restricting ourselves to power transformations is that they will not help us to analyse data like those shown in Figure 3.1 (which seems to indicate a slight upturn in the mean towards the left-hand end). The reason is that all the power transformations (including log) are monotonic (when applied to positive data) — i.e., they are either increasing or decreasing. For instance, the graph of $x^a$ against $x$, where $a$ is a positive constant, is increasing. (It slopes upwards from left to right for all positive $x$.) The same is true of log $x$. If $a$ is negative, the graph of $x^a$ against $x$ is decreasing. In no case does the graph increase and then decrease (or vice versa).
Section 3.5 Transformations

Example 3.2 Studying memory retention

Table 3.10 contains the first few points of a famous dataset on memory retention, collected in an experiment by a psychologist named Strong. Subjects memorised a list of disconnected items, and then average percentage memory retention ($p$) was measured against passing time ($t$, measured in minutes). The measurements were taken five times during the first hour after the items were memorised, and then at various times up to a week later.

**Table 3.10** Average percentage memory retention ($p$) over time ($t$)

<table>
<thead>
<tr>
<th>$p$</th>
<th>$t$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.84</td>
<td>1</td>
</tr>
<tr>
<td>0.71</td>
<td>5</td>
</tr>
<tr>
<td>0.61</td>
<td>15</td>
</tr>
<tr>
<td>0.56</td>
<td>30</td>
</tr>
<tr>
<td>0.54</td>
<td>60</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

A scatterplot of these data is shown in Figure 3.3.

![Figure 3.3 Scatterplot of average memory retention against time](image)

The relationship between the variables is obviously not linear. However, it is monotonic, in that $p$ generally decreases as $t$ increases. Thus it may be possible to ‘straighten out’ the plot by transforming one or both of the variables, using a transformation from the ladder of powers. In this case, psychological theory suggested that a model of the form $p = C \exp(-\beta t)$ might be appropriate, where $C$ and $\beta$ are unknown positive constants. Taking logs of both sides of this equation leads to

$$\log p = \log C - \beta t,$$

which is a linear equation relating $\log p$ to $t$. If this model fits, then a plot of $\log p$ against $t$ should be linear. Figure 3.4 shows the resulting scatterplot.
Figure 3.4 Log(average memory retention) against time

So much for the theory in this case! The plot is clearly still nowhere near straight. But what about the empirical approach of finding an appropriate power (i.e. trying out some transformations and seeing what happens)? You are asked to try this approach in the next exercise. However, before you attempt it, note that, when trying to straighten out a relationship, it is often good practice to concentrate on transforming the explanatory variable rather than the response variable.

Exercise 3.21 Seeking transformations to straighten relationships

Look back at the original scatterplot of $p$ against $t$ (Figure 3.3). A transformation of $t$ can move points to the right or left relative to one another in the scatterplot. In this case, we need to bring the points at the lower right of the scatterplot back towards the others. In other words, we need to bring the high values of $t$ closer together compared with the low values. This can be achieved by moving down the ladder of powers (see Subsection 1.2.3). Load the file strong.gsh into GenStat, and find an appropriate transformation for these data, using the Transformation dialogue box you met in Exercise 2.14. Power and log transformations appear as options in this dialogue box. Fit a straight line to the transformed data, and check the appropriateness of the simple linear regression model by producing residual plots.

Finding an appropriate transformation for the memory retention data was reasonably straightforward because there was only one major problem in the original scatterplot: the curvature. For the data in the next exercise, there are more things to take account of.

Exercise 3.22 Modelling the tensile strength of cement

The tensile strength of cement depends on (among other things) the length of time for which the cement is dried or ‘cured’. In an experiment, different batches of cement were tested for tensile strength after different curing times. The data are given in Table 3.11.
Table 3.11  Tensile strength of cement for different curing times

<table>
<thead>
<tr>
<th>Curing time (days)</th>
<th>Tensile strength (kg/cm²)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>13.0 13.3 11.8</td>
</tr>
<tr>
<td>2</td>
<td>21.9 24.5 24.7</td>
</tr>
<tr>
<td>3</td>
<td>29.8 28.0 24.1 24.2 26.2</td>
</tr>
<tr>
<td>7</td>
<td>32.4 30.4 34.5 33.1 35.7</td>
</tr>
<tr>
<td>28</td>
<td>41.8 42.6 40.3 35.7 37.3</td>
</tr>
</tbody>
</table>

Produce a scatterplot of these data, treating curing time (variates curetime) as the explanatory variable. Comment on how appropriate the simple linear regression model would be for these data.

Can problems with the model like those you found in Exercise 3.22 be dealt with by transformation? It is often a good strategy in cases like this to deal with the non-constant variance first. The problem is with the variance in the y direction, the direction of the response variable, so a transformation of the response variable may correct things. The aim in this case is to reduce the spread of the higher values of the tensile strength variable compared with the lower values. Such a transformation is one towards the left on the ladder of powers — something like a square root, a logarithm or a reciprocal (i.e. power of $-1$).

Exercise 3.23  Transforming to stabilise variance

For the data in Table 3.11, try, by trial and error, to find an appropriate transformation of the tensile strength that produces a constant variance. (Such a transformation will, typically, also help with the normality of the responses.)

The problem after transforming the response variable is that the relationship is even less linear than it was originally (compare the scatterplots in Solutions 3.22 and 3.23). A transformation of the explanatory variable (curing time) may, however, straighten it out. Again the requirement is to reduce the spread of the higher values of curing time compared with the lower values, so again we want to move down the ladder of powers.

Exercise 3.24  Analysing log(cement strength)

(a) Find an appropriate transformation of curing time that makes a plot of log(tensile strength) against transformed curing time reasonably linear.

