

Chapter Nine

Simple Linear Regression

Consider the following three scenarios:

1. The CEO of the local Tourism Authority would like to know whether a family's annual expenditure on recreation is related to their annual income. This information could be used to tailor marketing campaigns to certain consumer segments.
2. A food company is interested in determining a shelf-life for a new chilled food product and hence they would like to quantify the relationship between microbial activity and time.
3. A car buyer is interested in purchasing a second hand car and would like to ascertain the relationship between a car's age and advertised purchase price.

What do these three scenarios have in common? The answer is that they all involve the quantification of the relationship between two variables. This relationship is of interest as it allows us to gain an understanding of the problem, to make predictions, or assess new data in light of the relationship. These problems are typically faced in many areas, including business, finance, genetics, food science, and many others.

Objectives: In this chapter you will learn

- when fitting a regression model is sensible;
- how to fit a regression model;
- how to assess whether a regression model is valid;
- what to do when the regression model assumptions are not valid; and
- how to interpreting the output produced by the software;
- how to use a regression model to make predictions.

Expenditure (\$)	Income (\$)	Exp. (\$)	Inc. (\$)
2400	41200	1450	37500
2650	50100	2020	36900
2350	52000	3750	48200
4950	66000	1675	34400
3100	44500	2400	29900
2500	37700	2550	44750
5106	73500	3880	60550
3100	37500	3330	52000
2900	56700	4050	67700
1750	35600	1150	20600

Table 9.1: Data on the annual expenditure on recreation and annual income of 20 families. The data can be found in [tourism.csv](#).

9.1 The Data

Suppose you are a consultant to the local Tourism Authority and the CEO of the Authority would like to know whether a family's annual expenditure on recreation is related to their annual income. In addition, if there is a relationship, he would like you to build a statistical model which quantifies the relationship between the two variables. A data set consisting of a random sample of 20 families, collected last year (Table 9.1) is available to help you with the assessment.

Notice that the problem revolves around two quantitative variables, namely a family's annual expenditure on recreation and their annual income. Since annual expenditure is likely to depend on a family's annual income, rather than the other way around, we refer to expenditure as the **dependent variable**, while income is considered to be the **independent variable**. The dependent variable, often denoted by Y is also referred to as the **response variable** while the independent variable, often denoted by X , is also referred to as the **explanatory** or **predictor variable**.

All problems addressed in this chapter deal with only two variables — the [next chapter \(Reference\)](#) looks at extending the methods developed here to situations where there are more than one explanatory variables.

Graphical summary of the data

The first step in any analysis is to visually examine the data. In the case of two quantitative variables the most appropriate graphical display is the **scatter plot**. The scatter plot allows investigation of the relationship between two variables ([link to Chapter 2](#)) — the independent variable is plotted along the horizontal axis and the dependent variable is plotted on the vertical axis. A scatter plot is frequently also referred to as a plot of Y versus X .

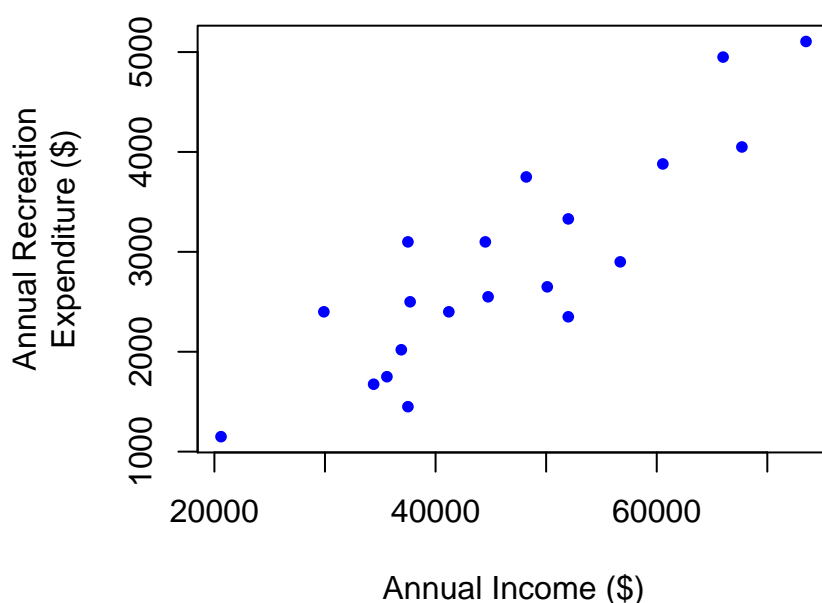


Figure 9.1: Scatter plot of annual expenditure on recreation versus annual income for a sample of 20 families (Tourism data).

A scatter plot of the data in Table 9.1 is shown in Figure 9.1. From Figure 9.1 it can be seen that the relationship between income and recreation expenditure has the following characteristics.

Direction: Positive, *i.e.* as income increases so does recreation expenditure;

Shape: Roughly linear, *i.e.* the points appear to fall along a straight line; and

Strength: Reasonably strong, *i.e.* there is considerable scatter about a straight line.

Clearly, the characterisation of the strength of the relationship is rather subjective and a numerical estimate of the strength is preferable. Given that the relationship between income and recreation expenditure appears linear, the strength of this linear relationship can be numerically summarized using the correlation, ρ , which is discussed in the following section.

Numerical summary of the data — Correlation

After investigating the data visually, a numerical summary of the strength of the association between the two variables is often desired. This can be achieved with the **population correlation coefficient**, ρ , which measures the strength of the *linear* association between two variables, X and Y . Since X and Y are two quantitative variables, ρ is also known as the **Pearson correlation coefficient** or **Pearson's product-moment correlation coefficient**.

Recall from [Chapter 2 \(reference\)](#) that ρ can take values between -1 and 1 and the interpretation of ρ is as follows:

- A negative value indicates a decreasing relationship between X and Y , that is, as X increases, Y decreases.
- A positive value indicates an increasing relationship between X and Y , that is, as X increases, so does Y .
- A value of 0 indicates that there is *no linear relationship* between the two variables — this however does not imply that there is *no relationship*.
- The correlation does not give an indication about the value of the slope of any linear relationship.

The type of relationship, and hence whether a correlation is an appropriate numerical summary, can only be assessed with a scatter plot.

In general, the whole population cannot be measured, but only a sample of n paired observations $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ is available. Table 9.1 give such pairs, where, for example, $(x_1, y_1) = (41200, 2400)$ denotes the pair of observations in the first row. From this sample the **sample correlation coefficient**, r , can be calculated. The mathematical formula used to calculate the sample correlation is

$$r = \frac{1}{n-1} \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right),$$

where \bar{x} and \bar{y} are the sample means and s_x and s_y are the sample standard deviations.

Of course, few people would calculate a correlation coefficient by hand these days, when it can easily be calculated by software. For example, the R-function `cor()` will calculate the sample correlation.¹

```
> cor(Tourism$income, Tourism$expend)
[1] 0.8752564
```

A Word of Caution:

Some caution must be exercised when interpreting correlations and scatter plots as there is no requirement for the two variables under investigation to have any sensible link whatsoever. For example, the average salary for teachers in Australia over the past 20 years might be highly correlated with the average price of one kilogram of cranberries in New York for the past 20 years.

¹In the Microsoft Excel or OpenOffice Calc the function `correl()` can be used.

But, these variables are not related to each other in any meaningful way. When a correlation exists between two seemingly unrelated variables, the correlation is said to be **spurious**!

Reference to a experimental design on the discussion of causation, confounding and common response.

9.2 The Simple Linear Regression Model

Statistics is all about taking large amounts of data and summarizing these data as concisely as possible. Consequently, when observing a linear relationship between two variables, it is only natural to ask the question “What is the equation of the linear relationship between the explanatory and response variables?”

Quantifying the relationship will allow us to

- better understand the functional relationship between X and Y , in particular, how quickly does Y increase for every unit increase in X , and
- make predictions about Y for a new value of X .

The linear regression model has the form:

$$Y_i = \beta_0 + \beta_1 X_i + E_i \quad (9.1)$$

where

- Y denotes the dependent variable;
- X denotes the independent variable;
- β_0 denotes the y -intercept;
- β_1 denotes the slope of the line; and
- E denotes a random error.

As usual, the realizations of Y_i and X_i are denoted by lower-case letters, *i.e.* y_i and x_i . However, the realization of E_i is denoted by r_i , the i -th **residual** (see Section 9.2).

