Chapter 5: Regression Analysis: fitting equations to data

5.0 Introduction

Regression analysis is, perhaps, the most important data modelling technique in statistics. Here, only the most elementary applications will be considered, but this introduction will provide a basis on which you can develop your knowledge should you require more sophisticated analyses for your work.

Regression is concerned with fitting equations to data. In the simplest case this means that we have a variable X which we believe is related to a second variable Y and which we wish to use to predict Y; the relationship between the two variables is believed to be linear. We do not assume that X is a perfect predictor of Y: in the systems we study there is always chance variation, which prevents us from making perfect predictions. The various ideas will be developed in a series of examples; most of the ideas will be introduced in the first example and then further illustrated in the others, but each example contains something new.

In some cases we will know in advance that a relationship exists; basic scientific and technological theory will provide the grounds for such knowledge. Our first example is of this type; it concerns establishing a quantitative relationship between the shear strength of spot welds and their diameter. The resulting equation will allow us to predict shear strength for welds of different diameters. Since the system is affected by chance variation (a given weld diameter does not always result in a weld with the same shear strength), a method for placing error bounds around our predictions will be discussed.

Example 2 is an epidemiological study of the possible relationship between dietary fat levels and prostate cancer rates across countries. Here the first question is 'is there a relationship between the two variables?'; if there is, then we will want to quantify the strength of the relationship. Many studies in the social and biological sciences are of this type – essentially either exploratory or attempting to quantify the extent of an assumed (but not necessarily proven relation).

Example 3 is concerned with the stability of a pharmaceutical product. Drugs may lose their potency over time; this requires, first, that a shelf-life be established and, then, that the potency be monitored over the lifetime of the drug. We will see how prediction error bounds can be used to monitor drug potency over time: the bounds are used in essentially the same way as the control charts discussed in Chapter 1 for monitoring industrial production, measurement systems and retail sales, and in Chapter 2 for monitoring the relative bias between two laboratories.

Example 4 introduces the idea of instrument calibration. Most (probably all – I cannot think of an exception!) scientific and technological measurement involves
an instrumental response which is an indirect measure of the quantity of interest. Thus, very many (bio-)chemical measurement systems produce a trace output and the areas under the peaks give an indirect measure of the quantities of different substances in the material being measured. Similarly, physical measurements often result in an electrical response which, in turn, produces a deflection on a dial. A quantitative relationship must be established between these indirect responses and standard amounts of the quantities that are to be measured routinely, before the instrument can be used: this process is known as ‘calibration’.

These four examples introduce and illustrate the various technical features of regression analysis, including the underlying statistical model and residual analysis to validate this model, parameter estimates, confidence intervals and tests on parameters, making predictions, and describing how well the model summarises the observed data. The last section is different – it discusses some pitfalls in using regression analysis naively – the title says it all: ‘Regression Health Warnings’.

Finally, why is it called ‘regression’? The ideas were developed during the nineteenth century by people interested in heredity and were applied to datasets such as measurements of the heights of fathers and their sons. It was found that while tall fathers tended to have tall sons, the sons tended to be shorter than the fathers. Similarly, the sons of short fathers tended to be short, but taller than their fathers. This phenomenon was described as ‘regression to the mean’. The word ‘regression’ became attached to the body of techniques used in fitting quantitative relations; it has no intrinsic meaning beyond being a label for these ideas and techniques.

5.1 Fitting and interpreting regression models

Example 1: Strengths of Spot Welds

<table>
<thead>
<tr>
<th>Strength</th>
<th>3327</th>
<th>3240</th>
<th>3632</th>
<th>3435</th>
<th>4362</th>
<th>4236</th>
<th>4490</th>
<th>5556</th>
<th>5070</th>
<th>5796</th>
<th>6630</th>
<th>6410</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diameter</td>
<td>5.0</td>
<td>5.0</td>
<td>5.5</td>
<td>5.5</td>
<td>5.5</td>
<td>6.0</td>
<td>6.0</td>
<td>6.5</td>
<td>6.5</td>
<td>6.5</td>
<td>7.0</td>
<td>7.0</td>
</tr>
</tbody>
</table>

Table 5.1.1: Weld diameter (mm) and shear strength (N) for 12 spot welds

Table 5.1.1 shows data from a study carried out to determine the relationship between shear strength (Newton) and weld diameter (mm) of spot welds\(^1\). The data are displayed as a scatterplot in Figure 5.1.1.

---

\(^1\) Spot welding is a type of resistance welding used to weld various sheet metal products. Typically the sheets are in the 0.5-3.0 mm thickness range. The process uses two shaped copper alloy electrodes to concentrate welding current into a small "spot" and to simultaneously clamp the sheets together. Forcing a
For prediction purposes, we might expect that the strength of the weld would be proportional to its area, which would be proportional to the square of weld diameter. Accordingly, instead of plotting strength against diameter, as in Figure 5.1.1, we might plot it against diameter squared. Figure 5.1.2 shows a scatterplot of diameter squared versus diameter.

large current through the spot will melt the metal and form the weld. The attractive feature of spot welding is a lot of energy can be delivered to the spot in a very short time (ten to one hundred millisecond). That permits the welding to occur without excessive heating to the rest of the sheet. (Wikipedia)
Note that over the restricted range of the study (diameters of 5–7 mm), the relationship is effectively linear, so there is no real benefit to be gained from using the square of the diameter: it is simpler to use the diameter itself as our predictor variable, since it is in terms of weld diameter that engineers would think about weld size. Note, though, that this would not be the case if the range of diameters studied was much wider.

It is clear from Figure 5.1.1 that there is a general tendency for strength to increase with diameter (as would be expected). To describe the relationship quantitatively, we draw a straight line through the datapoints. There are many possible ways to do this (simply fitting a line “by eye” will be perfectly adequate for many purposes), but the standard approach is to fit a line by “least squares”; this is done in Figure 5.1.3 (we will discuss the details later).

The line can be thought of as an “average line” through the data – at any given diameter, the strength of replicate welds will vary about this line; this is the line that (approximately) joins up the means of the strength values for different weld diameters.

Interpreting the fitted line

Obviously, the fitted line is subject to chance variation – adding extra data points or deleting some of those currently available will change the fitted line. However, for the moment we will ignore this and, taking the line at face value, will interpret the results as we have them.
The slope of the fitted line is 1649 – this is the amount by which the strength would be expected to increase, on average, as the diameter is increased by 1 mm over the linear range of the study, i.e., between 5 and 7 mm. If the line is projected backwards towards zero diameter, the intercept at zero is – 5220. Clearly, negative strength is meaningless, as is a weld with zero diameter. We might, however, have expected the strength to be close to zero for very small weld diameters. If this were true for a very large dataset, then the negative fitted intercept in the current small study would simply be due to chance variation – we will investigate this below. Alternatively, the relationship between strength and diameter might be non-linear at very low diameter values, and it might take the linear shape suggested by Figure 5.1.3 as the diameter increased. Thus, a shape similar to that sketched in Figure 5.1.4 could produce a negative intercept when data were obtained only in the linear region. Such a shape would suggest that welds need to be of a minimum size before the linear relationship comes into force.

![Figure 5.1.4: Sketch of possible non-linear relationship](image)

The fact that the overall relationship is non-linear does not diminish the usefulness of the fitted linear section of the curve, provided we are only interested in making predictions within the diameter range studied (5-7 mm).

**Statistical Model**

The descriptive analysis given above will be sufficient for many purposes. If, however, we want to take account of the chance variation in, for example, quantifying the average change in strength for every extra millimetre in diameter, or in making predictions of strength at given diameter values, then we need a statistical model for the underlying data generation process.

The statistical model that underlies the regression analysis is defined as follows:
Independent observations
Normal distribution of $Y$ at any given $X$
Constant standard deviation
Means of $Y$ distributions can be joined by a straight line

The model is summarised in Figure 5.1.5.

![Figure 5.1.5: The statistical model underlying regression](image)

Note the correspondence between this model and those underlying the two-sample t-test (see page 34 of Chapter 2) and one-way ANOVA (see pages 6, 7 of Chapter 4). One-way ANOVA generalises the two-sample t-test model to many groups. Here, the means of the many groups (defined by the different values of the predictor variable, $X$) are assumed to be connectable by a straight line – otherwise, the model is the same as before.

The “true” or underlying relationship between the predictor, $X$, and the response, $Y$, is through the mean of the $Y$ values at any given $X$, i.e., $\mu_Y = E(Y)$:

$$\mu_Y = E(Y) = \beta_0 + \beta_1 X$$

The least squares line is an estimate of this “running mean”:

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

where the ‘hats’ indicate that we have estimates (i.e., values subject to chance variation) rather than model parameters, which are considered constants.
Individual observations vary randomly around the true line, with a standard deviation, $\sigma$, which does not depend on the X value. This means that the scatter around the line is the same for (relatively) large welds as it is for smaller diameter welds.

The model for individual observations can be written as:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$  \hspace{1cm} (5.1)

where $i$ labels the observations from 1 to $n$, and $\epsilon_i$ is an independent random ‘error’ term which comes from a Normal distribution with zero mean and constant standard deviation, $\sigma$. The mean is zero, as at any given value of $X = X_i$ the mean of the $Y$ values is defined by the first part of the model:

$$\mu_Y = E(Y_i) = \beta_0 + \beta_1 X_i$$

Note that this results in the responses (here weld strengths) being viewed as the sum of two components: the first part (above) is systematic and the second is a chance component, $\epsilon_i$.

Response = Systematic component + Chance component  \hspace{1cm} (5.2)

**Using the model**

**The slope coefficient**

The model allows us to put error bounds around the point estimates provided by the fitted line. For example, suppose the engineer asks ‘by how much, on average, does the strength increase for every extra mm of diameter?’ The fitted line provides a point estimate: $\hat{\beta}_1 = 1649$, but it is obvious that this would change with the addition of new data or the deletion of any of the current data points. Can we obtain an interval estimate of the true slope coefficient?

A 95% confidence interval for the true slope is given by:

$$\hat{\beta}_1 \pm t_{c}SE(\hat{\beta}_1)$$  \hspace{1cm} (5.3)

1649 ± 2.23(146)

1649 ± 326
We estimate the true slope to be between 1326 N and 1975 N: this is our answer to the engineer’s question.

Note that this interval has exactly the same form as the one and two-sample confidence intervals of Chapter 2: it is calculated as the point estimate plus or minus the estimated standard error, multiplied by the appropriate t-distribution critical value. In Chapter 2 we obtained a confidence interval for a population or process mean based on sample data using:

$$\bar{y} \pm t_c S\hat{E}(\bar{y})$$

where the estimated standard error of $\bar{y}$ was $s/\sqrt{n}$. In the current context the estimated standard error of $\hat{\beta}_1$ is:

$$S\hat{E}(\hat{\beta}_1) = \sqrt{s^2 \left(\sum_{i=1}^{n} (X_i - \bar{X})^2\right)} = s \sqrt{\frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n}}$$  \hspace{1cm} (5.4)

where $s$ measures the variation of the data around the line (the running mean, $\hat{y}$) just as in the single sample case $s$ measured the variation around the sample mean, $\bar{y}$. We will discuss this further later. Note also that in this case $s$ (and therefore the t-distribution we use in calculating the confidence interval) has $(n-2)=10$ degrees of freedom – this will be discussed later, also.