(b) Fit a regression line with log(tensile strength) as the response variable and curing time transformed as in part (a) as the explanatory variable. Produce residual plots to investigate the fit of the model.

Exercise 3.25  Output of a wind generator

The data underlying Table 3.12 were obtained in a study of a wind generator; they record the direct current (DC) output at different wind speeds.

Table 3.12  Output of a wind generator for various wind speeds

<table>
<thead>
<tr>
<th>DC output (miles per hour)</th>
<th>Wind speed</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.123</td>
<td>2.45</td>
</tr>
<tr>
<td>0.500</td>
<td>2.70</td>
</tr>
<tr>
<td>0.653</td>
<td>2.90</td>
</tr>
<tr>
<td>0.558</td>
<td>3.05</td>
</tr>
<tr>
<td>1.057</td>
<td>3.40</td>
</tr>
</tbody>
</table>

Dataset name: cemstren.

Transformation is not the only way of dealing with non-constant variance. A method called weighted least squares provides one alternative. Another alternative, using a non-normal error distribution, is dealt with in Units 8 and 9.

Dataset name: wind.
Unit 3  Simple linear regression

(a) Produce a scatterplot of these data with DC output, \( y \), as the response variable and wind speed, \( x \), as the explanatory variable. Hence comment on the relationship between the DC output and the wind speed.

(b) Decide the appropriate direction on the ladder of powers in which you might transform \( y \) to straighten the curve in the plot. Try to find an appropriate transformation, and plot the transformed data. Are the assumptions of simple linear regression appropriate for the transformed data?

(c) Repeat part (b), but this time leave \( y \) untransformed and transform \( x \).

(d) Choose the most appropriate transformation from those you investigated in parts (b) and (c). Fit a regression line to the transformed data, and check the model using appropriate residual plots.

Exercise 3.26  Estimating Janka hardness

Janka hardness is an important structural property of timber, but is difficult to measure directly. However, it is related to the density of the timber, which is comparatively easy to measure, and therefore it would be useful to fit a model allowing Janka hardness to be predicted from the density. The Janka hardness and density of 36 Australian eucalypt hardwoods form the dataset underlying Table 3.13.

Table 3.13  Janka hardness of different densities of Australian eucalypt hardwoods

<table>
<thead>
<tr>
<th>Janka hardness</th>
<th>Density</th>
</tr>
</thead>
<tbody>
<tr>
<td>484</td>
<td>24.7</td>
</tr>
<tr>
<td>427</td>
<td>24.8</td>
</tr>
<tr>
<td>413</td>
<td>27.3</td>
</tr>
<tr>
<td>517</td>
<td>28.4</td>
</tr>
<tr>
<td>549</td>
<td>28.4</td>
</tr>
<tr>
<td>..</td>
<td>..</td>
</tr>
</tbody>
</table>

Produce a scatterplot of these data with density as the explanatory variable. Explain why simple linear regression might be inappropriate for these data without transformation. Carry out appropriate transformations so that a simple linear regression model does fit the transformed data reasonably well. Produce residual plots to investigate how well the model fits.

Notice that, as in Solution 3.26, a conclusion that there is nothing amiss with (some of) the flagged residuals is often to be expected, since we should expect the occasional datapoint to be flagged anyway.

We have dwelt on transformations at some length, because the ability to transform data appropriately is a very important tool in the statistical modeller’s toolbox. Two aspects have been emphasised: the desire to make the variance constant and the need to straighten out the mean. Transformations in the \( y \) and \( x \) directions, respectively, are useful ways of coping with each of these in turn (if indeed transformations are powerful enough to produce a linear regression model for the data). Note that transformations of \( y \) and/or \( x \) retain the linearity in the parameters of the regression equation.
3.6 Comparing slopes

The next topic covered in this unit is a method of carrying out an approximate hypothesis test to compare the slopes of two different regression lines. This topic is useful for two main reasons. In some examples, there are important questions whose answers depend on whether different regression lines have the same slope, and as you will see, these questions can be answered using simple extensions of statistical ideas you have already met. Furthermore, developing these ideas will begin to illustrate how the basic regression models covered in this unit can be extended to the more complicated models covered in the rest of this course.

Example 3.3 Impact of calcium concentration and location on mortality

The data in Table 3.14 were collected in an investigation of environmental causes of disease. They comprise the annual mortality rate per 100,000 for males, averaged over the years 1958–64, and the calcium concentration (in parts per million) in the drinking water supply for 61 large towns in England and Wales. (The higher the calcium concentration, the harder the water.) The data also identify which towns are at least as far north as Derby.

Table 3.14 Mortality rate and water calcium concentration in towns in England and Wales

<table>
<thead>
<tr>
<th>Mortality rate (per 100,000)</th>
<th>Calcium concentration (parts per million)</th>
<th>North? (1 is yes)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1247</td>
<td>105</td>
<td>0</td>
</tr>
<tr>
<td>1668</td>
<td>17</td>
<td>1</td>
</tr>
<tr>
<td>1466</td>
<td>5</td>
<td>0</td>
</tr>
<tr>
<td>1800</td>
<td>14</td>
<td>1</td>
</tr>
<tr>
<td>1609</td>
<td>18</td>
<td>1</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

A scatterplot of mortality rate against calcium concentration is shown in Figure 3.5.

![Figure 3.5](image)
Unit 3 Simple linear regression

There is a clear relationship between mortality rate and calcium concentration, with lower mortality rates in areas with harder water, on average. But many of the areas with soft water in England and Wales are in the north, so one possible explanation for these data is that the mortality rate is higher in the north for reasons that have nothing to do with the water supply.