The assumptions of the of the linear regression model are:

1. The E_i are statistically independent of each other;
2. The E_i have constant variance, σ_E^2 , for all values of x_i ;
3. The E_i are normally distributed with mean 0;
4. The means of the dependent variable Y fall on a straight line for all values of the independent variable X .

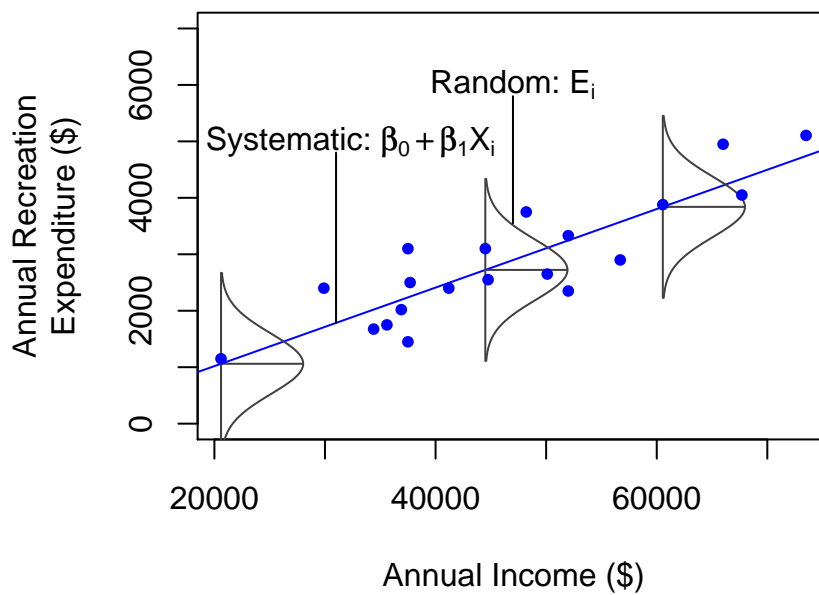


Figure 9.2: Graphical representation of a regression model overlaid on the Tourism data.

Note that the model in (9.1) consists of two components: a **systematic component**, $\beta_0 + \beta_1 X_i$, and a **random component**, E_i . This is common for statistical models — more complicated models can be created by changing either the form of the systematic component or the random component (for example see next Chapter).

The assumptions of the linear regression model are depicted graphically in Figure 9.2. From this graphic it can be seen that the regression model, the straight line, is the line connecting the average of the y values for each level of the independent variable, x . The actual y values for each level of x are normally distributed around the mean of y . In addition, the diagram also shows that the distribution of y is the same for each value of x , i.e. the variability is the same.

The model in (9.1) refers to the whole population, that is, β_0 , β_1 and σ^2 are unknown population parameters. The question of how these parameters are estimated from a sample is addressed in Section 9.2.

How the Line of Best Fit is found

Any straight line fitted to the data will take the form

$$y = b_0 + b_1 x$$

where b_0 and b_1 denote the intercept and the slope of the fitted line.² How good this line fits the data is determined by how close

²For an Java applet of this concept see http://onlinestatbook.com/stat_sim/reg_by_eye/index.html.

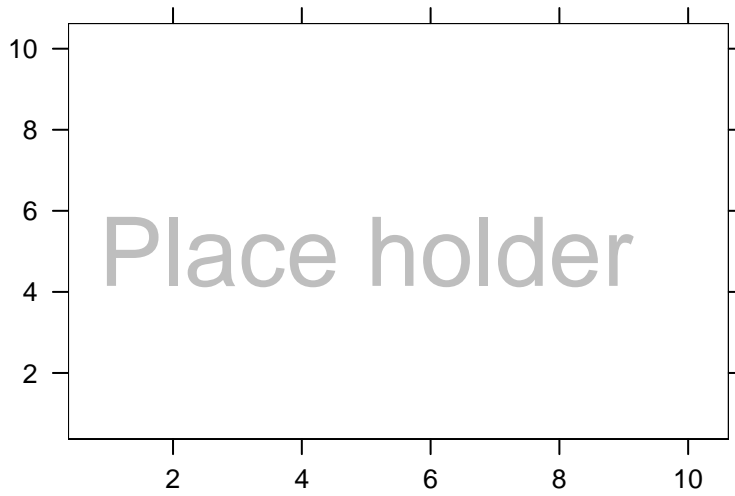


Figure 9.3: Graphical presentation of how the residual is calculated.

the data points (x_i, y_i) are to the corresponding points on the line, which are given by

$$\hat{y}_i = b_0 + b_1 x_i .$$

That is, for a given value x_i , the observed value of the dependent variable is y_i and the corresponding prediction from the line is \hat{y}_i (Figure 9.2). Consequently, to see how well the predicted value \hat{y}_i agrees with the observed value y_i the error, or residual, is calculated as the difference between the two, that is,

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

This is depicted in Figure 9.2. This residual r_i is

- negative if the observed value lies below the prediction, and
- positive if the observed value lies above the prediction.

The residuals need to be combined in some way to give an overall measure of how well the particular line fits the data. The important thing to remember is that positive *and* negative residuals need add to the **lack of fit**. Consequently, a commonly used measure of lack-of-fit is the **Residual Sum of Squares, RSS**, which is also known as **Sum of Squared Errors**,³ which is given by

$$\text{RSS} = \sum_{i=1}^n r_i^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2 . \quad (9.2)$$

Note that residuals of -2 and $+2$ have the same effect — they add an additional 4 units to the RSS.

³An alternative method is to use the absolute value of the residuals $|r_i|$, rather than the square, which is known as *least absolute value regression*. However, taking the square has some nicer mathematical properties.

So, what makes the *best* line? Well, the best line will result in the smallest RSS from all possible candidate lines $b_0 + b_1x$. This smallest RSS indicates that the correspondence between the data and the fitted line is as good as it can possibly be. This line is known as the **least squares regression (LSR) line**,⁴ and the estimates of the intercept and slope obtained from the LSR line are denoted by $\hat{\beta}_0$ and $\hat{\beta}_1$. The values of $\hat{\beta}_0$ and $\hat{\beta}_1$ are obtained from the output of the least squares regression routine from statistical software packages (see Section 9.2 and 9.3). For the data in Table 9.1 are $\hat{\beta}_0 = -274.88$ and $\hat{\beta}_1 = 0.066$.

The mathematical detail of how the least squares estimates are found is given below for those interested — these details are not essential for understanding the interpretation of the model. However, students intending to continue with their statistical learning should make themselves familiar with the approach.

Note that Equation (9.2) can be written as

$$\text{RSS} = \sum_{i=1}^n \left(y_i - (b_0 + b_1x_i) \right)^2,$$

and the aim is to find values for b_0 and b_1 such that RSS is minimized. In order to do this, the partial derivative of RSS with respect to each b_0 and b_1 are taken and equated to zero, *i.e.*

$$\begin{aligned} \frac{\partial \text{RSS}}{\partial b_0} &= \sum_{i=1}^n 2(y_i - (b_0 + b_1x_i)) = 0 \\ \frac{\partial \text{RSS}}{\partial b_1} &= \sum_{i=1}^n 2(y_i - (b_0 + b_1x_i))x_i = 0, \end{aligned}$$

The solution to these equations yields the **least squares regression estimates**

$$\begin{aligned} \hat{\beta}_0 &= \bar{y} - \hat{\beta}_1\bar{x} \\ \hat{\beta}_1 &= \frac{\sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2}, \end{aligned}$$

where \bar{y} and \bar{x} denote the mean of y and x , respectively.

Fitting the model

The steps required to fit a regression model will differ from one software application to another and not all possible software pro-

⁴The least squares regression model will also be referred to as the **linear regression model** since the only method of fitting a linear regression model will be via the method of *least squares*. An alternate approach is that of

grams can be discussed here. The following steps illustrate how to fit the model in R (`?`, `?`).

Firstly, the simple linear regression model is fitted using the `lm()` command,

```
> rec.lm <- lm(expend ~ income, data=Tourism)
```

which saves all the important information about the model in an object of the name `rec.lm`. This information can consequently be accessed using other commands, without the need to re-fit the model. For example, a summary of the fit and estimates of the regression coefficients are obtained using the `summary()` function.

```
> summary(rec.lm)
```

Call:

```
lm(formula = expend ~ income, data = Tourism)
```

Residuals:

Min	1Q	Median	3Q	Max
-895.09	-347.75	-26.82	368.02	863.70

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-372.645082	436.971630	-0.853	0.405
income	0.069572	0.009062	7.678	0.000000438 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 537.2 on 18 degrees of freedom

Multiple R-Squared: 0.7661, Adjusted R-squared: 0.7531

F-statistic: 58.95 on 1 and 18 DF, p-value: 0.000000438

In addition, the so-called **Analysis of Variance table** is obtained using the `anova()` function.

```
> anova(rec.lm)
```

Analysis of Variance Table

Response: expend

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
income	1	17010801	17010801	58.947	0.000000438 ***
Residuals	18	5194374	288576		

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Both pieces of output will be described in detail in Section 9.2.