**Regression Analysis: Shear Strength versus Diameter**

The regression equation is

Shear Strength = - 5220 + 1649 Diameter

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Coef</th>
<th>SE Coef</th>
<th>T</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>-5220.3</td>
<td>881.3</td>
<td>-5.92</td>
<td>0.000</td>
</tr>
<tr>
<td>Diameter</td>
<td>1649.0</td>
<td>146.0</td>
<td>11.30</td>
<td>0.000</td>
</tr>
</tbody>
</table>

$S = 342.307$ \hspace{1cm} R-Sq = 92.7\% \hspace{1cm} R-Sq(adj) = 92.0\%

Predicted Values for New Observations

<table>
<thead>
<tr>
<th>New Obs</th>
<th>Fit</th>
<th>SE Fit</th>
<th>95% CI</th>
<th>95% PI</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>4673.7</td>
<td>98.8</td>
<td>(4453.5, 4893.8)</td>
<td>(3879.8, 5467.5)</td>
</tr>
</tbody>
</table>

Values of Predictors for New Observations

<table>
<thead>
<tr>
<th>New</th>
<th>Diameter</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>6.00</td>
</tr>
</tbody>
</table>

Table 5.1.2: Partial regression output for the spot welds data
In practice the calculations are taken care of by the software; thus, Table 5.1.2 shows output that comes with the fitted regression line. The line of the output corresponding to the slope parameter (labelled Diameter) gives the fitted slope parameter \( \hat{\beta}_1 = 1649 \) and also its estimated standard error \( SE(\hat{\beta}_1) = 146 \). These values were inserted into expression (5.3) to obtain the confidence interval.

The output also contains a t-value (11.3) and a p-value (0.000, which means \( p<0.0005 \)). These refer to a t-test of the hypothesis that the true slope is zero – this is the default test in most software. The null and alternative hypotheses for this test may be stated as:

\[
H_0: \beta_1 = 0 \\
H_1: \beta_1 \neq 0.
\]

The test statistic is:

\[
t = \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)} = \frac{1649}{146} = 11.3.
\]

The critical values that enclose 95% of a t-distribution with \( n-2 = 10 \) degrees of freedom are \( \pm 2.23 \). Since the magnitude of our test statistic greatly exceeds 2.23, we reject \( H_0 \) and conclude that the long-run or true slope is not zero. This conclusion follows directly from the very small p-value (\( p<0.0005 \)).

A true slope of zero would mean that the average strength was constant, i.e., that there is no relationship between strength and weld diameter. In a technological context, such as the welding example, such a test is superfluous – we know, on physical grounds, that a relationship exists. In other contexts this test will be important. In the social and biological sciences, for example, the question posed by the analysis is often ‘is there a relationship?’, and the t-test on the slope provides the answer.

There is a one-to-one correspondence between the confidence interval and the t-test on the slope coefficient. If the t-test is statistically significant, as here, the confidence interval does not contain the value stated in the null hypothesis (zero in the current case). Conversely, if the confidence interval had contained zero, the t-test would have been non-significant. For this reason, the confidence interval tells us what the test would have told us; it also gives us an interval estimate of the magnitude of the effect we are studying, and is, therefore, more useful.
The intercept term

The line in Table 5.1.2 labelled ‘Constant’ refers to the intercept – similarly to the information on the slope, it provides both a t-test and the quantities required to calculate a confidence interval for the true or long-run intercept.

A formal test of the hypothesis that the true intercept is zero may be carried out as follows.

The null and alternative hypotheses are:

\[ H_0: \beta_0 = 0 \]
\[ H_1: \beta_0 \neq 0 . \]

The test statistic is:

\[ t = \frac{\hat{\beta}_0 - 0}{SE(\hat{\beta}_0)} = \frac{-5220.3}{881.3} = -5.92 . \]

As for the test on the slope, the critical values are ± 2.23. Since the magnitude of our test statistic greatly exceeds 2.23, we reject \( H_0 \) and conclude that the long-run or true intercept is not zero. This conclusion follows directly from the very small p-value (p <0.0005).

Since a negative intercept makes no physical sense, we conclude that the linear relationship does not hold close to zero. A relationship similar to that sketched in Figure 5.1.4 may hold, but we have no information below a diameter of 5mm and so we can only conjecture as to what might be the shape.

A 95% confidence interval for the true intercept parameter is given by:

\[ \hat{\beta}_0 \pm t_cSE(\hat{\beta}_0) \]

\[ -5220 \pm 2.23(881.3) \]
\[ -5220 \pm 1965 \]

Thus, we estimate the long-run mean weld strength to be between –7185 N and –3255 N when the diameter is zero. As stated already, neither the concept of a negative strength nor that of a zero diameter weld make sense. The fact that the interval does not contain zero is consistent with the result of the t-test: it implies that the observed intercept (–5220) is not just a chance fluctuation away from zero (if it were, we would not be concerned at the fact that it was negative, since we would interpret this as simply an expression of the chance variation in the
system). We interpret the negative result as indicating non-linearity, perhaps of the type suggested in Figure 5.1.4.

The estimated standard error of the fitted intercept coefficient $\hat{\beta}_0$ is given by:

$$S\hat{E}(\hat{\beta}_0) = \sqrt{s^2 \left( \frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^{n} (X_i - \bar{X})^2} \right)} \quad (5.6)$$

though we rarely need this formula, as the software automatically provides the calculated value.

**Making predictions**

One of the main reasons for developing regression relationships is to use them for making predictions. Suppose we predict the strength $Y$ at weld diameter $X=6$ mm; a point estimate is given by:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i \quad (5.7)$$

$$\hat{Y} = -5220.3 + 1649.0(6) = 4673.7$$

This value is subject to the chance variation in the study and would change with addition or deletion of data points. It makes sense, therefore, to put error bounds around this point predictor.

Figure 5.1.6 shows two sets of error bounds around the fitted line; for $X=6$ these give the values labelled as ‘95% CI’ and ‘95%PI’ in Table 5.1.2.
In Figure 5.1.6 the narrower bands are the confidence bounds. Suppose we made a very large number of spot welds at 6 mm, the fitted value (4673.3 N) is a point estimate of the long-run average strength for spot welds of this size. The confidence bounds (4453.5, 4893.8 N) have the usual interpretation of a confidence interval: taking the chance variation in our dataset into account, they are error bounds around our point estimate – we are 95% confident that this interval covers the true long-run mean.

If, instead of estimating the long-run mean at weld diameter 6 mm, what we want to do is predict the strength limits within which a single new weld with diameter 6 mm will lie, then we use the prediction interval bounds; these are wider (3879.8, 5467.5) than those intended to bracket the long-run mean at 6 mm; the confidence bounds are given by the fitted value at 6 mm ± 220 N, while the prediction bounds are given by the fitted value at 6 mm ± 794 N. The difference arises because a single value of Y is the sum of the true mean and the chance deviation around the mean, $\varepsilon_i$, due to the random component of the model. The confidence interval bounds express the uncertainty in our estimate of the true mean. However, even if the true mean were known, there would still be uncertainty in the value of Y, due to the chance component, $\varepsilon_i$. The difference between the bounds is quite large as the estimated standard deviation of individual strengths around the mean value for any given diameter is 342 N. The formulae for the two sets of bounds show explicitly how the difference arises.

The confidence interval bounds at $X=X_o$ are given by:

$$
Y_o \pm t_c \sqrt{s^2 \left( \frac{1}{n} + \frac{(X_o - \bar{X})^2}{\sum_{i=1}^{n} (X_i - \bar{X})^2} \right)}
$$

(5.8)

The prediction interval bounds at $X=X_o$ are given by:

$$
Y_o \pm t_c \sqrt{s^2 (1 + \frac{1}{n} + \frac{(X_o - \bar{X})^2}{\sum_{i=1}^{n} (X_i - \bar{X})^2})}
$$

(5.9)

The square root term in the confidence interval formula (5.8) is the standard error that accounts for the uncertainty in estimating the true mean of Y at $X=X_o$. The prediction interval standard error term (5.9) includes this component also, but there is an extra “1 +” in the formula, just before the $1/n$ term: this (which is really “$s^2 +$”) accounts for the additional uncertainty due to the single deviation, $\varepsilon_i$, present in the new observation.
Note that everything under the square root signs is constant except $(x_o - \bar{x})^2$—the further $x_o$ is from the mean, the larger this term becomes, and hence, the wider both the confidence and prediction bounds become. This explains why the bounds curve outwards in Figure 5.1.6 (the curvature is only very slight here but will be more marked in other cases—especially if we try to extrapolate beyond the range of the $X$ values in the study; see Figure 5.1.24 on page 39). Thus, predictions far from the mean of the data on which the line is based will be subject to greater uncertainty than those close to the mean.

When the sample size $n$ is large, the prediction formula is often simplified to:

$$\hat{y} \pm 2s.$$ 

For large $n$ both of the elements $\frac{1}{n} \sum \frac{(x_o - \bar{x})^2}{\sum (x_i - \bar{x})^2}$ in expression (5.9) will become small and $s^2$ will dominate: the quantity under the square root sign in (5.9) will reduce to $s$. For large degrees of freedom the $t$-value becomes more and more like the standard Normal distribution, 95% of whose area lies between $\pm 1.96$, which is rounded to 2.

**Data Collection**

Our discussion of the spot welds data has taken the dataset as given; but, in practice, the first question that arises asks ‘what data should be collected?’ The prediction formulae indicate that if our focus is on using the regression line for prediction, then our predictions will be most precise (have narrowest bounds) at the centre, $\bar{x}$. This suggests that our $X$ values should be chosen so that $\bar{x}$ coincides with, or is close to, the $X$ levels at which we will make predictions.

The slope coefficient measures the strength of the relationship, in the sense of quantifying by how much $Y$ changes, on average, for a unit change in $X$. How is the slope coefficient estimate affected by the data collection design? The formula for the estimated standard error of the fitted slope is:

$$S\hat{E}(\hat{\beta}_1) = \sqrt{s^2 / \sum (x_i - \bar{x})^2} = s / \sqrt{\sum (x_i - \bar{x})^2}$$

Note the implications of the denominator on the RHS for the uncertainty in the fitted slope—the larger the denominator, the smaller is the uncertainty in the fitted slope. The denominator is increased by:

(i) taking more observations (i.e., increasing $n$)
(ii) increasing the spread of the $X$ values
The s value is a measure of the chance component in our strength data – this is part of the system we are studying so we cannot change it; all we can do is choose our X values with a view to minimizing the effects of s on our inferences.

Figure 5.1.7 illustrates the effects of the selection of the X levels for the study.