We can investigate this further by identifying on the plot those points that correspond to towns in the north, as in Figure 3.6, where the towns in the north are shown by crosses.

**Figure 3.6** Mortality rate for different water calcium concentrations in northern and southern towns

This shows that in fact there is a range of calcium concentrations in the north and in the south. The mortality rate is generally rather higher in northern towns; but within each group of towns, the mortality rate is lower in towns with harder water.

**Exercise 3.27** Fitting separate regression lines

Load the dataset `water.gsh` into GenStat. Fit a regression line to the data on mortality rate and calcium concentration for the northern towns only.

One way to do this is as follows. Make sure the Spreadsheet window is active. Choose the `Spread|Restrict/Filter|To Groups (factor levels)` menu item. In the resulting dialogue box, choose north in the Factor field. Set the Restriction Type to Include. Select 1 in the Selected Levels box. This all tells GenStat to include in its analyses only those rows of the spreadsheet where north = 1, i.e. just the northern towns. Click on OK. The rows corresponding to southern towns disappear from the spreadsheet. (You can get them back, if you need to, using the `Spread|Restrict/Filter|Remove All` menu item.) Now GenStat will act as if the southern towns were not in the dataset, and you can fit the regression line in the usual way. Note the estimated standard error of the slope of the regression line.

Now go through the same procedure, but this time using the data for the southern towns only. When doing this, in the `Restrict Units on Factor north` dialogue box, make sure you select Replace with New as the Existing Restrictions option. (Otherwise you will be selecting only those towns that are in the north and the south, that is, no towns at all!)
The regression lines from Exercise 3.27 are shown on the scatterplot in Figure 3.7.

Figure 3.7  Separate regressions of water calcium concentrations on mortality rate

Perhaps the most obvious feature of these lines is that they are close to being parallel. In other words, they appear to have almost the same slope. The question then arises of whether the slopes differ significantly, or whether the small difference between them could be attributed to sampling error.

A hypothesis test can be constructed as follows. An obvious estimate of the true difference between the slopes is the difference between the slopes of the fitted lines, which is \((-1.931) - (-2.093) = 0.162\). Since the two slope estimators are calculated from independent datasets, they are independent. Each of the slope estimators has a normal distribution, so their difference has a normal distribution too. Using results in Subsection 1.2.4, under the null hypothesis of equal slopes, the mean of the difference between the slopes of the fitted lines is zero. The variance of the difference is the sum of the variances of the two slopes; and though these are not known exactly, we have sample estimates for them, namely the squares of their estimated standard errors. Thus the estimated variance of the difference in slopes is \(0.848^2 + 0.566^2\), or 1.03946. The estimated standard deviation of the difference between slopes is then \(\sqrt{1.03946}\), or 1.0195.

Exercise 3.28  Comparing slopes

Without doing any calculations, what do you think the outcome of this test is likely to be? In what way is the information given insufficient for you to calculate the \(p\) value precisely?

This hypothesis test, as presented here, is indeed a little crude and approximate; however, to improve it we need to make a further assumption about the data, namely that the error variances are the same in both groups. The technique for performing the test in the light of this further assumption will be covered in later units.
3.7 Looking forward: extensions to simple linear regression

Much of this course extends simple linear regression by adding more explanatory variables. To model many datasets fully, one would need a model with more than one explanatory variable. For instance, in Exercise 3.20 there was a single response variable, abrasion loss \( (Y) \), and two explanatory variables, hardness \( (x) \) and tensile strength \( (z) \). The most straightforward model for data like these is

\[
Y_i = \alpha + \beta_1 x_i + \beta_2 z_i + \epsilon_i, \quad i = 1, 2, \ldots, n,
\]

where \( \alpha, \beta_1 \) and \( \beta_2 \) are unknown constants and the \( \epsilon_i \) are normally distributed random variables with the same variance \( \sigma^2 \), the \( \epsilon_i \) being assumed independent. If there are more than two explanatory variables, the model is extended by adding on more terms in the obvious way. The process of fitting models like this is called multiple regression and is the topic of Unit 5.

The other main extension of simple linear regression in the earlier part of the course is to categorical explanatory variables. All the explanatory variables in the examples in this unit have been quantitative. But there is nothing in the regression model to say that this has to be the case. Regression with categorical explanatory variables is a way of solving a surprisingly large range of problems, as you will see. One example is the following.

In Example 3.3, you saw that a plausible model was that, in both the south and the north, the mortality rate, \( Y \), is linearly related to calcium concentration, \( x \), and that both regression lines have the same slope, but their intercepts are different. Using the approach in Section 3.6 we can therefore deal with these data by using two separate, independent models:

\[
E(Y) = \alpha_S + \beta x \quad \text{for the south};
\]
\[
E(Y) = \alpha_N + \beta x \quad \text{for the north};
\]

where \( \alpha_S, \alpha_N \) and \( \beta \) are three parameters to be estimated from the data. But suppose we are particularly interested in the difference, \( \delta \) say, between the intercepts. Since the regression lines are parallel, this quantity represents the vertical distance between them; in other words, it is the excess of the mortality rate in the north compared with the south, when comparing two towns with the same calcium level in the drinking water supply. We could write the models as:

\[
E(Y) = \alpha_S + \beta x \quad \text{for the south};
\]
\[
E(Y) = \alpha_S + \delta + \beta x \quad \text{for the north};
\]

where this time the three parameters to be estimated are \( \alpha_S, \delta \) and \( \beta \).