However, before attempting to interpret the output it is important to check that the fitted model is actually appropriate. This is done through the use of model diagnostics. Some commonly used diagnostic tools are discussed in Section 9.2. But before delving into the diagnostics a more detailed discussion of residuals is warranted, since these play a fundamental role in model diagnostics.

Types of Residuals

Residuals play an important part in assessing whether the assumptions of the least squares regression model have been met. This assessment process is known as *model diagnostics*, and model diagnostics will be discussed in detail in Section 9.2. For a more complete treatment of regression diagnostics see ?).

Recall from Section 9.2 that the definition of the residual is the difference between the observed value y and predicted value \hat{y} , that is, the i -th residual is given by

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

These **ordinary** residuals can be obtained from the fitted model object using the `resid()` function.

```
> resid(rec.lm)
```

While most model diagnostics can be performed on these ordinary residuals, they have one undesirable property: they do not all have the same variability. This difference in variability is not due to the model assumption being inappropriate, but instead due to the least squares fitting process. It can be shown that the i -th residual r_i has variance

$$V[r_i] = \sigma_E^2(1 - h_i) ,$$

where h_i is known as the **hat value** denoting the **leverage** of the i -th observation, and is given by

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} .$$

The hat values capture how far each observation is from the mean; observations far away from the centre of the data will have large hat values while observations close to the mean will have small hat values. Note that hat values only rely on the independent variable and do not involved the dependent variable in any way. For a discussion on the impact of large leverage points on the model see Section 9.2.

In order to obtain residuals that have the same variability, each residual needs to be scaled by some value, usually involving the variability in some way. There are two similar approaches that can be used for this scaling operation, resulting in standardised residuals and studentised residuals, respectively. These are discussed below.

The Standardised Residual

The **standardised residual**, r'_i is found by dividing the residual by an estimate of its standard deviation, that is,

$$r'_i = \frac{r_i}{s_r \sqrt{1 - h_i}} , \tag{9.3}$$

where s_r is an estimate of the error variability, σ_E , and is obtained from the model fit as

$$s_r = \sqrt{\frac{\text{RSS}}{n - 2}}.$$

Note that $\text{RSS}/(n - 2)$ is also referred to as the **mean square error (MSE)**. — the MSE is discussed further in Section 9.2.

The standardised residuals can be obtained from the fitted model object using the `stdres()` function.

```
> stdres(rec.lm)
```

However, since the calculation of RSS includes the i -th residual r_i the numerator and denominator in Equation (9.3) are not independent. Consequently, they standardised residuals do not follow a t -distribution.

To overcome this problem, an estimate of s_e which is independent of r_i is needed, giving rise to the studentised residual which is discussed in Section 9.2.

The Studentised Residual

Consider refitting the model without the i -th observation, which results in an estimate of the error variability denoted by $s_{r(-i)}$. Then, r_i (from the original fit) and $s_{r(-i)}$ (from the leave-one-out fit) are now independent and consequently the **studentised residual** is given by

$$r_i^* = \frac{r_i}{s_{r(-i)}\sqrt{1 - h_i}}. \quad (9.4)$$

The r_i^* now follow a t -distribution with $(n - 1) - 2$ degrees of freedom.

These studentised residuals now have constant variance of 1 and are the preferred type of residual used in many model diagnostics. They can be obtained from the fitted model object using the `studres()` function.

```
> studres(rec.lm)
```

Model Diagnostics

Recall that there are several assumptions that need to be met for the linear regression model to be appropriate (Model (9.1)). In the following sections each of the four assumptions will be looked at in turn and a way to assess the assumption will be described.

Most of the methods described in the following sections use one of the three types of residuals from the model to help assess the

assumptions. These ordinary, standardised, and studentised residuals can be obtained from the fitted model with the functions `resid()`, `stdres()`, and `studres()`, respectively.

Independence of Errors

This assumption can only be assessed by knowing how the data were collected. In essence, this assumption tries to ensure that each data point is unique in its own right and that no two data points convey the same information. As an example of two data points which are not independent, consider the expenditure information obtained separately from people from the same family, *e.g.* husband and wife. It could reasonably be expected that both would come up with very similar, if not identical, values for the annual income and recreation expenditure. Consequently, the answers obtained from husband and wife would not be independent (of each other).

In contrast, consider the two families living in different parts of the country. Knowing the income and expenditure from the first family would generally not give us any information about the income and expenditure of the other family — the two families can be considered to be independent.

Constant Variance of Errors

The assumption of constant variance of the errors states that the variability of the errors remains the same, irrespective of the value of the independent variable x . Since the residuals are estimates of the true errors, plotting the residuals versus the corresponding values of x should give a plot which shows roughly equal scatter of the residuals for all values of x , *i.e.* a band of points with about constant width should be visible. This type of plot is known as a **residual plot** — using the studentised residuals instead of the ordinary residuals results in the **studentised residual plot**. For the least squares regression model of the recreation expenditure data the studentised residual plots is shown in Figure 9.2, which can be obtained using the following commands.

```
> plot(Tourism$income, studres(rec.lm),  
+      pch=20, col="blue", ylim=c(-3,3),  
+      xlab="Annual Income ($"),  
+      ylab="Studentised Residuals(rec.lm)")  
> abline(h=0)  
> abline(h=c(-2,2), col="grey")
```

From Figure 9.2 it can be concluded that the assumption of constant variability is appropriate since the residuals fall within a horizontal band which does not appear to change in width. Note that guides at ± 2 have been added to the plot. Based on the

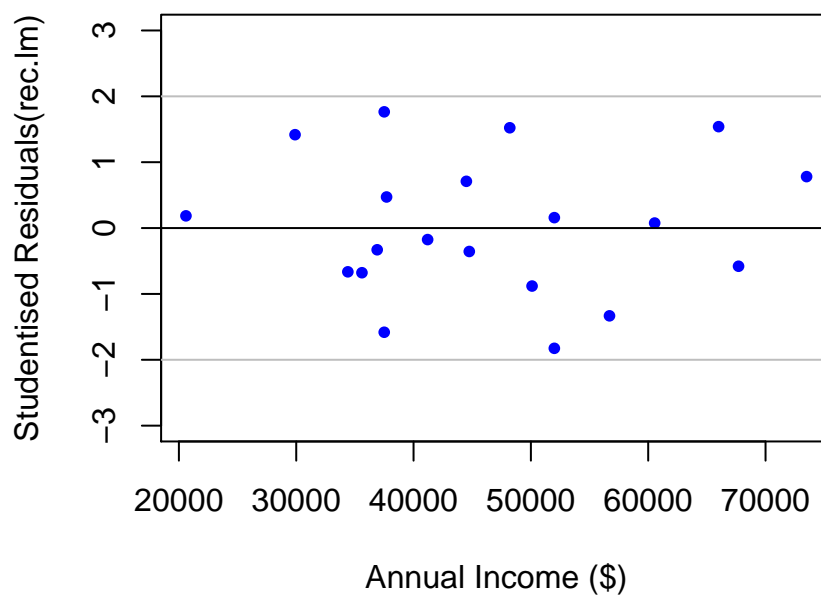


Figure 9.4: Studentised residual plot for the linear regression fitted to the Tourism data.

68-95-99.7% rule, approximately 95% of points should fall within those lines.⁵

To illustrate when this assumption is *not* met, consider the plots in Figure 9.2 — the top plot shows the scatter plot of Y versus X , while the bottom plot shows the corresponding studentised residual plot after fitting a least squares regression model to the data. In this second plot, a distinctive “*fanning out*” pattern can be observed, indicating that the variability of the residuals increases considerably with increasing values of X . This fanning pattern is also evident in the original scatter plot, but it is far more pronounced in the residual plot.

A Word of Caution:

Sometimes a “*diamond*” pattern can be observed in residual plots, showing less variability on the very left and right hand side and more variability near the middle. While this pattern may indicate a variability problem, it also often arises due to fewer observations being observed in the extremes and many more near the middle. If this is the case, then you will need to use your judgment as to whether this pattern is due to a lack of data or a change in variability.