![Figure 5.1.7: Different choices of X levels](image)

Compare layouts A and B in Figure 5.1.7: A is clearly a bad data collection strategy – because the X values are so close together, any small chance variation in the corresponding Y values will have a big influence on the fitted slope. Thus, deleting the point marked as X would cause the fitted line to flip from position 1 to position 2. Because the X values are well spread out in B, the chance variation in Y has less influence in determining the fitted slope.

For a given number of data points, splitting the data into two extreme sets, as in C, will maximise the denominator and thus reduce the uncertainty in the fitted slope value. This, however, is at the expense of providing no information on possible departures from linearity. Such an arrangement should only be considered where the linear relationship has previously been established, and there is no possibility that in the current study the relationship could change.

**Assessing the Model**

We have now seen how to use our model to answer practical questions, but we have not yet considered whether or not our model is good – does it fit the data closely and are the model assumptions, on which our tests and confidence intervals are based, valid? Obviously, if the model assumptions do not hold, we cannot rely on the inferences we have made, based on a bad model.

To consider model quality we will first discuss residual analysis for assessing the model assumptions – this follows directly from the residual analyses carried out in the context of two-sample tests in Chapter 2 and those described in the context of ANOVA in Chapter 4.
We will then return to the Analysis of Variance technique, discussed in Chapter 4; this will lead us to a breakdown of the total variation in the data into components associated with the systematic regression effect and the chance variation around the line. This in turn will lead us to a simple summary measure ($r^2$ or the ‘coefficient of determination’) of the proportion of the total variation which may be considered systematic. An alternative (equivalent) measure, the correlation coefficient will then be described.

**Residual Analysis**

As with all statistical methods, the conclusions drawn from a regression analysis may be wrong if the assumptions underlying the model do not hold. It is worthwhile, therefore, to carry out some simple checks on these assumptions. The statistical model that underlies linear regression implies that all the systematic variation is embodied in the straight line and, consequently, no further systematic variation should be discernable in the data. This means that if a scatterplot of the data that includes the fitted regression line is inspected, the data should vary randomly around the fitted line. Our ability to detect departures from random scatter will be much improved if, instead of plotting the raw data, we use residual plots for model validation. The residuals

$$e_i = Y_i - \hat{Y}_i = Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i)$$

are the vertical deviations, $e_i$, of the data points from the fitted regression line. These residuals are estimates of the error terms $\varepsilon_i$ and should have approximately the same properties. Accordingly, they should show no systematic trends or outliers – if they do, they call into question the assumptions underlying the regression model. Experience suggests that departures from assumptions frequently involve some form of relationship between residuals and fitted values and are likely to be detectable in a scatterplot of residuals versus fitted values, as illustrated below. The fitted value is equivalent to the X value for simple linear regression, but in multiple regression, where there are more X variables, the fitted value is preferred for plotting purposes, as it represents the systematic component of the model.

Figure 5.1.8 shows some schematic diagrams illustrating possible results from plotting residuals versus fitted values or other variables.
In A we see the kind of plot expected when the model assumptions hold (the model assumes constant standard deviation at all levels of X and, therefore, at all levels of the fitted values). Panel B shows the scatter increasing with the size of the fitted value – suppose Y is the time taken by students to solve problems and X is the level of complexity of the problem, then a scatterplot similar to B might be expected, indicating that experimental subjects take similar times to solve easy problems, but that more complex problems result in greater time differences. I encountered type C in a psychological study where scores were quite variable for 7 year old children, but 10 year old children all got almost full marks on a test that was supposedly suitable for children up to the age of 12 (it had been developed in the USA and, clearly, was unsuitable for Irish children, for some reason). Figure D illustrates non-linearity, E correlation in time and F a systematic relationship with some other X variable: all of these, B – F, violate the assumption that all the systematic variation is captured by the regression line and that the scatter around the line is purely chance variation whose magnitude does not depend on X. We will encounter real examples of some of these patterns in later sections.

Normal plots of residuals are used to assess the Normality assumption underlying the regression model. We saw in Section 4 of Chapter 1 examples of the types of departure from Normality that such plots may detect.

We return now to our spot welds data and draw plots of the residuals from the simple linear regression line. Note, however, that with so few observations we cannot expect to demonstrate beyond all doubt the validity of the model – only gross departures from the model would be likely to be detected with as few as a dozen observations.
Figure 5.1.9 is a plot of the residuals versus the fitted values. If there were very many values we would expect a roughly rectangular scatter of the data points around the horizontal zero line – as predicted by the assumption of constant standard deviation for different weld sizes. Figure 5.1.9 does not show any systematic departure from such a pattern, so we can have some confidence in this assumption.

Figure 5.1.10 is a Normal plot of the residuals – it suggests random scatter around a straight line, as would be expected from Normal data. The Anderson-Darling test statistic has an associated p-value of 0.289; since this is larger than the conventional cut-off value of 0.05, we cannot reject the null hypothesis that the data come from a Normal distribution.
We cannot check the assumption of data independence, as we do not have the time order in which the observations were made, so we will have to assume that the study was carried out in such a way as to make this a plausible assumption.

**Analysis of Variance**

We saw earlier that the regression model views data as the sum of two components:

\[
\text{Response} = \text{Systematic component} + \text{Chance component}.
\]

The systematic component (the line) describes that part of the variation which is predictable, while the error or chance component is unpredictable. To address the question ‘how much or what fraction of the variation in the data is predictable?’, we return to the technique of ANOVA, introduced in Chapter 4 for the analysis of comparative studies involving several groups.

We saw there that ANOVA provides a global test of the null hypothesis that all the long-run means are the same. It does this by dividing the total sum of squares (and degrees of freedom) into components associated with between-group and within-group variation. The within-group variation was assumed to be due to chance.

In a regression context, ANOVA similarly breaks the total sum of squares (and degrees of freedom) into two components, one associated with the fitted line (assumed systematic) and the other, with variation around the line (assumed random).
Figure 5.1.11 shows that the deviation of any data point from the mean of all the data, i.e., \( y_i - \bar{y} \), is the sum of two parts, viz., the deviation of the point from the corresponding point on the regression line, \( y_i - \hat{y}_i \), and the deviation, \( \hat{y}_i - \bar{y} \), of that fitted point from the overall mean:

\[
(Y_i - \bar{Y}) = (Y_i - \hat{Y}_i) + (\hat{Y}_i - \bar{Y}).
\]

When the two sides of this equation are squared and summed over all the data points, the cross-product of the two terms on the right-hand side sums to zero.

\[
\sum_{i=1}^{n} (Y_i - \bar{Y})^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 + \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2 \quad (5.10)
\]

\[
\text{SSTO} = \text{SSE} + \text{SSR}
\]

Thus, the total variation in \( Y \), as measured by the ‘total sum of squares’, SSTO, is partitioned in two components, the ‘error or residual sum of squares’, SSE, and the ‘regression sum of squares’, SSR. The error sum of squares measures the chance variation of the data points around the regression line. The regression sum of squares is a measure of the extent of the variation in \( Y \) as \( X \) varies – if it is a large fraction of the total sum of squares then \( X \) is a good predictor of \( Y \), if it is small then knowing \( X \) is not helpful in predicting \( Y \).
The ‘least squares’ line is the fitted line that results in the smallest possible value of SSE (i.e., of all possible lines it gets closest to the data in the sense of leaving as little as possible of the variation ‘unexplained’ – when variation is described by squared deviations). This is the criterion which is almost universally applied when fitting regression lines.

Ideally, we would measure the variation in the data around the long-run mean, \( \mu \), but since this is unknown we use \( \bar{Y} \) in its place; in doing so we use some of the information contained in the data and ‘lose’ a degree of freedom in the process. Thus, SSTO has \( n-1 \) degrees of freedom. Similarly, SSE has \( n-2 \) degrees of freedom, since two degrees of freedom are ‘lost’ in estimating both the intercept and slope that are used in calculating the fitted values, \( \hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i \). Consequently, SSR has only one degree of freedom.

This decomposition of sums of squares and degrees of freedom gives us an ANOVA table – the resulting table for the spot welds data is shown in Table 5.1.3; such tables are produced automatically as part of standard output for regression analyses.

### Analysis of Variance

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>1</td>
<td>14955605</td>
<td>14955605</td>
<td>127.64</td>
<td>0.000</td>
</tr>
<tr>
<td>Residual Error</td>
<td>10</td>
<td>1171743</td>
<td>117174</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>11</td>
<td>16127349</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 5.1.3: ANOVA output for weld strength

The table includes an F-test with an accompanying p-value. In multiple regression (where there are more than one predictor variables) the F-test checks for a relationship between the response, \( Y \), and the set of \( X \) variables. Where there is only one \( X \) variable, the F-test is redundant, as it is identical to the t-test on the slope coefficient: the F-value is, in fact, the square of the corresponding t-value (apart from rounding error).

The ANOVA gives us the standard deviation, \( s \), of the observed data points around the regression line: \( s = \sqrt{\text{MSE}} \), where the mean square error is \( \text{SSE}/(n-2) \). This is an important summary number (note that it appeared in all our standard error formulae) as it describes the chance variation in the system we are studying. If \( s \) is large predictions will have wide error bounds. The ANOVA also gives us the information required for calculating \( r^2 \), which is widely used as a measure of ‘goodness of fit’.
Coefficient of Determination

The ANOVA table leads to a measure of the closeness of agreement between the data points and the corresponding fitted values on the regression line; this is called the coefficient of determination ($r^2$, labelled R-sq in Table 5.1.2, page 8) and is defined as:

$$r^2 = \frac{SSR}{SSTO} = 1 - \frac{SSE}{SSTO}$$  \hspace{1cm} (5.11)

This summary statistic measures the proportion of the total variation that is associated with the regression line, as opposed to the variation (presumed random) of the data points around the line. Since it is a proportion it must lie between zero and 1, though packages often multiply by 100 and quote the result as a percentage; thus, $r^2$ is given as 92.7% in the spot welds output of Table 5.1.2. Values of $r^2$ close to 1 indicate good predictive ability. When all the data points lie on the fitted line, as shown in the left-hand panel of Figure 5.1.12, the

In engineering and physical or biological sciences it is often possible to say that variation in a response $Y$ is caused by variation in a predictor variable $X$. In the social sciences, on the other hand, such statements are generally more difficult to make with any certainty. Where two variables change together it is often better, then, to say variation in $Y$ is ‘associated’ with variation in $X$ – this makes no claims as to causality.
error sum of squares is zero – in such a case the line is a perfect predictor of $Y$, given an $X$ value, and $r^2$ is equal to 1. On the other hand, if there is no relationship between $Y$ and $X$, the fitted line will neither increase nor decrease as $X$ varies – $\hat{Y}$ will coincide with $\bar{Y}$, as shown in the right-hand panel of Figure 5.1.12, and $SSR = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 = 0$, which means $r^2=0$. As $r^2$ increases from zero to 1 the data points lie correspondingly closer to the fitted line and the predictive ability of the regression equation increases accordingly.