Now define a new variable \( z \) which takes two possible values, 0 for towns in the south and 1 for towns in the north. We can now combine the two models into one:

\[
E(Y) = \alpha_S + \delta z + \beta x.
\]

This single model incorporates both of the separate models above; and it has exactly the same form as the multiple regression model that was described for the abrasion loss data. Thus multiple regression methods with categorical explanatory variables can be used to answer questions about parameters such as differences between regression intercepts. You will see several examples of this general approach in the rest of the course.

The standard multiple regression models, like those for simple linear regression, involve an assumption that the random terms in the model are normally distributed. In later units of the course, you will see how this assumed distribution can be replaced by different ones, such as Poisson or Bernoulli, in order to accommodate responses that are in the form of counts or are binary.
3.8 Correlation

Regression models are concerned with explanatory relationships in which values of one variable, \( y \), are considered to be dependent on the values of one, or more, explanatory variables, \( x \). That is, there is a directionality in the relationship, ‘from \( x \) to \( y \)’. However, there need be no such directionality in a relationship. In many cases, it is more appropriate to consider \( x \) and \( y \) to be on an equal footing: certainly, there is some kind of relationship between \( x \) and \( y \) (for instance, high values of \( x \) may tend to go with high values of \( y \)), but it is not sensible to think of \( y \) as being ‘driven’ by \( x \) or indeed of \( x \) as being driven by \( y \).

As well as relationships of this sort between observed data values, similar relationships often exist between quantities calculated from data. For instance, in Section 3.2 it was pointed out that \( \hat{\alpha} \) and \( \hat{\beta} \) are (in general) dependent, and in later units you will see that residuals are related to each other.

**Example 3.4 Renal clearance in heart patients**

Digoxin is a drug that is eliminated in the urine largely unchanged. Table 3.15 gives the first five data pairs from a dataset of measurements, for each of 35 heart patients, of the ‘renal clearance’ of both digoxin and the naturally occurring substance creatinine. Each variable has been logged, the original units being ml/min/1.73 m\(^2\).

<table>
<thead>
<tr>
<th>Log of creatinine clearance</th>
<th>Log of digoxin clearance</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.970</td>
<td>2.862</td>
</tr>
<tr>
<td>3.207</td>
<td>3.550</td>
</tr>
<tr>
<td>3.277</td>
<td>2.434</td>
</tr>
<tr>
<td>3.437</td>
<td>3.378</td>
</tr>
<tr>
<td>3.444</td>
<td>2.632</td>
</tr>
</tbody>
</table>

There does not seem to be any obvious directionality to any relationship there might be between digoxin clearance and creatinine clearance; it seems best to treat them on an equal footing. Figure 3.8 is a scatterplot of the logs of digoxin clearance and creatinine clearance.

Dataset name: *digoxin*.  
Note that since we are just considering the association between the variables and not a regression relationship, the choice of which variable to put on which axis is arbitrary; it would have been equally justifiable to plot the variables the other way round.

The scatterplot suggests quite a strong relationship between digoxin and creatinine clearances. The relationship is positive in the sense that large values for digoxin tend to go with large values for creatinine (and small with small). The strength of this association can be measured by the Pearson product-moment correlation coefficient, or correlation coefficient (or even just correlation) for short. In this case, the value of the correlation coefficient turns out to be 0.836.

The formula for the sample version \( r \) of the correlation coefficient is given by

\[
\frac{1}{n - 1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})
\]

where \( s_x \) and \( s_y \) are the sample standard deviations of the \( x_i \) and the \( y_i \), respectively. The correlation coefficient takes values between \(-1\) and \(+1\). The sign of \( r \) reflects the type of relationship: positive if high values of one variable go with high values of the other, negative if high values of one go with low values of the other. The stronger the relationship, the closer the value of \( r \) to 1 (if positive) or \(-1\) (if negative). No association means \( r = 0 \). The value of 0.836 found for the digoxin example above reflects the strong positive association between the variables seen in Figure 3.8.

There are, however, caveats about the use and interpretation of \( r \). First, the Pearson correlation coefficient is a measure of linear association between variables; that is, it is high (in absolute value) when the relationship between the variables is strong and roughly linear (in the straight-line sense). However, it is distinctly possible to have a very strong relationship between variables which is not well reflected by a straight line, in which case \( r \) could well be small. Second, if \( X \) and \( Y \) are independent, then the population correlation coefficient, the population version of \( r \), is zero. However, because of the first caveat, the converse does not hold: it is perfectly possible to have a zero correlation coefficient but a strong dependence between variables. Nonetheless, \( r \) is still a useful measure.

**Exercise 3.29 Calculating correlations**

Let us consider the two variables hardness and strength from the rubber dataset; these refer to the hardness, in degrees Shore, and to the tensile strength, in kg/cm², respectively, of samples of rubber. In Exercise 3.20, these variables were considered as explanatory variables for the response variable abrasion loss. But hardness and tensile strength are not necessarily unrelated themselves. However, there does not seem to be any obvious directionality to any relationship they may have; it seems best to treat them on an equal footing.

Load the rubber data into GenStat again. Make a scatterplot of hardness against tensile strength. Use the following method to calculate the correlation coefficient. Choose the Stats\Summary Statistics\Correlations menu item. Enter the two variables whose correlation is required, hardness and strength, in the Data field. Make sure the Correlations item is checked, and click on Run. Look in the Output window. What do the scatterplot and correlation coefficient tell you about the relationship between hardness and tensile strength?

One way of assessing the smallness of a correlation is to make a hypothesis test of whether the (population) correlation is zero, but this will not be described here. We shall make only subjective assessments of the size of the correlation in this course.