Normality of Errors

The assumption of normality is important as it allows us to make inferences about the parameters, *i.e.* regression coefficients. As seen in [Chapter ?? — reference needed](#), the Normal Quantile-

⁵Note that this assessment is only approximate, since the studentised residuals actually follow a t -distribution.

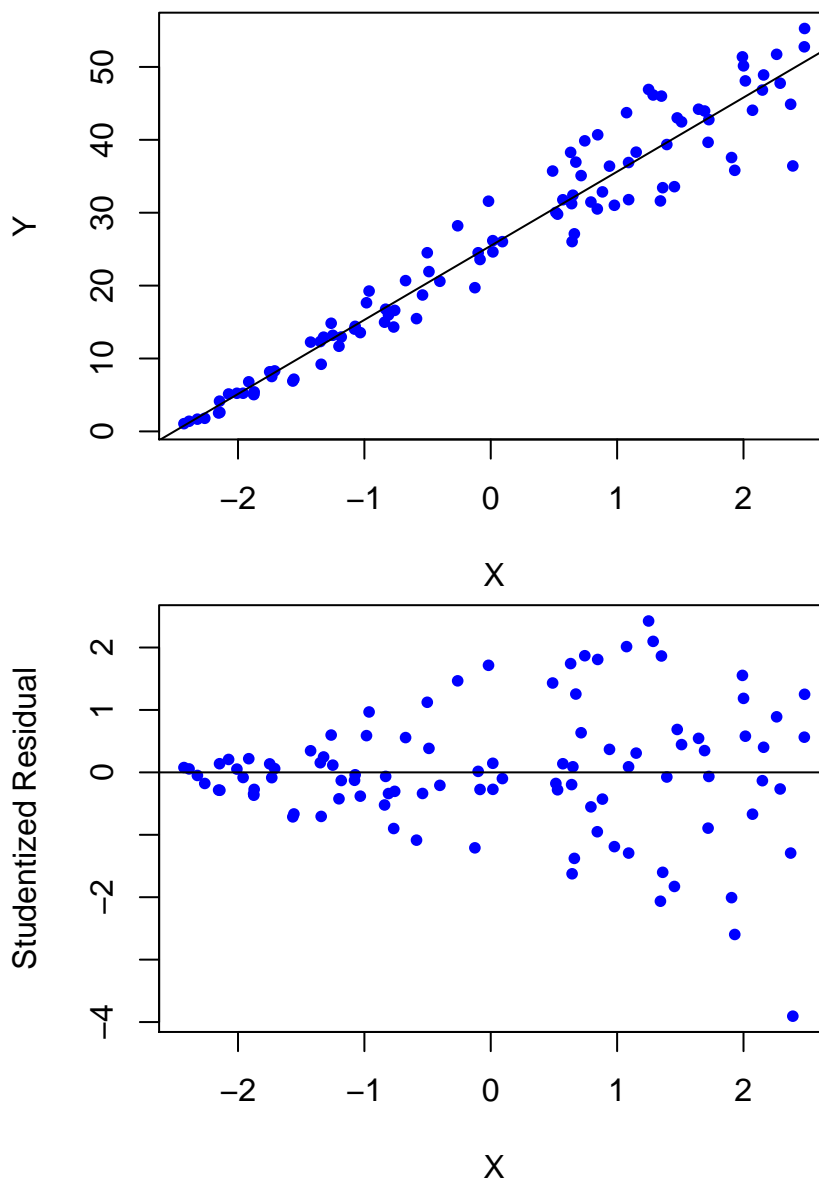


Figure 9.5: Example of non-constant variability; (top) scatter plot of fictitious data, and (bottom) corresponding studentised residual plot.

Quantile plot (`qqnorm()`) can be used to assess whether a data set is consistent with the assumption that the data are a random sample from a Normal distribution. So, for the recreation expenditure data, the commands

```
> qqnorm(resid(rec.lm), col="blue", pch=20, ylab="Residuals")
> qqline(resid(rec.lm))
```

result in the plot shown in Figure 9.2. From this plot it can be seen that the residuals fall generally fairly close to a straight line and thus it can be concluded that the assumption of normality is satisfied.

Note that while the residuals in this example fall very close to the straight line this may not always be the case. This leads to the question “*How close must the residuals be to the straight line in*

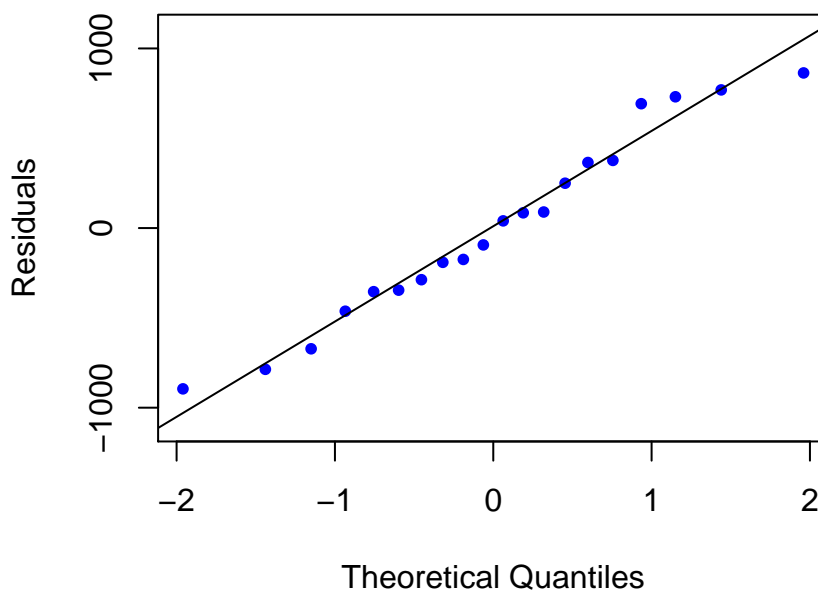


Figure 9.6: Normal quantile-quantile plot of the residuals from the linear regression fitted to the Tourism data.

order for the normality assumption to be acceptable?” There are, as usual, several ways of dealing with this question and two of the approaches are considered here — a test of significance and a graphical assessment.

Firstly, the **Shapiro-Wilk test** (`shapiro.test()`) for normality can be used to assess whether the residuals are significantly different from a Normal distribution. For example, performing the test with

```
> shapiro.test(resid(rec.lm))

Shapiro-Wilk normality test

data:  resid(rec.lm)
W = 0.9642, p-value = 0.6303
```

shows that the residuals can be considered to come from a normal distribution (due to the large P-value).

An alternative, graphical approach is based on the Q-Q plot — it tends to be favoured over the statistical test by many statistical professionals.

This approach is implemented in the `qq.plot()` function, which is part of the `car` package (`car::qq.plot()`). The idea behind this approach is as follows:

1. Generate many random samples of the same size as the data from a Normal distribution with the same mean and standard deviation as the data.
2. Sort each random sample in increasing order.
3. Across all random samples, calculate the 2.5 and 97.5 percentiles for the smallest value. Repeat for the second smallest value, third smallest value, *etc.*.

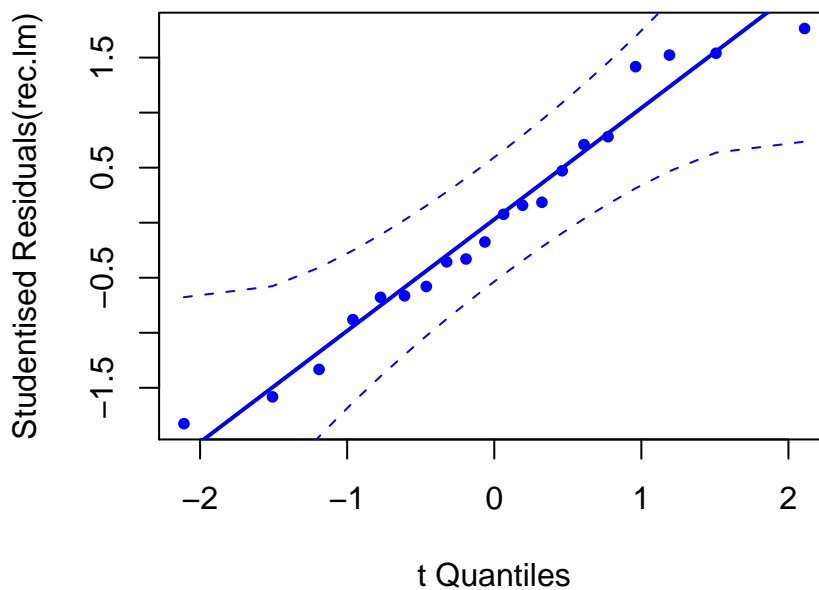


Figure 9.7: Normal quantile comparison plot obtained using the `qq.plot` function in the `car` package.