**Correlation coefficient**

The square root of the coefficient of determination is numerically equal to a summary measure called the **correlation coefficient** or, sometimes, Pearson’s product-moment correlation coefficient. The correlation coefficient is calculated as:

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

and was designed to measure the extent to which two variables, $X$ and $Y$, vary together. There is no assumption, as in regression, that one variable (the independent or predictor variable, $X$) determines the other (the response or dependent variable, $Y$). The index simply measures the ‘co-variation’ of the two variables, treating them symmetrically.

When the means are subtracted to give the deviations $(X_i - \bar{X})$ and $(Y_i - \bar{Y})$, the origin from which variation is measured shifts from $(0,0)$ to $(\bar{X},\bar{Y})$, as shown in Figure 5.1.13.

![Figure 5.1.13: Defining the correlation coefficient](image)
If the data points fall mainly in quadrants I and III, then there is a ‘positive’ relationship, i.e., the products \((X_i - \bar{X})(Y_i - \bar{Y})\) are mainly positive and \(X\) and \(Y\) tend to increase together; if they mainly fall into quadrants II and IV, the products are negative and the relationship is said to be ‘negative’ – as \(X\) increases \(Y\) decreases, and vice versa. The correlation coefficient lies between \(\pm 1\) – if all points fall on a line sloping upwards to the right, \(r = +1\), or, if downwards to the right \(r = -1\). If the data points fall into all four quadrants, so that the positive products (quadrants I and III) cancel the negative products (quadrants II and IV), then \(r = 0\), indicating no relationship between the two variables.

![Scatterplots to illustrate the implications of different values of r](image)

Figure 5.1.14: Scatterplots to illustrate the implications of different values of \(r\)

Figure 5.1.14 shows examples of the kinds of values of \(r\) that are given by different scatterplots. Note that even quite high values of \(r\) correspond to scatterplots that suggest \(X\) would not be an especially good predictor of \(Y\). This suggests that a scatterplot should always be drawn when assessing relationships between variables – single numerical summaries, such as \(r\) or \(r^2\), are inadequate on their own. Looking forward to Figures 5.2.8 and 5.2.9 (page 50) you will see a correlation of \(r=0.997\), even though the scatterplot (and, more particularly, the residual plot) shows the relationship to be non-linear. The correlation coefficient
is not, as is commonly thought, a measure of linearity – it measures the extent to which data points are close to a straight line, which is not the same thing.

The dangers of calculating correlation coefficients without drawing scatterplots are nicely illustrated by some data [1] kindly provided by Donata Dubber, an environmental scientist in Trinity College, Dublin. Donata studies sludge quality in wastewater treatment plants. One indicator she uses is the Specific Sludge Volume Index (SSVI), which is a measure of how well and how quickly the sludge settles, so that the supernatant, the final effluent, can be discharged. If the SSVI value is high, the settleability of the sludge is poor and organic particles might be discharged with the final effluent, indicating that the treatment process has been less efficient. It has been suggested in the literature that some free swimming protozoan species (unicellular micro-organisms) might have an impact on the settleability of the sludge (and therefore on SSVI). To investigate this, she counted the numbers of the species Trachelophyllum pusillum present and measured the SSVI of sludge sampled from 11 Irish wastewater treatment plants. The data are plotted in Figure 5.1.15.

The correlation coefficient for the full dataset (A) is 0.98 (p<0.0005)\(^3\) which suggests a strong relationship between the two variables. When the obvious outlier (due to operational problems at the plant) is removed (B) the correlation

\[ r = 0.78 \] (p-value not provided)

The p-values refer to a statistical test that the population correlation coefficient is zero; for dataset A this is strongly rejected, but for B the data are consistent with there being no relationship in the population of pairs of values. The test is equivalent to testing for a zero slope in a regression context.
coefficient drops to 0.19 (p=0.6), and the apparent relationship between the two variables has disappeared.

**Exercise 5.1.1**

Neter, Wasserman and Kutner [2] report the data of Table 5.1.4 on a study of the relationship between numbers of visitor days, \( Y \), and numbers of miles, \( X \), of intermediate level ski trails in a sample of New England ski resorts during normal snow conditions. A regression analysis was carried out in Minitab; partial output is shown in Table 5.1.5. Note that the output contains predictions at \( X=10 \).

<table>
<thead>
<tr>
<th>Miles-of-trails</th>
<th>Visitor Days</th>
</tr>
</thead>
<tbody>
<tr>
<td>10.5</td>
<td>19929</td>
</tr>
<tr>
<td>2.5</td>
<td>5839</td>
</tr>
<tr>
<td>13.1</td>
<td>23696</td>
</tr>
<tr>
<td>4.0</td>
<td>9881</td>
</tr>
<tr>
<td>14.7</td>
<td>30011</td>
</tr>
<tr>
<td>3.6</td>
<td>7241</td>
</tr>
<tr>
<td>7.1</td>
<td>11634</td>
</tr>
<tr>
<td>22.5</td>
<td>45684</td>
</tr>
<tr>
<td>17.0</td>
<td>36476</td>
</tr>
<tr>
<td>6.4</td>
<td>12068</td>
</tr>
</tbody>
</table>

Table 5.1.4: Ski trail data

**Regression Analysis: Visitor days versus Miles of trails**

VisitorDays = -364 + 2033 Miles-of-trails

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Coef</th>
<th>SE Coef</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>-364</td>
<td>1071</td>
</tr>
<tr>
<td>Miles-of</td>
<td>2032.53</td>
<td>89.91</td>
</tr>
</tbody>
</table>

Analysis of Variance

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>SS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>1</td>
<td>1617707400</td>
</tr>
<tr>
<td>Residual Error</td>
<td>25326684</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>1643034085</td>
<td></td>
</tr>
</tbody>
</table>

Predicted Values for New Observations

<table>
<thead>
<tr>
<th>New Obs</th>
<th>Fit</th>
<th>SE Fit</th>
<th>95.0% CI</th>
<th>95.0% PI</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>19961</td>
<td>563</td>
<td>(18664, 21259)</td>
<td>(15658, 24265)</td>
</tr>
</tbody>
</table>

Values of Predictors for New Observations

<table>
<thead>
<tr>
<th>New Obs</th>
<th>Miles of</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>10.0</td>
</tr>
<tr>
<td>2</td>
<td></td>
</tr>
</tbody>
</table>

Table 5.1.5: Minitab partial regression output for ski trail data
Figure 5.1.16: A scatterplot of the residuals versus the fitted values (ski data)

Figure 5.1.17: A Normal plot of the residuals for the ski trails data

- A t-test of the hypothesis that no relation exists between Visitor Days and Miles-of-trail is required. Specify the model underlying the regression analysis and express this hypothesis in terms of the model parameters. Choose an appropriate significance level and carry out the test. What do you conclude?
• Calculate a 95% confidence interval for the slope of the regression relation. Interpret this interval in the context of the data given above.
• The full standard output contains a p-value associated with the intercept parameter. State the null and alternative hypotheses corresponding to this p-value. Carry out a test of the null hypothesis and interpret your result in the context of the data given above. Comment on the appropriateness of this test.
• The output gives both a confidence interval and a prediction interval for X = 10 (miles of trails). Interpret these intervals in the context of the study.
• Fill in the numbers of degrees of freedom in the ANOVA table.
• Calculate $r^2$. Does the regression line fit the data closely?
• Figures 5.1.16 and 5.1.17 show residual plots of the ski trails data. What can we learn from these plots (bearing in mind the small size of the dataset)?

Example 2: Prostate Cancer Death Rates

Weiss [3] reports data on prostate cancer death rates (per 100,000) and dietary fat consumption (g/day) for 30 countries; the data are based on an article in Advances in Cancer Research, Reddy (1980). The data are shown in Table 5.1.6.

<table>
<thead>
<tr>
<th>Country No.</th>
<th>Diet-Fat</th>
<th>D-rate</th>
<th>Country No.</th>
<th>Diet-Fat</th>
<th>D-rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>38</td>
<td>0.9</td>
<td>16</td>
<td>97</td>
<td>10.1</td>
</tr>
<tr>
<td>2</td>
<td>29</td>
<td>1.3</td>
<td>17</td>
<td>73</td>
<td>11.4</td>
</tr>
<tr>
<td>3</td>
<td>42</td>
<td>1.6</td>
<td>18</td>
<td>112</td>
<td>11.1</td>
</tr>
<tr>
<td>4</td>
<td>57</td>
<td>4.5</td>
<td>19</td>
<td>100</td>
<td>13.1</td>
</tr>
<tr>
<td>5</td>
<td>96</td>
<td>4.8</td>
<td>20</td>
<td>134</td>
<td>12.9</td>
</tr>
<tr>
<td>6</td>
<td>47</td>
<td>5.4</td>
<td>21</td>
<td>142</td>
<td>13.4</td>
</tr>
<tr>
<td>7</td>
<td>67</td>
<td>5.5</td>
<td>22</td>
<td>119</td>
<td>13.9</td>
</tr>
<tr>
<td>8</td>
<td>72</td>
<td>5.6</td>
<td>23</td>
<td>137</td>
<td>14.4</td>
</tr>
<tr>
<td>9</td>
<td>93</td>
<td>6.4</td>
<td>24</td>
<td>152</td>
<td>14.4</td>
</tr>
<tr>
<td>10</td>
<td>58</td>
<td>7.8</td>
<td>25</td>
<td>129</td>
<td>15.1</td>
</tr>
<tr>
<td>11</td>
<td>95</td>
<td>8.4</td>
<td>26</td>
<td>156</td>
<td>15.9</td>
</tr>
<tr>
<td>12</td>
<td>67</td>
<td>8.8</td>
<td>27</td>
<td>147</td>
<td>16.3</td>
</tr>
<tr>
<td>13</td>
<td>62</td>
<td>9</td>
<td>28</td>
<td>133</td>
<td>16.8</td>
</tr>
<tr>
<td>14</td>
<td>96</td>
<td>9.1</td>
<td>29</td>
<td>132</td>
<td>18.4</td>
</tr>
<tr>
<td>15</td>
<td>86</td>
<td>9.4</td>
<td>30</td>
<td>143</td>
<td>12.4</td>
</tr>
</tbody>
</table>

Table 5.1.6: Prostate cancer death rates and dietary fat consumption for 30 countries

Figure 5.1.18 displays a least squares line fitted to the scatterplot of the data, which shows a general tendency for death rate to increase with dietary fat; however, there is not a one-to-one correspondence between them. The fitted
line estimates the mean death rate, averaged over (potentially) a very large number of countries, at any given dietary fat level. The scatter around the line shows how individual countries vary around this average level.