The main role of correlation in this course will be in respect of relationships between explanatory variables. In particular, as already suggested in Unit 1, if explanatory variables exhibit strong relationships with one another, it may well be
Summary

that we need only include a small subset of these explanatory variables in the regression model; the reason is that there is little extra information to be gained by including further variables that are closely related to those already included, since they behave much like repeats of those already there.

Finally, be aware that correlation between variables is not necessarily indicative of a causative effect. (Nor indeed are regression relationships necessarily causative.) For example, in the USA, there is a high positive correlation between teachers’ pay and alcoholism in teachers, yet the alcoholism is not caused directly by increasing pay nor are teachers paid more for being alcoholic! Such an observed relationship probably arises (perhaps in a complicated way) from other variables measuring stress levels and/or social activity. High correlation between smoking and lung cancer is now generally accepted to reflect a causative effect (i.e. smoking causes lung cancer), but one argument against this was that perhaps both variables are closely linked to a third variable reflecting lung types that had a propensity both to make people smoke and to develop lung cancer.

Causation, which is not an entirely well-defined concept, has to be established by routes other than correlation and regression. Typically, the aim is to carry out a study such that any posited causal relationship can be the only plausible explanation for the effects of interest.

Learning outcomes

You have been working to develop the following skills.

- Fit a simple regression line by least squares.
- Interpret parameter estimates and make inferences about them using hypothesis tests and confidence intervals.
- Obtain and interpret estimates and confidence intervals for the mean of the response variable.
- Obtain and interpret predictions and prediction intervals for values of the response variable.
- Check the assumptions of a regression model using residual plots.
- Select suitable transformations from the ladder of powers to improve the suitability of data for modelling via the simple linear regression model.
- Informally compare the slopes of different regression lines.
- Calculate and interpret the correlation between two variables.
Solutions to exercises

Solution 3.1

The scatterplot looks like this.

Note that stylistic elements (such as line styles and the colours of symbols) in the plots reproduced in this unit and in later units may differ from the GenStat defaults that you see on your screen.

The relationship between the variables appears well described by a straight line, abrasion loss decreasing with increasing hardness. The scatter about the main trend of the data looks reasonably even.

Solution 3.2

As mentioned in Solution 3.1, the relationship between the two variables looks linear, and since it looks like the scatter about any sensible fitted line is not greater at some parts of the line than others, the assumption of constant variance looks tenable. It is harder to check on the normality assumption and the independence assumption simply by looking at the scatterplot, but there is no obvious reason to doubt them. Overall, the simple linear regression model would appear to be appropriate.

Solution 3.3

The scatterplot, on scales showing the origin, is shown below. The points do indeed appear to lie close to a straight line through the origin.
Solution 3.4

The scatterplot is shown below. Note that, on this scatterplot, times long ago are far to the right, which conflicts with the usual convention for plots with time on the horizontal axis. If you desire a more conventional plot, you could plot extinction rate against \((-\text{time})\) or perhaps against \((265 - \text{time})\).

There is a noticeable upward trend from left to right in the plot; in other words, extinction rates seem to be getting smaller as time moves towards the present day. However, there are several marked outliers (possible ‘mass extinctions’, including the one at 65 MYBP, corresponding to the end of the Cretaceous era, when the dinosaurs also became extinct). Generally, it would seem inappropriate to assume normality for these data. It also appears that the variability of the data is greater for greater extinction rates. Overall, the simple linear regression model is not adequate.

Solution 3.5

(a) The part of the output that you ought to recognise is the bit starting **Estimates of parameters.** The estimate for the intercept, \(\hat{\alpha}\), is in the **estimate** column and the **Constant** row; the value is 550.4. The estimate for the slope, \(\hat{\beta}\), is in the same column, but in the **hardness** row; its value is \(-5.337\). The rows have these labels because we multiply the values in hardness by \(\hat{\beta}\) and add a constant, \(\hat{\alpha}\), to get fitted values.

Thus the fitted line is

\[
\text{abrasion loss} = 550.4 - 5.337 \times \text{hardness},
\]

where loss is measured in grams per hour and hardness in degrees Shore.

You might well have worked out that s.e. means ‘standard error’ and that some sort of \(t\)-test is referred to in the \(t(28)\) and \(t \text{ pr.}\) columns. We shall come back to these columns, and the rest of the output, later.
(b) When the line is drawn on the scatterplot, it appears to fit reasonably well.

\[ y = 100.0021 - 0.002904x, \]

where \( y \) is the percentage of non-contaminated peanuts and \( x \) is the aflatoxin level. When plotted on the scatterplot, as above, the line appears to fit reasonably well.
Solution 3.7

This time, because there is no constant (intercept), there is only one parameter estimate given in the output, in the before row of the Estimates of parameters section. Thus the fitted line is

\[ y = 0.9574x, \]

where \( y \) is the total energy consumption after insulation and \( x \) is the total energy consumption before insulation.

On the scatterplot, the fit looks reasonably good.