4. Plot the Q-Q plot as usual for the data
5. For each data point on the plot add the corresponding percentile points to the plot — these points indicate the **envelope**.

An alternative is to calculate theoretical envelopes based on the assumed distribution. However, irrespective of which of these two approaches is used, most data points should fall within the corresponding envelope, as this indicates that the data are consistent with the distributional assumption, *i.e.* the assumption of normality.

Since the residuals in a linear model do not all have the same variance, it is preferable to use the Studentised Residuals to generate these Q-Q plots. This is what the `qq.plot()` function does for a least squares regression model; it also use the theoretical envelope calculation approach by default. For example,

```
> qq.plot(rec.lm, labels=FALSE, col="blue", pch=20)
```

results in the plot shown in Figure 9.2. The option `labels=FALSE` has been provided to omit interactive identification of points.

Since all points fall inside their respective envelopes in Figure 9.2 it can be concluded that the assumption of normality is satisfied.

Straight line is appropriate

The studentised residual plot is also used to assess the assumption that a straight line is appropriate. However, this time interest lies in any patterns which suggest that a straight line is not appropriate, for example, any evidence of curvature in the residual plot.

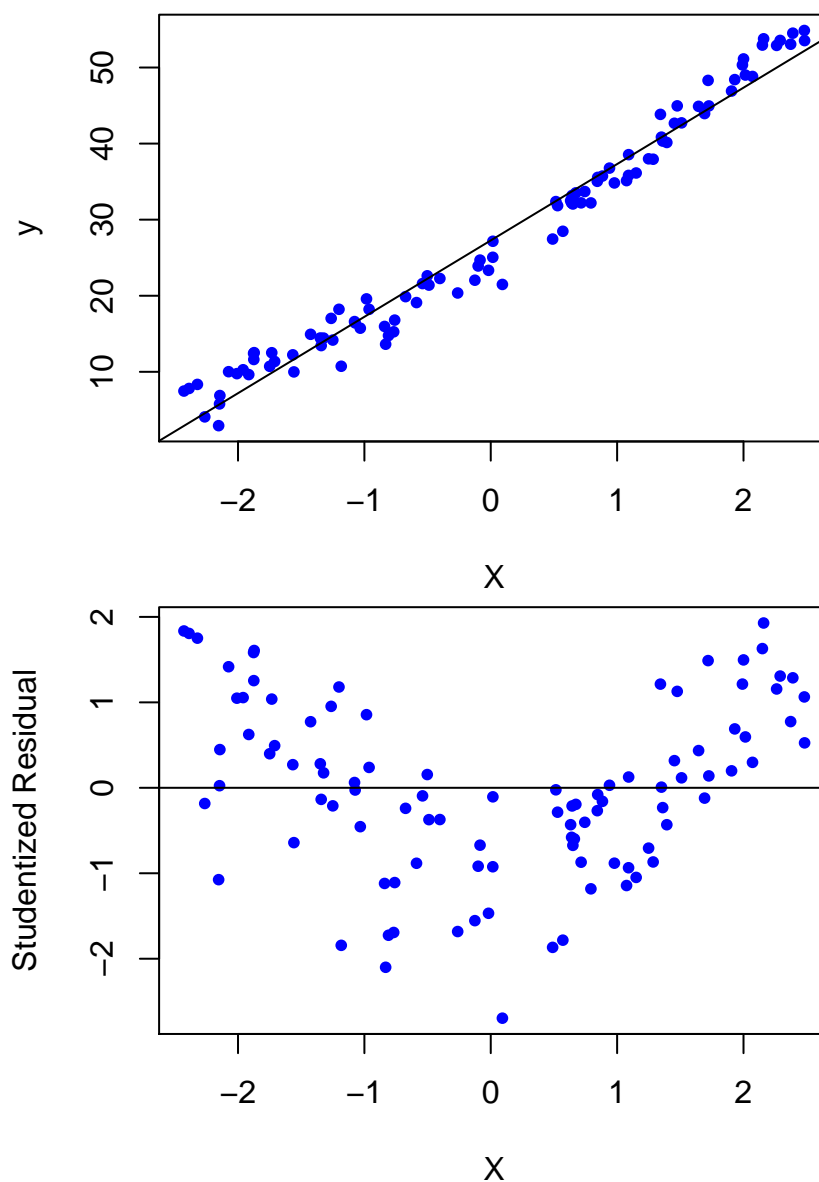


Figure 9.8: Example of nonlinearity; (top) scatter plot of fictitious data showing, and (bottom) corresponding studentised residual plot.

Similar to the discussion on constant variance, an example where the assumption is not met is useful, such as the plots in Figure 9.2. While the nonlinear nature of the data may be evident in the original scatter plot (top), this pattern is much more obvious in the studentised residual plot (bottom).

For the recreation expenditure data no pattern is apparent from the studentised residual plot shown in Figure 9.2. Consequently, it can be concluded that this assumption is also met.

Outliers

Outliers are data points which are not consistent with the rest of the data. These points usually indicate that something else is go-

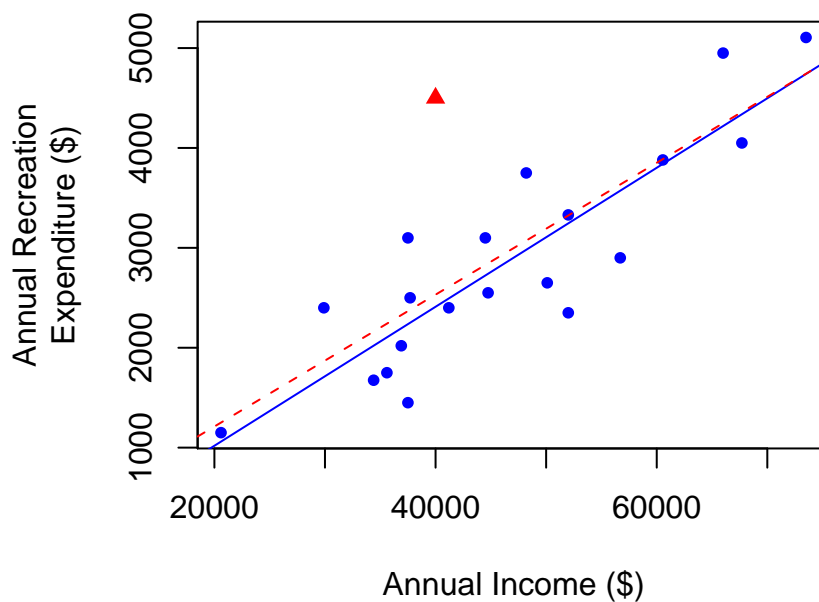


Figure 9.9: Example of an outlier and its effect on the least squares regression line. The solid blue and dashed red lines indicates LSR lines calculated excluding and including the outlier (red triangle).

ing on. In the context of a simple linear regression model, outliers are often defined as points with an unusually large residual. As such, a common rule-of-thumb is to class any point with a studentised residual outside the ± 2 range as a *potential outlier* requiring investigation. Note that these points must not automatically be deleted, but must be investigated to find the reason for why they are different. Finding the reason will allow you to either delete the point or change the model to better accommodate such points.

The reason for considering studentised residuals outside the ± 2 range as outliers is due to the fact that when the residuals are $N(0, 1)$ distributed then only 2.5% will fall below -2 and only 2.5% will fall above $+2$.⁶ Consequently, such points are unlikely, but not impossible. This gives further credence to only deleting points that can be identified as truly being different, as by chance 1 in 20 data points will give rise to large studentised residual purely by chance.

Outliers are generally located within the bulk of x values, but they have a rather different y -value than most other points (see Figure 9.2). As a consequence, they tend to shift the least squares regression line up or down a little, but do not have a major influence on the estimates of the regression coefficients.

A Word of Caution:

Some people think that because outliers are different from the rest of the data that outliers should be deleted. This is not the

⁶Note that the t -distribution is actually more appropriate for studentised residuals. However, the ± 2 bounds are still quick and useful guides, requiring no further calculation.

case! Outliers (and other data points) can only be removed if there is a good reason for why they should be excluded. For example, a data entry mistake is a good reason for excluding a point — correcting the mistake is even better.