*The slope coefficient*

The slope coefficient, shown in Table 5.1.7, suggests that prostate cancer death rate increases by approximately 0.113 for every extra unit of dietary fat. The t-test corresponding to the slope coefficient asks if the ‘true’ or long-run slope is zero:

\[ H_0: \beta_1 = 0 \]
\[ H_1: \beta_1 \neq 0. \]

If the slope is zero, it means that as dietary fat changes, the expected death rate remains the same, i.e., there is not a linear relationship between mean death rate and dietary fat.

\[ t = \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)} = \frac{0.11336}{0.01126} = 10.06. \]

The degrees of freedom for the t-test are always those of the s value that is used in calculating the standard error. Here, this is \(n-2=28\). For a two-tail test using a significance level of \(\alpha=0.05\) the critical values are \(\pm 2.05\). The observed t-statistic is 10.06 which is very large compared to the critical value, so we reject the null hypothesis and conclude that a relationship does exist between mean death rate and dietary fat.

The p-value in Table 5.1.7 corresponding to the test on the regression slope coefficient is given as \(p=0.000\) (which means that it is less than 0.0005). The p-value is the area in the tails further from the centre (zero) than the observed t-statistic (or its negative value). Since the t-curve used in the test is the sampling distribution of the t-statistic, *when the null hypothesis is true*, the p-value tells us that the chances of observing such a t-statistic in the absence of a relationship between the two variables would be less than 0.0005. Since we did observe such a highly unlikely result (unlikely, that is, if the null hypothesis were true), the null hypothesis is called into question – the improbable result suggests that the null hypothesis is implausible.
Regression Analysis: D-rate versus Diet-Fat

The regression equation is

\[ \text{D-rate} = -1.06 + 0.113 \times \text{Diet-Fat} \]

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Coef</th>
<th>SE Coef</th>
<th>T</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>-1.063</td>
<td>1.170</td>
<td>-0.91</td>
<td>0.371</td>
</tr>
<tr>
<td>Diet-Fat</td>
<td>0.11336</td>
<td>0.01126</td>
<td>10.06</td>
<td>0.000</td>
</tr>
</tbody>
</table>

\[ S = 2.29453 \quad R-Sq = 78.3\% \]

Analysis of Variance

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>1</td>
<td>533.31</td>
<td>533.31</td>
<td>101.30</td>
<td>0.000</td>
</tr>
<tr>
<td>Residual Error</td>
<td>28</td>
<td>147.42</td>
<td>5.26</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>29</td>
<td>680.73</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Predicted Values for New Observations

<table>
<thead>
<tr>
<th>New Obs</th>
<th>Fit</th>
<th>SE Fit</th>
<th>95% CI</th>
<th>95% PI</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>15.148</td>
<td>0.666</td>
<td>(13.783, 16.512)</td>
<td>(10.253, 20.042)</td>
</tr>
</tbody>
</table>

Values of Predictors for New Observations

<table>
<thead>
<tr>
<th>New Obs</th>
<th>Diet-Fat</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>143</td>
</tr>
</tbody>
</table>

Table 5.1.7: Regression output for the prostate cancer data
A 95% confidence interval for the true slope parameter is given by:

\[ \hat{\beta}_1 \pm t_c \cdot SE(\hat{\beta}_1) \]

\[ 0.113 \pm 2.05(0.0113) \]

\[ 0.113 \pm 0.023 \]

We estimate the long-run slope to be between 0.090 and 0.136. This is the amount by which the death rate increases, on average, for every extra unit of dietary fat consumed. Note that this interval does not include zero. There is a one-to-one correspondence between the confidence interval and the t-test on the slope coefficient. If the t-test is statistically significant, as here, the confidence interval does not contain the value stated in the null hypothesis (zero in the current case).

**The intercept coefficient**

The negative intercept coefficient is counter-intuitive – we cannot have a negative death rate for prostate cancer! The coefficient could be negative for two reasons. The true intercept could be zero or positive but small, and the negative coefficient could be the result of the chance variation in the data. A different set of countries might give a non-negative coefficient, in such a case. Alternatively, the relationship could be non-linear close to zero dietary fat, with an overall shape similar to that of Figure 5.1.4. Since the lowest dietary fat value in the dataset is 29, we cannot investigate the shape of the relationship at low dietary fat values.

The t-test corresponding to the intercept coefficient asks if the ‘true’ or long-run mean response (death rate) is zero when the X-variable (dietary fat) is zero:

\[ H_0: \beta_0 = 0 \]
\[ H_1: \beta_0 \neq 0. \]

Presumably, no country will have a dietary fat level of zero. So, the function of the intercept here is to set a base mortality level, which is incremented as fat consumption increases, rather than to describe the behaviour of countries with a dietary fat level of zero. We could also think of it as the value towards which death rates would be driven as fat consumption was decreased (again assuming the linear relationship holds at low fat levels).

We do not necessarily expect it to be zero, since factors other than fat consumption might be expected to influence prostate cancer death rates – if other factors do cause prostate cancer, then even in the absence of fat
consumption, we would expect a non-zero death rate. Statistical packages automatically produce tests like this and zero is usually the default value tested.

A significant test result would reject the hypothesis of a zero intercept parameter and a positive coefficient would suggest a ‘background’ prostate cancer death rate, even in the absence of dietary fat. A non-significant test statistic would mean that there is no basis for rejecting the null hypothesis and we would conclude that the empirical data suggest a negligible prostate cancer death rate in the absence of dietary fat (although recognising that we do not have dietary fat data close to zero and such a conclusion involves an extrapolation of the data).

\[
t = \frac{\hat{\beta}_0 - 0}{\text{SE}(\hat{\beta}_0)} = \frac{-1.063}{1.170} = -0.91.
\]

The observed t-statistic is –0.91 which is small compared to the critical values of ±2.05, so we cannot reject the null hypothesis. We conclude that the data are consistent with the long-run intercept being zero. If the linear model is appropriate close to zero fat levels, then there is no evidence of a background prostate cancer death rate at zero fat levels.

The p-value corresponding to the test on the regression intercept coefficient is given as p=0.371. In the context of the null hypothesis of the true intercept parameter being zero, this is interpreted as saying that the data are consistent with the null hypothesis. The large p-value corresponds to the small t-value being not statistically significant.

A 95% confidence interval for the true intercept parameter is given by:

\[
\hat{\beta}_0 \pm t_c \text{SE}(\hat{\beta}_0)
\]

-1.063 ± 2.05(1.170)

-1.063 ± 2.399

We estimate the long-run mean death rate to be between –3.46 and 1.33 per 100,000. The interval includes zero, which implies that the data cannot rule out a zero intercept. Again, there is a one-to-one correspondence between the confidence interval and the t-test on the intercept coefficient. Since the t-test is not statistically significant, the confidence interval covers the value stated in the null hypothesis (zero in the current case).
Predictions

When the data were collected the dietary fat value for the UK (presumably not too different to that for Ireland) was 143 g/day and the prostate cancer death rate was 12.4 per 100,000. Did the UK death rate conform to the general pattern as suggested by the study?

The fitted regression line predicts a death rate of 15.148 corresponding to this value:

\[ \hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i \]

\[ \hat{Y} = -1.063 + 0.11336(143) = 15.148 \]

Table 5.1.7 shows confidence and prediction interval bounds around this fitted value. The confidence interval bounds are error bounds on the long-run mean death rate at any given dietary fat level. We would expect the mean death rate, averaged over hundreds of countries, each having a dietary fat consumption level of 143, to be covered by this interval (in theory the average is over an infinite number of countries, but gathering such data might present some practical problems!). The confidence interval bounds are not useful when addressing our question regarding the UK death rate – we are concerned with a single country, not the average over very many countries.

To answer our question we turn to the prediction interval bounds. These give us the limits (based on our fitted regression model) within which we would expect a single new value of Y (death rate) to lie at \( X = 143 \) (dietary fat level). The prediction interval bounds corresponding to the UK dietary fat value of 143 are 10.25 and 20.04. Since the reported death rate for prostate cancer is 12.4, the UK value is well within the prediction bounds – there is no reason to suspect that the UK prostate cancer death rate is in any substantial way inconsistent with those of other countries.

Residual Analysis

The scatterplot of residuals versus fitted values of Figure 5.1.19 does not show any systematic variation and, in particular, the assumption of constant standard deviation around the line appears reasonable.

The Normal plot of the residuals, shown in Figure 5.1.20, is reasonably straight, although there is some suggestion of either curvature or possibly two lines with different slopes (adding further ‘explanatory’ variables might improve the plot – it is unlikely that dietary fat consumption is the sole determinant of prostate cancer death rates). The Anderson-Darling test statistic has an associated p-value of 0.49, which is quite large. We accept, at least provisionally, therefore, the null hypothesis of data Normality.
The residuals in these figures are ordinary or ‘raw’ residuals – they are simply the differences between the observed values and the corresponding fitted values on the regression line. Raw residuals are often modified to make them more useful. Because the size of the residuals depends on the context of the data and the measurement units, it is useful to divide the residuals by their standard deviations to scale them – they will then have the properties (approximately) of a standard Normal distribution\(^4\), i.e., approximately 95% of them should be less than 2 and only very rarely will they be greater than 3. This gives us rules of thumb for what constitute unusually large or small residuals. Because of the way the line is fitted, the standard errors of the residuals depend on their positions with respect to the X-axis. In fact, that part of the expression for the prediction interval bounds which follows the t-value (Expression 5.9, page 12) is the standard error for the corresponding residual. Thus, the residuals corresponding to those X values furthest from the mean of the Xs have the greatest standard errors, which is why the error bounds of Figure 5.1.6 curve outwards. When the residuals are divided by their standard errors they are said to be ‘standardised’ or ‘Studentised’ – note that there is not universal agreement on terminology, so you need to check what is meant by these words in any given context.

The s used in calculating the standard errors can be strongly influenced by unusual values. A second adjustment to the raw residuals to allow for this is to delete each Y value in turn, fit the regression, calculate the value of s, calculate the residual for the deleted value (the deviation of this value from the line fitted to the remaining data), calculate the standard error for the residual, using the s value calculated while its Y value was deleted, and then divide the residual by its standard error. These residuals allow for the fact that unusual observations can pull the fitted regression line towards them, thus producing small raw residuals, which give no indication of the exceptional nature of the corresponding observation.

Figures 5.1.21 and 5.1.22 use deleted Studentised residuals. They are virtually identical to Figures 5.19 and 5.20, but note the difference in the scales on the y-axes. The advantage of re-scaling is that while for Figures 5.1.19 and 5.1.20 it is not clear whether the smallest residual (–5) is unusual (there must always be a smallest value!), Figures 5.1.21 and 5.1.22 show that the standardised value is about –2.5: such a value might be expected to occur by chance in a dataset of size 30, so we would not be especially concerned about retaining it in the dataset used to build the model.