Solution 3.8

(a)

There is a lot of scatter in the numbers of taps at each caffeine dose, but (as far as can be seen) the scatter at each point looks reasonably normally distributed and the numbers of taps increase with dose in a reasonably linear manner. So fitting a straight line is not unreasonable.
(b) The fitted line is
\[ y = 244.75 + 0.0175x, \]
where \( y \) is the number of taps per minute and \( x \) is the caffeine dose. On the scatterplot, the fitted line goes where one might expect.

\begin{center}
\includegraphics[width=\textwidth]{scatterplot.png}
\end{center}

**Solution 3.9**

Relation (3.4) says that
\[ \frac{\hat{\beta} - \beta}{s/\sqrt{s_{xx}}} \sim t(n - 2), \]
where \( s_{xx} = \sum (x_i - \bar{x})^2 \). The denominator, \( s/\sqrt{s_{xx}} \), is called the standard error of \( \hat{\beta} \). The sample value of this is the value in the s.e. column of the Estimates of parameters section of the output. The hardness row is the one containing \( \beta \) and we can read off the standard error as 0.923, i.e.
\[ \frac{s}{\sqrt{s_{xx}}} = 0.923. \]
To calculate the \( t \)-statistic for testing \( H_0 : \beta = 0 \), we take \( \hat{\beta} \) and divide it by its standard error. This calculation gives the value printed in the t(28) column, namely −5.78. (The 28 is the number of degrees of freedom, \( n - 2 \), for this particular dataset.) This \( t \)-statistic can be compared with the t(28) distribution to give the \( p \) value. The \( p \) value is given in the t pr. column; it is < 0.001.

The same pattern applies to the Constant row: the standard error of \( \alpha \) is 65.8; the \( t \)-statistic for testing \( H_0 : \alpha = 0 \) is 8.37; the \( p \) value is < 0.001.

**Solution 3.10**

Both the v.r. value and the t(28) value (\( = \hat{\beta}/s.e.(\hat{\beta}) \)) in the output are test statistics for testing \( \beta = 0 \), and both yield \( p \) values < 0.001. The connection between the two is that
\[ \text{v.r.} = t(28)^2. \]
In this case, 33.43 \( \sim \) (−5.78)^2. The approximation is caused by rounding error. One way of noticing this might have been by observing that
\[ \log(\text{v.r.}) = 2\log(-t(28)). \]
So, really, the same test is being done in two ways, and exactly the same \( p \) value must ensue.
In general, it is a fact that if \( T \sim t(\nu) \) then \( T^2 \sim F(1, \nu). \)
**Solution 3.11**

On running the linear regression analysis, the value of the $t$-statistic for this test turns out to be 3.57. There are 30 datapoints. Comparing the $t$-statistic with a $t$-distribution with $30 - 2 = 28$ degrees of freedom, the $p$ value is very small at 0.001. The hypothesis of zero slope is clearly rejected. Alternatively, the v.r. value is $12.77 = 3.57^2$, allowing for rounding error, again on 28 degrees of freedom, and again with a $p$ value of 0.001. These data do provide evidence of an effect of caffeine on finger-tapping rate.

**Solution 3.12**

(a) From the GenStat output, $\hat{\alpha} = 550.4$, $\hat{\beta} = -5.337$, s.e.$(\hat{\alpha}) = 65.8$ and s.e.$(\hat{\beta}) = 0.923$.

(b) $t_{0.975}(28) = 2.048$. Remember that in the Probability Distribution Calculations dialogue box, quantiles are obtained by selecting the Equivalent Deviate option in the Calculation drop-down list.

(c) The 95% confidence interval for $\alpha$ is $\hat{\alpha} \pm t_{0.975}(28) \text{s.e.}(\hat{\alpha}) = (415.6, 685.2)$. The 95% confidence interval for $\beta$ is $\hat{\beta} \pm t_{0.975}(28) \text{s.e.}(\hat{\beta}) = (-7.227, -3.447)$.

**Solution 3.13**

(a) The scatterplot looks like this.

![Scatterplot](image)

The relationship looks reasonably linear (or, at any rate, it is not immediately obvious that it is curved) and it is plausible that a straight line through the origin would fit reasonably well. However, the points are all rather a long way from the origin, so it is difficult to judge from the scatterplot alone whether a line through the origin would be appropriate.

(b) The (unconstrained) fitted line is

$$y = 36.9 + 3.41x,$$

where $y$ denotes the average gas consumption in kWh and $x$ denotes the average temperature difference in °C.

The confidence interval for the intercept is found in the same way as was done in Exercise 3.12. The estimate of $\alpha$ is 36.9, and its (estimated) standard error is 17.0. The appropriate $t$ percentage point is $t_{0.975}(13) = 2.160$. Therefore the 95% confidence interval for the intercept of the regression line is $(0.18, 73.62)$. The interval is wide, largely because the data (and hence $\pi$) are a long way from the origin (where $x_0 = 0$), and so the data do not determine the value of the intercept at all accurately. However, the interval
does not include the value zero, and so provides some evidence that fitting a line through the origin is inappropriate for these data.

(An alternative type of model, which is often appropriate in situations like this, is to formulate a non-linear model which goes through the origin and behaves like a straight line in the vicinity of the data, but which joins the origin to this straight line using a curve.)

**Solution 3.14**

(a) The 90% confidence interval is given by

\[
\hat{y}_0 \pm t_{0.95}(32) \text{s.e.}(\hat{y}_0).
\]

GenStat gives \( t_{0.95}(32) \) as 1.694. So the 90% confidence interval for the mean response at an aflatoxin level of 13.2 is (99.95, 99.97). This range is an interval for \( \alpha + 13.2\beta \).

(b) The 90% prediction interval is also given by

\[
\hat{y}_0 \pm t_{0.95}(32) \text{s.e.}(\hat{y}_0),
\]

though this time \( \text{s.e.}(\hat{y}_0) = 0.04027 \). Thus the 90% prediction interval for the predicted percentage of non-contaminated peanuts at an aflatoxin level of 13.2 is (99.89, 100.03). This range is an interval for \( \alpha + 13.2\beta + \epsilon \). It is indeed rather wider than the confidence interval in part (a). Note that this prediction interval includes percentages of non-contaminated peanuts over 100%, an indication that the model could be improved upon.