In terms of the Tourism example, the outlier shown in Figure 9.2 may be due to a family who had saved up for some time to go on a big holiday.

From Figure 9.2 it can be seen that there are not studentised residuals which fall outside the ± 2 limits, and consequently no points require further investigation.

Leverage Points

Leverage points are data points which can have a large effect on the least squares regression coefficients. Unlike outliers, leverage points usually fall outside the bulk of x values. As a result they have a large hat value, h_i . The average hat value, \bar{h} , can be shown to equal $(k + 1)/n$, where k , the number of parameters, equals 2 in the simple linear regression model.

A common rule-of-thumb for hat values, due to (?), states that points greater than $2\bar{h}$ are to be investigated, but other cut-offs have also been suggested by (?). In particular, for small data sets a cut-off of $3\bar{h}$ may be more appropriate. Points falling above the cut-off are referred to as leverage points and they come in the following two varieties.

Good leverage points: These are leverage points which have a small studentised residual, *i.e.* one that falls within the ± 2 range.

Bad leverage points: These are leverage points which have a large studentised residual, *i.e.* one that falls outside the ± 2 range.

An example of good and bad leverage points is given in Figure 9.2 in the context of the recreational expenditure data. The good leverage point (green star) has little influence on the straight line — it fits the general linear pattern. The bad leverage point (red triangle) however does not fit the linear pattern and has a considerable effect on the straight line. While good leverage points *can* be investigated, bad leverage points *must* be investigated (see Section 9.2). This bad leverage point may indicate a data entry mistake, or that the linear relationship does not hold over all values of x , but only a limited range. For example, this bad leverage point may indicate that the relationship between income and recreation expenditure levels off near \$ 70,000-80,000 annual income.

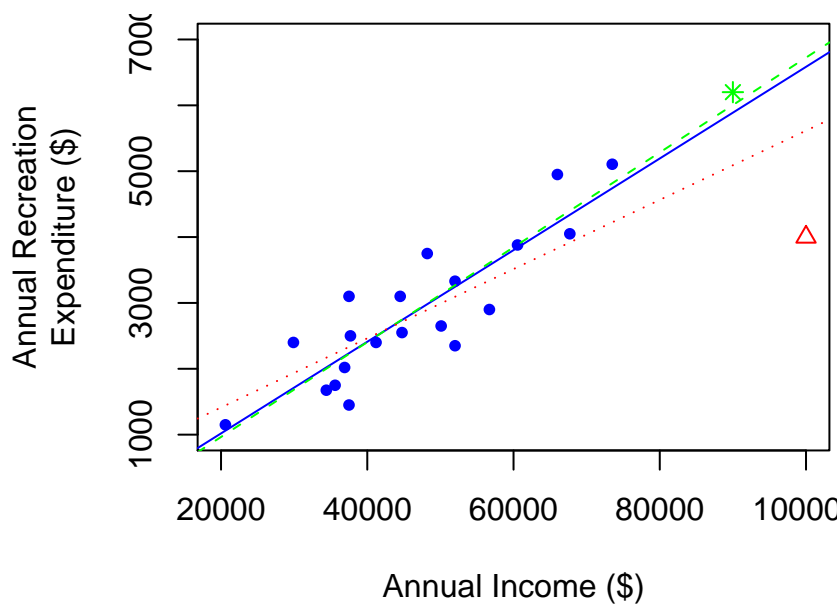


Figure 9.10: Example of leverage points and their effect on the least squares regression line. The solid blue line indicates the LSR line for the original data; the green dashed line shows the LSR line when a good leverage point (green star) is added; the red dotted line shows the LSR line when a bad leverage point (red triangle) is added.

The hat values can be obtained in R using the `hatvalues()` function in the `car` package `??`).

```
> hatvalues(rec.lm)

      1      2      3      4      5      6
0.05759070 0.05396938 0.05903502 0.15969912 0.05098969 0.07136380
      7      8      9     10     11     12
0.25950818 0.07236139 0.08039221 0.08297383 0.07236139 0.07549075
     13     14     15     16     17     18
0.05095811 0.09073491 0.12713737 0.05074214 0.10725323 0.05903502
     19     20
0.17951696 0.23888681

> c(2,3)*(2+1)/length(hatvalues(rec.lm))

[1] 0.30 0.45
```

And a graphical version is shown in Figure 9.2, which can be obtained with the following commands.

```
> plot(hatvalues(rec.lm), ylim=c(0,0.5),
+      xlab="Observation", ylab="Hat value",
+      col="blue", pch=20)
> abline(h=c(2,3)*(2+1)/length(hatvalues(rec.lm)),
+      col="red", lty=2)
```

Since none of the points fall above the $2\bar{h}$ cut-off it can be concluded that there are no leverage points in the recreational expenditure data set.

The Model Outputs

Two pieces of output are routinely examined when fitting a least squares regression model. These are obtained using the `anova()`

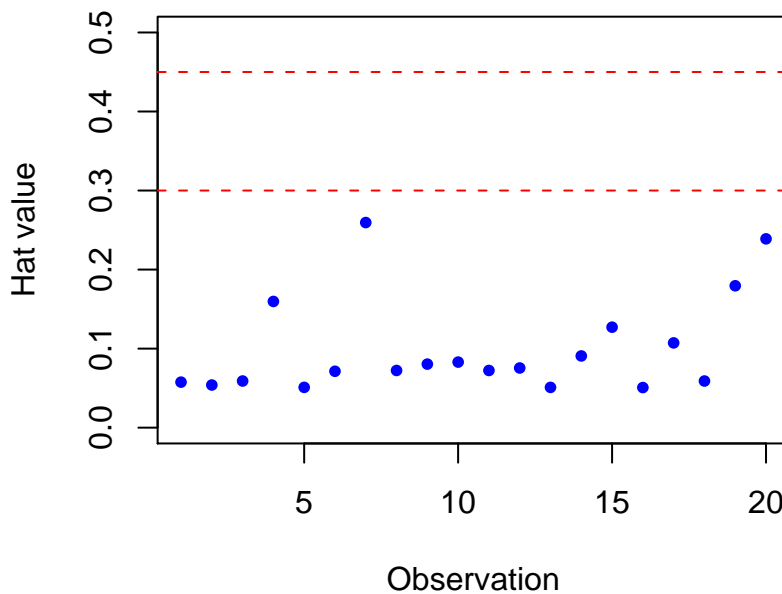


Figure 9.11: Leverage plot to assess the hat values of the observations in the Tourism data.

and `summary()` functions (see Section 9.2). Both pieces are discussed further in the following sections.

It should be noted that the output from the simple linear regression model should only be examined once all the assumptions are met.

The ANOVA Table

Recall from Model (9.1) in Section 9.2 that the model can be split into two parts — a systematic part and a random part. When fitting a model to data these can also be viewed as a *explained* and *unexplained* components, that is

$$\text{Data} = \text{Explained} + \text{Unexplained}.$$

Remember that not all of the y values are the same — there exists variability between the observations. Some of this variability can be explained through the straight line, *e.g.* people with more income spend more on recreation. However, not all the variability in the dependent variable can be explained — two families with similar incomes may spend different amounts on recreation — which makes up the unexplained component.