\(^4\)This will be true where there are large sample sizes; for small sample sizes the t-distribution is the relevant reference distribution – hence the name ‘Studentised’.
Figure 5.1.19: Scatterplot of residuals versus fitted values (Cancer data)

Figure 5.1.20: Normal plot of residuals (Cancer data)
Figure 5.1.21: Scatterplot of deleted Studentised-residuals versus fitted values (Cancer data)

Figure 5.1.22: Normal plot of deleted Studentised-residuals (Cancer data)
Example 3: Stability testing of drugs


Table 5.1.8: Potency monitoring test results

<table>
<thead>
<tr>
<th>Month</th>
<th>0</th>
<th>3</th>
<th>6</th>
<th>9</th>
<th>12</th>
<th>18</th>
<th>24</th>
</tr>
</thead>
<tbody>
<tr>
<td>Potency%</td>
<td>102.6</td>
<td>100.4</td>
<td>98.4</td>
<td>99.4</td>
<td>99.8</td>
<td>97.8</td>
<td>97.6</td>
</tr>
</tbody>
</table>

Table 5.1.8 shows data from a stability testing programme in which samples of a drug product were stored at 25°C and 60% relative humidity [5]. Determinations of the potency of the drug product were made at seven time points, viz., zero, 3, 6, 9, 12, 18 and 24 months after manufacture. The reported results are ‘percentage of label claim’.

Figure 5.1.23: A regression line fitted to the potency data

Potency = $101.2 - 0.1683 \text{ Month}$

Figure 5.1.23 shows a regression line fitted to the scatterplot of the data; there appears to be a general tendency for the potency values, $Y$, to decrease with time, $X$. Table 5.1.9 gives standard regression output for the regression of drug potency on time: this will allow us to carry out a test to determine whether the decline in potency is systematic or just an artifact of the chance variation in the system of measurements. The fitted slope of $-0.168$ percentage points per
Month has an associated t-value of \(-3.17\) and a corresponding p-value of 0.025. These refer to the default test of the null hypothesis that the true slope is zero; this hypothesis is rejected and we conclude that the potency does, indeed, decay with time.

A confidence interval for the true slope, $\beta_1$, will provide a more realistic assessment of the change in potency with time than does the regression coefficient, $\hat{\beta}_1$, on its own. This interval is given by:

$$\hat{\beta}_1 \pm t_{\alpha/2}S\hat{E}(\hat{\beta}_1)$$

where the t value has $n-2=5$ degrees of freedom. For the potency data the 95% confidence interval for $\beta_1$ is given by:

$$-0.16826 \pm 2.57 (0.05302)$$

$$-0.16826 \pm 0.13626$$

**Regression Analysis: Potency versus Month**

The regression equation is

$\text{Potency} = 101 - 0.168 \text{Month}$

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Coef</th>
<th>SE Coef</th>
<th>T</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>101.159</td>
<td>0.685</td>
<td>147.58</td>
<td>0.000</td>
</tr>
<tr>
<td>Month</td>
<td>-0.16826</td>
<td>0.05302</td>
<td>-3.17</td>
<td>0.025</td>
</tr>
</tbody>
</table>

$s = 1.09873$  
$R^2 = 66.8\%$  
$R^2(\text{adj}) = 60.2\%$

**Analysis of Variance**

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>1</td>
<td>12.158</td>
<td>12.158</td>
<td>10.07</td>
<td>0.025</td>
</tr>
<tr>
<td>Residual Error</td>
<td>5</td>
<td>6.036</td>
<td>1.207</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>6</td>
<td>18.194</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Predicted Values for New Observations**

<table>
<thead>
<tr>
<th>New Obs</th>
<th>Fit</th>
<th>SE Fit</th>
<th>95% CI</th>
<th>95% PI</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>98.635</td>
<td>0.485</td>
<td>(97.389, 99.881)</td>
<td>(95.548, 101.722)</td>
</tr>
</tbody>
</table>

**Values of Predictors for New Observations**

<table>
<thead>
<tr>
<th>New</th>
</tr>
</thead>
<tbody>
<tr>
<td>Obs</td>
</tr>
<tr>
<td>Month</td>
</tr>
<tr>
<td>1 15.0</td>
</tr>
</tbody>
</table>

**Table 5.1.9**: Regression output for the potency data
The interpretation of this interval is that batch potency decreases by between 0.03% and 0.30% of the label claim for every month elapsed from manufacture; this estimate has an associated confidence level of 95%.

Table 5.1.9 also contains a t-test on the intercept (Constant). This, however, is not a useful, or even sensible, test in the current context, as the default null hypothesis is that the true intercept is zero, i.e., the average batch potency at time of manufacture was zero – a rather strange hypothesis in the context of drug manufacture! A more appropriate test would ask if the potency was 100% at time of manufacture, i.e., was the target potency achieved.

The null and alternative hypotheses are:

\[ H_0: \beta_0 = 100 \]
\[ H_1: \beta_0 \neq 100 \]

The test statistic is

\[ t = \frac{\hat{\beta}_0 - 100}{\text{SE}(\hat{\beta}_0)} = \frac{101.159 - 100}{0.065} = 1.69. \]

If a significance level of \( \alpha = 0.05 \) is to be used, then for a two-tailed test the critical values for \( n-2=5 \) degrees of freedom are \( \pm 2.57 \). The estimated intercept coefficient of 101.159 is said to be not statistically significantly different from 100. The conclusion from the regression analysis is that the batch potency was on-target (or close to target as other null hypothesis values close to 100% would similarly not be rejected by the test) at time of manufacture.

**Determining shelf-life**

If it is desired to estimate the mean batch potency at any particular time point \( X_o \) then, as we saw in Example 1, the expression shown below may be used to obtain a \( (1-\alpha) \)% confidence interval for the mean batch potency at time \( X_o \):

\[
\hat{Y}_o \pm t_c \sqrt{s^2 \left( \frac{1}{n} + \frac{(X_o - \bar{X})^2}{\sum_{i=1}^{n} (X_i - \bar{X})^2} \right)}
\]

where \( t \) is the critical value for the t-distribution with \( n-2 \) degrees of freedom which encloses \( (1-\alpha) \)% of the curve, and \( s \) estimates \( \sigma \). For example, the estimated mean batch potency at fifteen months after manufacture is given by:
\[ \hat{y}_{15} = 101.159 - 0.16826(15) = 98.635 \]

and the 95% confidence bounds are:

\[
98.635 \pm 2.57 \sqrt{1.207 \left( \frac{1}{7} + \frac{(15 - 10.286)^2}{429.429} \right)}
\]

giving an interval estimate between 97.4% and 99.9% of label claim. This expression is also used in determining the shelf-life of the drug, but instead of ‘forward prediction’ (from X to Y) the international guideline involves ‘inverse prediction’ (from Y to X).

The ICH guidelines define the shelf-life of a product to be the time point at which the 95% lower confidence bound for the batch potency intersects the ‘acceptable lower specification limit’. This means that the shelf life \( X_o \) is determined by solving the equation below which describes the point of intersection of the lower confidence bound and the specification limit, as shown in Figure 5.1.24.

\[
(\hat{\beta}_0 + \hat{\beta}_1 X_o) - t_c \left[ \frac{s^2 \left( \frac{1}{n} + \frac{(X_o - \bar{X})^2}{\sum_{i=1}^{n} (X_i - \bar{X})^2} \right)}{n} \right] = Y_{LSL} \quad (5.14)
\]

where \( Y_{LSL} \) is the ‘acceptable lower specification limit’. This is taken to be 90% in Figure 5.1.24.

This equation can be solved easily using a spreadsheet: the left-hand-side of the equation may be set up as a formula and then a series of values can be substituted for \( X_o \) until the result equals the right-hand-side, \( Y_{LSL} \). Alternatively, the value can be read from a plot similar to Figure 5.1.24, when this is provided by the statistical software. For the current example, Figure 5.1.24 indicates that the shelf life will be approximately 45 months, if the acceptable lower specification limit is taken to be 90% of label claim. Minitab calculates a lower bound of 89.99 at 44.25 months. Note that the very wide bounds at this point are an expression of the fact that the intersection with the lower specification limit occurs quite far to the right of the mean of the data values on which the regression line is based.

\[ ^* \text{Note that to obtain a one-side 95% lower bound the critical t-value is that which has 5% of the area in the left-hand tail of the t-curve; this is the same value (t =–2.015 in the current example) that is required for a two-sided 90% confidence interval.} \]
'Inverse prediction' is routinely used where regression lines are used for calibration – we will see examples of this in the next section.

**Monitoring product stability**

Expression (5.13) gives confidence limits within which the *mean* batch potency is expected to lie at any given time point; this is the agreed basis for defining shelf-life. Suppose we want to monitor product stability over time; what we will require are limits at a given time point within which we predict a new measurement to lie, if product stability/degradation rate remains unchanged. Expression (5.15), below, gives ‘prediction limits’, as discussed earlier – a single new potency measurement at any given time point would be expected to lie within these limits.

\[
\hat{Y}_o \pm t_c \sqrt{s^2 (1 + \frac{1}{n} + \frac{(X_o - \bar{X})^2}{\sum_{i=1}^{n} (X_i - \bar{X})^2})}
\]  

(5.15)

Figure 5.1.25 shows prediction limits for product potency, based only on the data up to 18 months after manufacture. Note that the fitted regression line is different from that of Figures 5.1.23 or 5.1.24 – this is to be expected, since only six of the seven data pairs have been used in fitting Figure 5.1.25.
The observed test result of 97.6 at 24 months is well inside the 95% prediction limits (91.4, 101.5) derived from the data up to 18 months. This indicates that the 24 month test result is entirely consistent with the prior data: the prediction limits bracket the test result values that might be expected based on the earlier measurements. Thus, there is no cause for concern – the new results do not call into question the established degradation relationship.

Using the regression line and prediction limits in this way is essentially equivalent to using control charts for monitoring system stability, as described in Chapters 1 and 2. If the 24 month test result had fallen below the lower prediction limit, this would be a cause for concern, as it would suggest either an acceleration in degradation or, perhaps, problems with the analysis of the test portion – an immediate investigation would be indicated.

The routine use of control charts in industry (see Chapter 1) is based typically on 3-sigma limits, i.e., the control limits are placed three standard errors on either side of the mean about which the plotted points are expected to vary at random. The standard deviation of observations is often based on more than 100 observations and, assuming Normality, the limits should enclose approximately 99.7% of plotted values. These charts are used in a context where there could be many points plotted every day, so if 95% bounds were chosen we would run into the problem of multiple tests (every time we plot a point we essentially test the hypothesis that the data are varying at random around the chart mean), and false alarm signals (points outside the limits), as discussed in Chapter 4.
case such as the present one, where the centre line and prediction bounds are based on only a small number of observations and where we are only interested in assessing one or a small number of new observations, the use of the t-distribution and 95% error bounds is safer.

Exercise 5.1.2:

The data shown below are the 24-hour percentage dissolution results for a drug product tested in a USP Type 1 dissolution apparatus [5]. The analysis was carried out using a UV spectrophotometer. The data come from a stability monitoring programme – the product was stored at 25°C and 60% relative humidity and was measured at the six time points indicated in Table 5.1.10. Each test result is the mean of six replicates (though this is not relevant to your calculations).