**Solution 3.15**

(a) The residual plot is as follows (though different colours and line styles are shown here).

The plot of residuals against fitted values does not indicate any serious grounds for concern about the model. There is no clear sign of a pattern in it. The curved line across the middle gives an indication that the residuals tend to be rather higher at the ends of the range than in the middle, but the tendency is not very great.
(b) A normal probability plot of the residuals is produced.

![Normal Probability Plot]

It looks rather curved, indicating that the residuals are rather skewed. Perhaps transforming the data (see Section 3.5) would have led to a more normal distribution for the residuals; however, this improvement in modelling accuracy would have led to a more complicated model.

**Solution 3.16**

The composite residual plot is as follows.

![Composite Residual Plot]

The fitted values plot and the normal probability plot look almost exactly the same as (though smaller than) those in Exercise 3.15. The histogram shows fairly clearly the skewness of the residuals. Thus the composite plot does not (in this case) really tell us anything about the model we had not already noticed.
Solution 3.17

The composite residual plot looks like this.

The plot of residuals against fitted values shows no particularly marked sign of curvature, but, like the original scatterplot (Solution 3.6), it shows that there is a tendency for the residuals to be smaller at high fitted values, which correspond to low aflatoxin levels. This is reflected in the probability plot, which is reasonably straight but indicates that the extreme residuals (both high and low) are not spread out as much as they would be if they came from a normal distribution. The histogram looks a little skew. Thus the simple linear regression model is not perfect for these data, but the fit is not too bad.

Solution 3.18

(a) The scatterplot is as shown below.

As one might expect, the two variables are positively related (high values of one go with high values of the other). There is some evidence of curvature, in
that the chemical test values increase less rapidly for high magnetic test values than they do for low magnetic test values.

The equation of the fitted regression line is

\[ y = 8.96 + 0.5866x, \]

where \( y \) is the chemical test value and \( x \) is the magnetic test value.

In the composite residual plot, shown below, the plot of residuals against the fitted values emphasises the way the data appear to deviate from a straight line towards the right-hand end.

The probability plot and histogram give no clear signs of lack of normality.

(b) The index plot, shown below, gives an indication that the residuals tend to be negative for those points early in time order and positive for most of the second half of the data. Therefore, there is evidence of some sort of lack of independence between successive residuals.

(c) The simple linear regression model does not fit these data particularly well.
Solution 3.19

The scatterplot is as follows.

![Scatterplot](image)

Apart from one very high outlier, the data generally follow the pattern assumed in the simple linear regression model. The regression line is

$$y = 0.2069 + 0.003107x,$$

where $y$ is consumption in pints per capita and $x$ is temperature in °F. GenStat warns us about the outlier by producing a message saying that its standardised (deviance) residual is large.

The composite residual plot (not shown) shows nothing untoward apart from the single high outlier that we noticed in the original scatterplot. However, you may have noticed that the plot of the residuals in time order, shown below, is much too smooth for independent, random residuals. For instance, there is a noticeable run of relatively large negative residuals near the start. (The blue curve on the plot does not help to show this interesting structure!) This plot casts doubt on the assumption of independence in the simple linear regression model.

![Residuals Plot](image)
Solution 3.20

The plot of the residuals against tensile strength is as shown below.

There is a clear negative relationship between the residuals from the fitted line and the tensile strength of the rubber samples. That is, rubber samples with high tensile strength are much more likely to have negative residuals. Putting it another way, suppose you had two rubber samples with the same hardness but with very different strengths. According to the regression model of abrasion loss against hardness, they would be expected to have the same abrasion loss. But this residual plot indicates that the sample with higher tensile strength is much more likely to have a negative residual; in other words, its abrasion loss is likely to be less than that of the sample with lower tensile strength. Overall, therefore, the simple linear regression of abrasion loss against one explanatory variable, hardness, is not adequate to describe the data. We need a more complicated model, involving the other variable, tensile strength, as well.

Solution 3.21

Transforming $t$ to $\log t$ straightens the plot very effectively.

The fitted line is

$$p = 0.8464 - 0.07923 \log t.$$
In fitting this model, GenStat warns you that the residuals do not appear to be random. The composite residual plot is as follows.

Though the fitted values plot shows some evidence of curvature, generally the model fits reasonably well. (The curvature in the fitted values plot matches the warning message about non-random residuals in the output.)

**Solution 3.22**

The scatterplot of the data indicates that a simple linear regression model is inappropriate, for two reasons. First, the relationship between the variables is far from linear. Second, it appears that the variability of the tensile strength measurements is greater for high strengths than for low strengths, so the assumption of constant variance is inappropriate.
Solution 3.23

A logarithmic transformation of tensile strength appears to deal with the variance, as shown in this scatterplot. (You might instead have felt that a square root or reciprocal transformation is best — there is very little to choose between square root, reciprocal and log in this case.)

![Scatterplot](image)

Solution 3.24

(a) Square root and log transformations still produce a noticeably curved scatterplot, but transforming curing time to its reciprocal (i.e. a power transformation with power −1) does the trick, as the following scatterplot shows.

![Scatterplot](image)

Now the plot looks reasonably linear, and the variance looks reasonably constant. However, the transformation has been arrived at by a roundabout route; it is quite possible that a better transformation exists.

(b) The regression line is
\[
\log y = 3.6878 - \frac{1.1455}{x},
\]

where tensile strength \( y \) is measured in kg/cm\(^2\) and curing time \( x \) in days.

The residual plots (not shown) are not totally reassuring about the appropriateness of the simple linear regression model. In particular, the
histogram and the normal probability plot indicate skewness in the
distribution of the residuals. However, these departures from the simple
linear regression model do not appear very large, and it may be appropriate
to use the model with these transformed data.