The **Analysis of Variance (ANOVA) Table** shows the split of the total variability in the dependent variable into explained and unexplained components.

```
> anova(rec.lm)
```

```
Analysis of Variance Table
```

```
Response: expend
```

```
      Df    Sum Sq Mean Sq F value    Pr(>F)
income    1 17010801 17010801  58.947 0.000000438 ***
Residuals 18  5194374   288576
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The following is an explanation of the various parts of the `anova()` output.

- **Response: expense**

This line states that the dependent variable in the least squares regression model is called **expense**.

- The **income** line gives all the information about the explained component of variability, *i.e.* the variability in **expense**, the dependent variable, explained by the linear regression on **income**, the independent variable.

DF: Denotes the degrees of freedom for the explained component; equals $k - 1$, where k is the number of parameters estimated as part of the model. In the case of a simple linear regression model the degrees of freedom are always 1 since there are 2 parameters estimated, intercept and slope.

Sum Sq: The amount of variability in **expense** explained by the linear regression, also known as the **Regression Sum of Squares** or **Model Sum of Squares**. Calculated as

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

Mean Sq: The regression or model mean squares (MS) equals the regression sum of squares divided by the regression degrees of freedom, 1.

F value: This is a test statistic used to assess whether a significant amount of variability in **expense** can be explained by the simple linear regression on **income**. It is calculated as

$$F = \frac{\text{Regression MS}}{\text{Residual MS}}$$

The F statistic has an F-distribution with $(k - 1, n - k) = (1, n - 2)$ degrees of freedom. The F statistic can only take on positive values and large values indicate that the linear regression explains a significant amount of variability. Here the value of the F statistic is 58.947.

Pr(>F): The P-value gives the probability of observing an F statistic as large, or larger, than the one observed

purely by chance, *i.e.* when there is no linear relationship between the **expense** and **income**. Here the P-value is less than 0.0001, indicating that there is very strong evidence for a linear relationship between **expense** and **income**.

- The **Residual** line gives all the information about the unexplained component of variability, *i.e.* the variability in **expense**, the dependent variable, which is not explained by the linear regression on **income**, the independent variable.

DF: Denotes the degrees of freedom for the unexplained or residual component; equals $n - k$, where k is the number of parameters estimated as part of the model. In the case of a simple linear regression model the degrees of freedom are always $n - 2$ since there are 2 parameters estimated, intercept and slope.

Sum Sq: The amount of variability in **expense** unexplained by the linear regression, known as the **Residual Sum of Squares**. Calculated as

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

Mean Sq: The residual mean squared error equals the residual sum of squares divided by the residual degrees of freedom, $n - 2$. Is often used as an estimate of the error variability, σ_E .

Given that the P-value in the ANOVA table is very small and since the diagnostics show no problems with the model assumptions, it can be concluded that the linear regression on annual income explains a significant amount of the variability in annual expenditure. Consequently, it is now OK to look at the equation of the straight line which was fitted, as well as additional summary statistics. These issues are addressed in the following section.

The Model Summary

Looking at the model summary is only appropriate when

1. the model assumptions have been met, and
2. the least squares regression line explains a significant amount of the variability in the dependent variable.

The summary is obtained as follows.

```
> summary(rec.lm)
```

```
Call:
lm(formula = expend ~ income, data = Tourism)

Residuals:
    Min       1Q   Median       3Q      Max
-895.09 -347.75 -26.82  368.02  863.70

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -372.645082  436.971630  -0.853    0.405
income       0.069572    0.009062   7.678 0.000000438 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 537.2 on 18 degrees of freedom
Multiple R-Squared:  0.7661,    Adjusted R-squared:  0.7531
F-statistic: 58.95 on 1 and 18 DF,  p-value: 0.000000438
```

The output of the `summary()` command consists of four pieces of information, namely **Call**, **Residuals**, **Coefficients**, and some additional **Summary Information**. What follows is an explanation of the various parts of the output.

Call: Shows the command used to generate the model.

Residuals: Five number summary statistics of the residuals.

Coefficients: A summary table for the coefficients β_0 and β_1 . Information about the intercept (β_0) is given in the line starting with **(Intercept)**, while information about the slope is shown in the line starting with **income**.⁷

The estimated intercept and slope are $\hat{\beta}_0 = -372.65$ and $\hat{\beta}_1 = 0.070$ (exact value is 0.069572), respectively. Consequently, the estimated regression line is

$$\text{Ann. Recr. Expenditure} = -372.65 + 0.070 \times \text{Ann. Income}.$$

How to interpret the coefficient estimates and how to make inferences about the corresponding regression parameters will be discussed in Section 9.3.

Summary Information: Additional summary output.

Residual standard error: This is the estimate of the error variability, σ_E , and is calculated as the square root of the residual mean squares, *i.e.*

$$s_r = \hat{\sigma}_E = \sqrt{\text{RMS}},$$

where the $\hat{\sigma}_E$ is used to denote that $\hat{\sigma}_E$ is an estimate of σ_E . This estimate is based on $n - k$ degrees of freedom (here 18).

⁷In general, this line is denoted with the name given to the independent variable.

Multiple R-Squared: Denoted as r^2 (the squared of the correlation coefficient) and sometimes referred to as the **coefficient of determination**. It gives the proportion of the total variability in the dependent variable y which is explained (or accounted for) by the linear regression on the independent variable x . It can be shown that

$$r^2 = 1 - \frac{\text{Residual SS}}{\text{Total SS}} = \frac{\text{Regression SS}}{\text{Total SS}} ,$$

where **Total SS** (TSS) denotes the total sum of squares and is given by

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

In the example output above, $r^2 = 0.7661$ and indicates that 76.61% of the variability in annual recreation expenditure can be explained by the linear regression on annual income.

Adjusted R-squared: This is an “adjusted” form of the r^2 statistic. It is used in multiple linear regression problems, *i.e.* when there are more than one independent variables.

F-statistic: The F-statistics used to assess whether the least squares regression line explains a significant amount of the variability in the dependent variable; same value as the F-statistic in the `anova()` output.

p-value: The P-value corresponding to the F-statistics; same as the P-value in the `anova()` output.

In addition to looking at the tabulated output, it is often useful to see the scatter plot together with the fitted regression line. Such a plot is shown in Figure 9.2, which can be generated as follows.

```
> plot(Tourism$income, Tourism$expend,
+       ylab="Annual Recreation\nExpenditure ($)",
+       xlab="Annual Income ($)", pch=20, col="blue")
> abline(coef(rec.lm))
```

Now that the output generated by `summary()` has been explained, it is time to look at how this output can be used. In particular, attention will be paid to making inferences about the regression parameters, β_0 and β_1 .

9.3 Making inferences

In the previous section, the typical output from a least squares regression has been explained. In this section the pieces of output

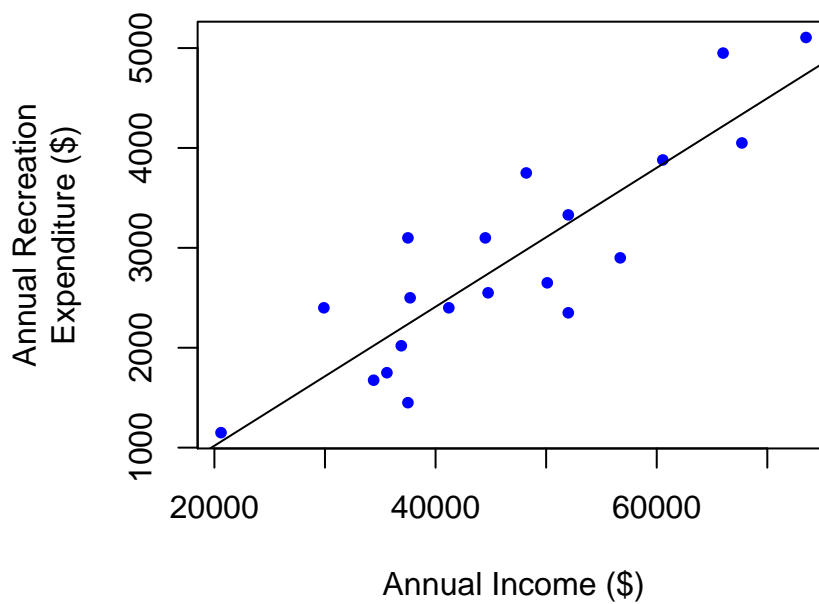


Figure 9.12: Scatter plot of annual expenditure on recreation versus annual income with least squares regression line added.

are used to draw conclusions about the underlying parameters. This work will require you to be familiar with details of hypothesis testing ([Reference to appropriate chapter](#)).

The Intercept

Repeated below is the line from the `summary()` command dealing with the intercept.

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -372.645082  436.971630  -0.853    0.405
...
```

As stated in Section 9.2, the estimate of the intercept is $\hat{\beta}_0 = -372.65$ (\$), and indicates the expected annual expenditure on recreation for a family with an annual income of \$ 0. Clearly, this seems like a silly result. However, it must be kept in mind that the minimum annual income observed in the data set is \$ 20,600, which is a long way from \$ 0, and hence such **extrapolation** must be treated with caution. There is no guarantee that the straight line model holds for annual incomes below \$ 20,000.

Furthermore, it must be kept in mind that -372.65 is only an estimate for the underlying population parameter, β_0 — a different estimate would be obtained if a different sample had been collected. Consequently, $\hat{\beta}_0$ has a standard error, $\sigma_{\hat{\beta}_0}$, denoting the variability of the estimate of the intercept. Under the assumptions of the linear regression model (9.1), it can be shown that the distribution of the $\hat{\beta}_0$ is given by

$$\hat{\beta}_0 \sim N(\beta_0, \sigma_{\hat{\beta}_0}^2),$$

where

$$\sigma_{\beta_0} = \sigma_E \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2}} .$$

The estimate of the standard error σ_{β_0} is denoted by s_{β_0} . It is given in the output above under the column heading **Std. Error** and takes the value 436.97 (\$) in the example.