<table>
<thead>
<tr>
<th>Month</th>
<th>0</th>
<th>3</th>
<th>6</th>
<th>9</th>
<th>12</th>
<th>18</th>
</tr>
</thead>
<tbody>
<tr>
<td>% Dissolution</td>
<td>82.7</td>
<td>82.55</td>
<td>73.8</td>
<td>70.52</td>
<td>64.28</td>
<td>57.82</td>
</tr>
</tbody>
</table>

Table 5.1.10: Stability monitoring results

- Draw a scatterplot of the data
- Fit a least squares regression line to the data
- Calculate a 95% confidence interval for the slope and interpret the interval.
- Is the slope statistically significantly different from zero?
- Calculate a 95% confidence interval for the intercept and interpret the interval.
- Is the intercept statistically significantly different from 100%.
- Calculate a 95% confidence interval for the mean batch potency at 6 months and confirm that the measured result at 6 months is not within the interval.
- Within what bounds would you expect the measured dissolution rate to be at 24 months? Use a 95% prediction interval.

5.2 Calibration

Regression analysis is widely used in science and technology to establish calibration relations. Scientific measurement is usually indirect – the response that is observed is not the quantity of direct interest, but it is closely related to it and can be used to arrive at a value for the quantity under study. For example, in measuring pollutants in water, light is shone through a vial of water and the amounts of light that are absorbed by the water at particular wavelengths indicate the amounts of particular trace compounds in the water (agricultural fertilizer residues, for example). In order to use the instrumental responses to measure
the pollutants, a relation between the response (light absorption) and the pollutant concentration in a set of standard solutions has to be established first. This process is known as ‘calibration’. We encounter the results of calibration exercises in our everyday lives, also – the thermometers we use to take temperatures, the speedometers in our cars are but two examples of instrumental responses (length of a column of mercury, the deflection of a dial) which have to be calibrated in order to give indirect measures of the quantities of interest (body temperature or car speed).

**Example 4: HPLC analysis of red dye concentrations**

Hunter [6] reports the data in Table 5.2.1 as coming from a calibration study of a HPLC system. The concentrations are for naphthionic acid (FD&C Red No. 2) and the responses are chromatographic peak area.

<table>
<thead>
<tr>
<th>Conc.</th>
<th>0.18</th>
<th>0.35</th>
<th>0.055</th>
<th>0.022</th>
<th>0.29</th>
<th>0.15</th>
<th>0.044</th>
<th>0.028</th>
</tr>
</thead>
<tbody>
<tr>
<td>Peak Area</td>
<td>26.666</td>
<td>50.651</td>
<td>9.628</td>
<td>4.634</td>
<td>40.21</td>
<td>21.369</td>
<td>5.948</td>
<td>4.245</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Conc.</th>
<th>0.044</th>
<th>0.073</th>
<th>0.13</th>
<th>0.088</th>
<th>0.26</th>
<th>0.16</th>
<th>0.1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Peak Area</td>
<td>4.786</td>
<td>11.321</td>
<td>18.46</td>
<td>12.87</td>
<td>35.19</td>
<td>24.245</td>
<td>14.18</td>
</tr>
</tbody>
</table>

Table 5.2.1: Red dye calibration data

Table 5.2.2 gives standard regression output for the dye data, while Figure 5.2.1 gives a scatterplot of the data, with a superimposed regression line. It is clear from the scatterplot that the data are all close to a straight line. This is to be expected in a calibration context – if it were not the case the calibration exercise would be pointless, as the response would not then be a useful (inverse) predictor of the concentration. Scientists typically require the regression to produce a high $r^2$ or, equivalently, a high correlation coefficient, before they consider the calibration equation useful. Here we get a value of $r^2 = 0.994$, which is acceptably high; note that this was calculated as the ratio of the regression sum of squares and the total sum of squares, as given by the ANOVA table.
Regression Analysis: Peak-Area versus Conc.

The regression equation is
Peak-Area = 0.566 + 140 Conc.

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Coef</th>
<th>SE Coef</th>
<th>T</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>0.5665</td>
<td>0.4734</td>
<td>1.20</td>
<td>0.253</td>
</tr>
<tr>
<td>Conc.</td>
<td>139.759</td>
<td>2.889</td>
<td>48.38</td>
<td>0.000</td>
</tr>
</tbody>
</table>

S = 1.09273   R-Sq = 99.4%

Analysis of Variance

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>1</td>
<td>2794.3</td>
<td>2794.3</td>
<td>2340.19</td>
<td>0.000</td>
</tr>
<tr>
<td>Residual Error</td>
<td>13</td>
<td>15.5</td>
<td>1.2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>14</td>
<td>2809.8</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 5.2.2: Regression output for the dye data.

Figure 5.2.1: Scatterplot of the dye data, with a least squares regression line
For the red dye data the estimated calibration relation,

\[ \hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x \]

is

Peak Area = 0.5665 + 139.8 Conc.

When a test portion is measured and a response, say \( Y_0 \), is obtained, the calibration equation may be used to estimate the unknown concentration of dye in the test portion.

\[ \hat{x}_o = \frac{Y_0 - \hat{\beta}_0}{\hat{\beta}_1} \]  \hspace{1cm} (5.16)

Thus, if a test portion gives a response of 30, the calibration line predicts a concentration of:

\[ \hat{x}_o = \frac{30 - 0.5665}{139.8} = 0.21 \]

The inverse prediction is illustrated in Figure 5.2.2.
When we want to put error bounds around our prediction $X_o$, we use the cutpoints on the prediction error bounds around the fitted regression line, but this time the cutpoints are determined by the horizontal line beginning at the observed $Y$ value on the vertical axis. This is illustrated for the dye data in Figure 5.2.3.

An approximate algebraic expression for these bounds is:

$$
\hat{X}_o \pm t_c \frac{s}{\hat{\beta}_1} \sqrt{\frac{\sum_{i=1}^{n} (Y_o - \bar{Y})^2}{\hat{\beta}_1^2 \sum_{i=1}^{n} (X_i - \bar{X})^2}}
$$

(5.17)

where $\hat{X}_o = \frac{Y_o - \hat{\beta}_o}{\hat{\beta}_1}$.

The approximation involved will be good provided the fitted line is very close to the data, as would be expected in scientific or technological calibrations.

For the dye data, a test sample response of $Y_o = 30$ gives an estimated concentration of $\hat{X}_o = 0.21$; the error bounds are $0.21 \pm 0.02$, so the estimated concentration is between 0.19 and 0.23.

Chapter 6 of Mullins [5] contains a more extensive discussion of calibration.
Residual Analysis

Figure 5.2.4: A scatterplot of the residuals versus the fitted values for the dye data

Figure 5.2.5: A Normal plot of the residuals for the dye data
Figures 5.2.4 and 5.2.5 show residual plots for the dye concentration calibration study. Both plots appear to be satisfactory; therefore, we accept that the standard modal assumptions hold in this case. Measurement systems often show increased variability in the responses as the X variable (and typically the Y variable too) increases. This occurs if the range of the measurements is large. Here, the concentration range is quite narrow (from approximately 0.05 to 0.35) and Figure 5.2.4 gives no indication of such an effect. Our next two examples, on the other hand, do demonstrate such behaviour.

**Examples 5 and 6: Two bio-analytical assay examples**

The two datasets in Tables 5.2.3 and 5.2.4 come from bio-analytical assay validation studies in which replicate measurements were made at several concentration levels [5]. The standards are spiked blood plasma samples (µgL⁻¹). The responses are 1000 times the peak height ratios, i.e., the ratios of the heights of analyte peaks to the peak height for the internal standard.

<table>
<thead>
<tr>
<th>10</th>
<th>20</th>
<th>50</th>
<th>100</th>
<th>150</th>
</tr>
</thead>
<tbody>
<tr>
<td>69.2</td>
<td>134.5</td>
<td>302.0</td>
<td>495.7</td>
<td>773.0</td>
</tr>
<tr>
<td>64.5</td>
<td>136.8</td>
<td>246.8</td>
<td>537.9</td>
<td>813.6</td>
</tr>
<tr>
<td>50.0</td>
<td>142.8</td>
<td>257.5</td>
<td>474.5</td>
<td>681.6</td>
</tr>
<tr>
<td>52.2</td>
<td>85.4</td>
<td>239.0</td>
<td>516.0</td>
<td>763.9</td>
</tr>
<tr>
<td>40.3</td>
<td>87.2</td>
<td>198.4</td>
<td>348.5</td>
<td>629.9</td>
</tr>
</tbody>
</table>

Table 5.2.3 Bio-analytical calibration data

<table>
<thead>
<tr>
<th>1</th>
<th>2</th>
<th>4</th>
<th>40</th>
<th>80</th>
<th>160</th>
</tr>
</thead>
<tbody>
<tr>
<td>10.6</td>
<td>15.6</td>
<td>34.8</td>
<td>241.2</td>
<td>490.6</td>
<td>872.6</td>
</tr>
<tr>
<td>11.3</td>
<td>13.4</td>
<td>34.0</td>
<td>229.7</td>
<td>497.0</td>
<td>891.5</td>
</tr>
<tr>
<td>9.1</td>
<td>21.5</td>
<td>27.0</td>
<td>262.1</td>
<td>487.4</td>
<td>907.9</td>
</tr>
<tr>
<td>16.9</td>
<td>16.3</td>
<td>29.2</td>
<td>242.0</td>
<td>502.4</td>
<td>870.3</td>
</tr>
<tr>
<td>13.1</td>
<td>21.4</td>
<td>30.4</td>
<td>250.2</td>
<td>462.3</td>
<td>857.0</td>
</tr>
<tr>
<td>11.0</td>
<td>24.6</td>
<td>32.0</td>
<td>251.0</td>
<td>471.6</td>
<td>918.8</td>
</tr>
<tr>
<td>13.6</td>
<td>15.2</td>
<td>34.9</td>
<td>247.9</td>
<td>487.4</td>
<td>864.4</td>
</tr>
</tbody>
</table>

Table 5.2.4 Bio-analytical calibration data
Both datasets show departures from the standard regression assumptions. In both cases the variability increases with increasing response level. The scatterplot for the first dataset, Figure 5.2.6 shows this, but the residual plot, Figure 5.2.7, is more effective in doing so. Note the scales on the two graphs: 5.31 ranges from zero up to 800 while 5.32 ranges from −100 to +100. This is the main reason why (in one-variable regression) residual plots are a more useful tool than the scatterplot of the raw data.

For the second dataset Figures 5.2.8 and 5.2.9 again show increasing variability. Figure 5.2.9 shows clearly, also, that the relation between the response and the concentrations is curved rather than linear. The $r^2$ value for the regression is 0.997, which shows that the data are very close to a straight line; it was pointed out earlier that $r^2$ is not a measure of linearity, as is evident from Figure 5.2.9.

Note that fitting a curved line, as shown in Figure 5.2.10, has removed the curvature in the residual versus fitted values plot (Figure 5.2.11) but it has not had any effect on the changing variability (it was not expected to do so!). Weighted least squares (WLS, see Section 6.6 of Mullins [5]) will allow for increasing variability in the responses, as the concentration increases.