Solution 3.25

(a) The scatterplot is as follows.

The relationship is clearly non-linear. (However, the scatter of points about
the curve in the y direction is fairly constant, so the variance does not appear
to vary with the response.)

(b) In straightening the curve by transforming y, we can only move the points in
the y direction (up or down). It is thus necessary to ‘pull up’ the points at
the top right in the scatterplot, or in other words to expand the larger values
of y relative to the smaller values. To do this, we must move up the ladder of
powers, considering transformations like $y^2$, $y^3$, and so on. A plot of $y^2$
against x, shown below, looks reasonably linear.

However, another difficulty has arisen regarding the simple linear regression
model. The spread of points about the line in the y direction now increases as
the response increases, because the transformation has spread out the larger
responses. The assumption of constant variance no longer seems appropriate.
(c) To straighten the plot by transforming \( x \), we need to move the points at the top right of the original scatterplot in part (a) towards the left, or in other words to expand the smaller values of \( x \) relative to the larger values. To do this, we must move down the ladder of powers, transforming \( x \) to something like \( x^{1/2} \), \( \log x \), \( x^{-1/2} \), \( x^{-1} \), . . . . Plotting \( y \) against \( x^{-1/2} \) produces a reasonably straight line, as shown below. (Plotting \( y \) against \( x^{-1} \) is more or less as good.)

This time, because the points have been moved only in the \( x \) direction, the scatter in the \( y \) direction has not been affected, and the assumption of constant variance looks appropriate.

(d) The ‘best’ transformation we have found leaves \( y \) unchanged and transforms \( x \) to \( x^{-1/2} \). For this transformation, the fitted line is

\[
y = 4.3897 - \frac{6.416}{\sqrt{x}}.
\]

The usual residual plots (not shown) provide no great cause for concern about the appropriateness of the model, though there is some slight suggestion of non-normality in the probability plot and histogram.

**Solution 3.26**

The scatterplot of the untransformed data is shown below.
The relationship looks reasonably linear, but there is a clear tendency for the points to be scattered further away from where the regression line would be for timbers with high values of hardness and density. Thus the assumption of constant variance about the regression line is not appropriate.

Transforming hardness in order to bring its higher values closer together, by a log, square root or reciprocal transformation, might reduce the variance of the higher values, but on its own it would produce a curved scatterplot. The curve can be avoided by transforming density as well, in the same direction on the ladder of powers. Plotting log(Janka hardness) against log(density) leads to the following scatterplot. (You are not expected to be able to jump directly to this.)

This is still very linear, but the problem of non-constant variance is no longer evident. For the transformed data, the fitted regression line is

\[ \log y = 0.015 + 1.8847 \log x, \]

where \( y \) is the Janka hardness and \( x \) is the density. GenStat flags one point as having a large standardised residual, but it does not show up on the plots as being grossly out of line with the others.

Plotting the residuals shows no obvious departure from the model. The normal probability plot is somewhat curved, indicating that perhaps the normality assumption is not exactly satisfied. (The plots are not shown.) But, overall, the simple linear regression model fits the transformed data well.

**Solution 3.27**

For the northern towns, the regression line is

\[
\text{mortality rate} = 1692.3 - 1.931 \times \text{calcium concentration}
\]

and the estimated standard error of the slope is 0.848.

For the southern towns, the line is

\[
\text{mortality rate} = 1522.8 - 2.093 \times \text{calcium concentration}
\]

and the estimated standard error of the slope is 0.566.
Solution 3.28

We have a normally distributed estimator for the difference between the slopes of the two regression lines, whose mean, under the null hypothesis of equal slopes, is 0, and whose estimated standard deviation is 1.0195. The observed value of this estimator is 0.162. Thus the observed value is well under one standard deviation away from its mean under the null hypothesis. It seems that the \( p \) value should be high, and there are no grounds for rejecting the null hypothesis of equal slopes.

We cannot calculate an exact \( p \) value from the given information merely using the normal distribution, because that calculation would not take into account the fact that the standard deviations involved have been estimated from the samples. (However, this should not actually make too much difference.)

Solution 3.29

The scatterplot (which you might have plotted with the variables the other way round) is not indicative of any strong relationship between hardness and tensile strength. This is reflected in the correlation coefficient, which is −0.2992. Any relationship between hardness and tensile strength is weak at most, the correlation being close to zero (and if there is any weak relationship it seems to be of the negative kind).
Index

anaerob 8
bristol 7, 11
cemstren 25
  comparing slopes 27
  composite residual plot 20
  conditional distribution 6
  confidence interval
    for regression coefficients 15, 16
    for the mean 17, 18
  correlation 31, 32
  correlation coefficient 32
cream 21
degrees of freedom 14
digoxin 31
error variance 6
half-normal plot 20
hardness 26
  index plot 21
  intercept 6
  iron 20
  ladder of powers 22, 24
  least squares 9, 10
  linear regression 4, 11
marine 9
  mean square 14
  monotonic 22, 23
  multiple regression 30
outliers 9, 21
peanuts 11, 18
  Pearson product-moment correlation coefficient 32
  percentage variance accounted for 15
  prediction interval 17, 18
regression coefficient 6
  regression through the origin 7
  residual 9, 18
  residual plot 19
  residual sum of squares (RSS) 10
rubber 5, 6, 11, 13, 16, 19–21, 32
  simple linear regression 4, 6, 11
  slope 6
strong 23, 24
  sum of squares 13
  summary of analysis table 13
taps 12, 15
temperat 16
  transformation 22
variance ratio 14
water 27, 28
wind 25