Typically WLS gives a fitted calibration line which is very close to that produced by ‘ordinary’ least squares (OLS). Consequently, the predicted value will be much the same under both approaches. However, if you want to put error bounds around the predicted $X_o$, then OLS will give equal error bounds on both sides of $(\bar{X}, \bar{Y})$, whereas WLS will take into account the fact that low responses are less variable than high responses. This is illustrated schematically in Figure 5.2.12.
Figure 5.2.6  A calibration line for the first bio-analytical assay

Figure 5.2.7  Residuals versus fitted values for the first bio-analytical dataset
Figure 5.2.8 The calibration line for the second bio-analytical dataset

Figure 5.2.9 Residuals versus fitted values for the second bio-analytical dataset
Figure 5.2.10  A curved response function for the second bio-analytical calibration

Figure 5.2.11  Residuals versus fitted values from the curved calibration line
Regression Health Warnings!

Once a regression line has been fitted to data there is a danger that it will acquire a life of its own – since it is based on mathematical techniques, it acquires the cachet of being ‘scientific’. No sensible scientist, though, will ignore the context in which a theory or an equation has been established. Unfortunately, it is all too easy to be deceived by the apparent objectivity of a regression equation.

Two distinguished American statisticians, Mostellar and Tukey [7], made a very interesting comment on regression analysis:

“Regression is probably the most powerful technique we have for analysing data. Correspondingly, it often seems to tell us more of what we want to know than our data possibly could provide. Such seemings are, of course, wholly misleading."

The full implications of their comment go beyond the simple model we have discussed. However, the essence of what they said is bound up with the context of the data analysis; two simple schematic examples will illustrates the dangers of uncritical use of regression analysis.

**Spurious Correlations**

Box [8] discusses a chemical process that uses a raw material which contains a variable percentage of an unknown impurity. The presence of this impurity has two consequences:

- it reduces the yield of the process;
- it causes frothing in the reactor vessel.
When frothing occurs, the operators respond by pumping an inert gas into the vessel: this controls the frothing and raises the pressure, but it has no effect on the yield of the process.

A review of historical data will result in a scatterplot of yield versus pressure similar to the schematic plot of Figure 5.2.13.

![Figure 5.2.13: Historical yields versus reactor vessel pressure](image)

A regression line fitted to the data will suggest a strong relation between yield and pressure, even though it is known on the basis of the chemistry of the process that pressure does not affect yield.

The problem here is that there is a relation between yield and pressure, but it is a **statistical** relation (the two variables vary together) not a **causal** relation. When the impurity level is high it results in low yield and high pressure (because the impurity causes frothing, the pressure is increased by the inert gas): these are the data points on the right-hand side of the diagram. When the impurity level is low the yield will be high, there will be little frothing and the pressure will be low (relatively): these are the data points on the left-hand side of the diagram. The statistical relation between yield and pressure is, therefore, mediated through this unknown third variable, impurity level. Box referred to this as a ‘lurking variable’.

The possibilities for third (or multiple other) variables to create apparent relationships are extensive in the social sciences. For example, socio-economic data very often appear as time series: wage rates, retail price indices, unemployment statistics etc. These variables move together in time (possibly in the same or opposite directions) and will, therefore, appear to be related (they are, statistically!) even though there may be no causal relation. In the physical and biological sciences, where controlled experiments can be carried out (generally, but not always) it is often possible to distinguish between statistical and causal relations, beyond reasonable doubt. In the social sciences purely...
observational data predominate and, consequently, distinguishing between statistical and causal relations is extremely difficult. Politico-economic arguments, for example, are largely based on different opinions as to possible causal mechanisms underlying observed economic relations: do high wage demands lead to higher retail prices, or vice-versa, for example? Our responses will depend on which side of the union/employer divide we sit.

Standardising for size is another obvious requirement to avoid spurious correlations. The numbers of babies born is likely to be strongly correlated with the numbers of lollipops sold in member countries of the European Union. The relation would (I expect!) disappear if we express both variables per 100,000 of population. Similarly, we might want to look at accident statistics in relation to numbers of vehicles or total journey lengths in different countries. Again, age-specific death rates may be more informative than crude (overall) death rates, when comparing countries with different age-structures.

**Extrapolation**

It is obvious (if we think about it!) that we are only entitled to make inferences within the range of the data on which our regression line is based. Thus, in the spot welds example, where the weld diameters in the study ranged from 5-7 mm, only a very foolish engineer would use our line to predict shear strengths of welds of diameter 17 mm. At least the engineer would have the possibility of carrying out further experimental runs at higher diameter values. In many situations no such option is available.

Consider the problem of forecasting demand for electricity in an economy where demand is growing, as shown schematically in Figure 5.2.14.

![Figure 5.2.14: Schematic relation between Demand and Time](image)

The predictor variable in our schematic example is simply time, but, in practice, the planners will have sophisticated econometric models containing many
variables thought to drive demand for electricity. Inevitably, if we are interested in future demand (for the purposes of planning new generation capacity, for example; this could require a planning/construction cycle of 5 or more years, depending on the technology) we will have to forecast the predictor variables also. Line 1 of Figure 5.2.14 simply projects the current linear relation into the future (assuming, for example, that industrial output will continue to grow, with consequent increasing energy demand). Line 2 shows a flattening of demand for electricity (perhaps due to an economic recession driven by strong increases in oil prices (as happened in the nineteen seventies).

Here, the energy modellers have no choice – they must extrapolate beyond the data on which the regression model is based. It is important, though, that they should recognize what they are doing – all forecasting is a form of crystal ball gazing (however sophisticated the mathematical clothes in which it is dressed)! Failure to acknowledge the inherent uncertainties involved in such exercises leads to the absurdly self-confident statements on how the economy is about to perform, that we hear from financial ‘experts’ on the radio, every day of the week.

We discussed earlier the extrapolation of the regression line back to the origin, in both the spot welds and the prostate cancer examples.
References


Outline Solutions

Exercise 5.1.1

The statistical model that underlies the regression analysis is defined as follows:

- Independent observations
- Normal distribution of $Y$ at any given $X$
- Constant standard deviation
- Means of $Y$ distributions can be joined by a straight line

The model is summarised in the diagram of Figure 5.1.1.1.

![Diagram of the statistical model underlying regression](image)

Figure 5.1.1.1: The statistical model underlying regression

The slope parameter ($\beta_1$) measures the amount by which the average value of $Y$ increases (decreases) for every extra unit of $X$. Therefore, our null hypothesis asserts that the ‘true’ slope (i.e., the model parameter as opposed to the estimate of this parameter obtained from the fitted regression line (the regression coefficient)) is zero: this would mean that there was no relation between the mean of the response ($Y$) and the predictor variable ($X$). The alternative hypothesis simply denies this assertion.

$$H_0: \beta_1 = 0$$

$$H_1: \beta_1 \neq 0.$$
The test statistic is:
\[
t = \frac{\hat{\beta}_1 - 0}{S\hat{E}(\hat{\beta}_1)} = \frac{2032.53}{89.91} = 22.61.
\]

If a significance level of \( \alpha = 0.05 \) is used, then for a two-tailed test the critical values for \( n-2 = 8 \) degrees of freedom are \( \pm 2.31 \). The estimated slope coefficient of 2032.5 is said to be statistically significantly different from zero. The conclusion from the regression analysis is that number of visitor days does depend on lengths of ski trails.

- A 95% confidence interval for the true slope is given by:
\[
\hat{\beta}_1 \pm t_cS\hat{E}(\hat{\beta}_1)
\]
\[
2032.53 \pm 2.31(89.91)
\]
\[
2032.53 \pm 207.7
\]
Thus, we are 95% confident that the average increase in visitor days (averaged over a large population of similar resorts) for every extra mile of ski trail, is between 1825 and 2240.

- The intercept parameter (\( \beta_0 \)) gives the long-run average (i.e., the average over a very large number of resorts) number of visitor days if there are zero miles of trail. This, of course, would mean that we were not discussing a ‘ski resort’ at all! Note that the smallest resort in the dataset has 2.5 miles of trail, so our dataset gives no information close to zero – interpreting the intercept, therefore, means that we are forced to assume that to extrapolate the straight line backwards to zero miles of trail is valid. The fact that the observed coefficient is negative could mean that the ‘true’ intercept is either zero or small and positive, and, that the observed negative value is just a result of the chance variation in the data. Alternatively, it could be that the straight-line relationship does not hold for resorts very close to zero miles of trail (if such exist!).

The null and alternative hypotheses are:
\[
H_0: \quad \beta_0 = 0
\]
\[
H_1: \quad \beta_0 \neq 0.
\]

The null hypothesis can be tested in exactly the same manner as was the slope parameter above, by simply replacing the slope coefficient and its
estimated standard error by the corresponding values for the intercept coefficient. This gives a t-value of \( t = \pm 0.34 \), which is compared to the same critical values as before. The degrees of freedom of the t-distribution are those of the value of the MSE (\( s^2 \)) that is used in calculating the relevant standard error, and this has \( (n-2=8) \) degrees of freedom in both cases. The observed t-value is not statistically significant: the data are consistent with a straight line that has a zero intercept.

Note that the results of the two tests could be deduced directly from the p-values normally shown in the full output table. The p-value is the probability of observing a more extreme test statistic than that observed in the study, if the null hypothesis is true. Thus, given our calculated t-values, we can deduce that the p-value for the intercept is large – meaning that the observed intercept coefficient is not far from zero when compared to the chance variation in the data, while that for the slope is very small, suggesting that the observed slope is not consistent with an underlying ‘true’ value of zero.

- The regression output gives two intervals corresponding to a fitted value of 19,961 visitor days at the \( X \) value of 10 miles of trail. The first of these, the ‘confidence interval’ estimates the long-run average value of \( Y \) when \( X=10 \); if data were assembled for many hundreds of ski resorts (in this area) each having 10 miles of ski trail, then the average number of visitor-days per resort would be expected to lie between the bounds of this interval, i.e., between 18,664 and 21,259 visitor days.

The second interval, the ‘prediction interval’, answers the question ‘within what range would we expect the visitor days to lie for a single new resort which has 10 miles of ski trail?’

The confidence interval is narrower because it refers to an average – individual values are subject to more variation than averages of similar quantities.

- The total degrees of freedom are \( (n-1)=10-1=9 \), those for error are \( (n-2)=10-2=8 \), leaving one degree of freedom for the regression sum of squares.

- The value of \( r^2 \) is:

\[
r^2 = \frac{SSR}{SSTO} = \frac{1617707400}{1643034085} = 0.985
\]

Since this value is quite close to 1, it suggests that ‘miles of trail’ is a good predictor of ‘visitor days’.
The residual plots are based on only 10 observations, so we need to be careful in our interpretations – chance patterns can be ‘over-interpreted’. For example, Figure 5.1.16 shows three values above zero, followed by four values below zero, followed by three values above zero. Does this pattern suggest curvature? I would hold an open mind on this (though a statistical test for curvature does not support such a view – but the test will not be powerful being based on so few data points) and proceed with the linear model. If more data became available it would be worthwhile reviewing this possibility.

The Normal plot of Figure 5.1.17 suggests that the assumption of Normality is reasonable.