Based on the general hypothesis testing framework

$$\text{Test statistic} = \frac{\text{Estimate} - \text{Hypothesised Value}}{\text{St. Error of Estimate}} \quad (9.5)$$

the intercept β_0 can be tested against any hypothesised value $\beta_{0,(0)}$ using the hypothesis

$$\begin{aligned} H_0 &: \beta_0 = \beta_{0(0)} , \\ H_a &: \beta_0 \neq \beta_{0(0)} . \end{aligned}$$

By default, most software packages will use $\beta_{0(0)} = 0$. This means that the null hypothesis states that a family with zero annual income has zero recreational expenditure, *i.e.* the regression line goes through the origin. The two-sided alternative states that the straight line does not go through the origin.⁸

Consequently, the test statistic is given by

$$t = \frac{-372.65 - 0}{436.97} = -0.853 ,$$

which is shown in the model summary output under the column heading **t value**. It can be shown that the test statistic follows a t -distribution with $(n - 2)$ degrees of freedom. Consequently, the two-sided P-value is 0.405, which is also shown in the summary output under the heading **Pr(>|t|)**. Consequently, it can be concluded that there is no evidence to suggest that the intercept β_0 is significantly different from zero. Of course, a value other than 0 can be used as the hypothesised value, but the associated test statistic has to be calculated by hand.

In addition to significance tests, confidence intervals with desired confidence levels can be calculated. These are based on the general confidence interval framework, *i.e.* a $100(1 - \alpha)\%$ confidence interval is given by

$$\text{Estimate} \pm t_{(1-\alpha/2)}^* \times \text{St. Error of Estimate} , \quad (9.6)$$

where $t_{(1-\alpha/2)}^*$ is the $1 - \alpha/2$ quantile (critical value) from the t -distribution with $n - 2$ degrees of freedom (in the simple linear regression context). In the example, the critical value for a 95% confidence interval ($\alpha = 0.05$) is found from the $t(18)$ distribution using the `qt()` function.

⁸For this example, a one-sided alternative hypothesis may actually make more sense. However, the two-sided alternative will apply more generally.

```
> qt(1-0.05/2, df=18)
```

```
[1] 2.100922
```

Consequently, the 95% confidence interval can be calculated. Alternatively, the function `confint()` can be used, which by default calculates a 95% confidence interval, as shown below.

```
> confint(rec.lm, parm="(Intercept)")

                2.5 %      97.5 %
(Intercept) -1290.68841071 545.39824669
```

It can be concluded that the 95% confidence interval for the population intercept is $(-1290.69, 545.40)$. It is worthwhile to note that this interval includes the value $\$ 0$, which of course is consistent with the findings of the significance test above.

In this example the intercept is not of particular interest, but this is not always the case. In particular, by rescaling the data it is possible to turn an otherwise uninteresting intercept into a useful piece of information. For more details on this see [Exercise ??](#).

The Slope

The slope in a regression model is often considered to be of great interest as it conveys how quickly the dependent variable changes in relation to the independent variable, and as such determines as to whether a regression line is useful.

Repeated below is the line from the `summary()` command dealing with the slope.

```
Coefficients:
                Estimate Std. Error t value    Pr(>|t|)
...
income         0.069572    0.009062   7.678 0.000000438 ***
```

From this output the estimate of the slope is $\hat{\beta}_1 = 0.070$ (Rec. \$/Income \$). This indicates that for every $\$ 1$ increase in annual income an average increase of $\$ 0.070$ in annual recreation expenditure is expected. Note that the slope only refers to the average increase — it does not imply that if a family were to receive a pay rise of $\$ 1000$, that they would spend an additional $\$ 69$ on recreation. In fact, they may spend more, or less.

Similar to the discussion on the intercept, the value 0.070 is only an estimate of the true, underlying population slope β_1 . Under the assumptions of the linear regression model (9.1), it can be shown that the distribution of the least squares estimate of the slope is given by

$$\hat{\beta}_1 \sim N(\beta_1, \sigma_{\beta_1}) ,$$

where

$$\sigma_{\beta_1} = \frac{\sigma_E}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}} .$$

The estimate of the standard error σ_{β_1} is denoted by s_{β_1} which takes the value 0.00906 (Rec. \$/Income \$) and is given under the column heading **Std. Error**.

Using the general hypothesis framework (9.5), the slope β_1 can be tested against any hypothesised value of interest $\beta_{1(0)}$, using the hypotheses

$$\begin{aligned} H_0 &: \beta_1 = \beta_{1(0)} , \\ H_a &: \beta_1 \neq \beta_{1(0)} . \end{aligned}$$

By default, most statistical software packages will test the slope against a hypothesised value of $\beta_{1(0)} = 0$, *i.e.* test whether there is a significant linear relationship between the dependent and independent variables. The result from this test is shown in the output above, that is, the test statistic is

$$t = \frac{0.069572 - 0}{0.009062} = 7.677 .$$

Note that the `summary()` output is based on more decimal places and hence differs slightly from the manual calculation (7.678 versus 7.677). The P-value associated with this test is also obtained from the t -distribution with $n - 2 = 18$ degrees of freedom, and is found under the heading $\Pr(>|t|)$. Here the P-value is much less than 0.0001, indicating that there is very strong evidence that the slope is different to zero.

Note:

The P-value from the output above is identical to the P-value given in the ANOVA table (see Section 9.2). This is no accident. In fact both have to be the same as they perform the same test, but in different ways. Note also that the F -statistic shown in the ANOVA table is the same as the square of the t -statistic shown in the output above, *i.e.* $7.678^2 = 58.947$. Again, this is no accident. The mathematical relationship for this is beyond the scope and purpose of this book, but more details can be found in [Need a reference here](#).

Of course, other values may be of interest. For example, consider that an economic analyst claims that the rate of recreation expenditure is ¢ 7.5 per \$ of annual income. Are the data consistent with this claim? To test this claim, consider the following two hypotheses.

$$\begin{aligned} H_0 &: \beta_1 = 0.075 , \\ H_a &: \beta_1 \neq 0.075 . \end{aligned}$$

Consequently, the test statistic takes the value

$$t = \frac{0.070 - 0.075}{0.00906} = -0.552 .$$

The associated two-sided P-value is 0.588, and hence there is no evidence to suggest that the rate of recreational expenditure is different to ¢ 7.5 per \$ of annual income.

In addition to hypotheses tests, confidence intervals can be calculated in a similar way to those calculated for the intercept (see (9.6)). Again, the critical value is obtained from the t -distribution with $n - 2$ degrees of freedom. In the recreation expenditure example this critical value equals 2.10, and the 95% confidence interval can be calculated as

$$0.070 \pm 2.10 \times 0.00906 .$$

Alternatively, the confidence interval can be obtained with the `confint()` function.⁹

```
> confint(rec.lm, parm="income")  
  
          2.5 %      97.5 %  
income 0.05053422 0.08860932
```

Note that the 95% confidence interval includes the previously hypothesised value of 0.075, which is consistent with the findings from the hypothesis test.

Note:

It can also be shown that the slope is related to the correlation via the following equality

$$\hat{\beta}_1 = r \frac{s_y}{s_x} ,$$

where r is the correlation coefficient and s_x and s_y are the sum of squares of the x 's and y 's, respectively, that is,

$$s_x = \sum_{i=1}^n (x_i - \bar{x})^2 \quad \text{and} \quad s_y = \sum_{i=1}^n (y_i - \bar{y})^2$$

Confidence Intervals for the Mean Response

Part of the reason why a linear regression model may be fitted is that the estimated equation can consequently be used to make predictions. For the recreational expenditure example a family with an annual income of $x = \$ 69,500$ would be expected to spend on average

$$\hat{\mu}_{y|x} = -372.65 + 0.070 \times 69500 = 4492.35 .$$

⁹The argument `parm` can be left unspecified, in which case the `confint()` function will calculate the desired confidence interval for all terms in the model.

Note that $\hat{\mu}_{y|x}$ is used to denote the estimated ($\hat{\cdot}$) mean (μ) of recreation expenditure (y), given ($|$) an annual income of $x = \$69,500$. But since the regression line was estimated from data, there is of course some variability around this estimate. This variability depends on

- the random error variability, σ_E ,
- the sample size, and
- how far the new observation is from the centre of the data.

Since σ_E is unknown it is estimated by the residual variability s_r . From the Central Limit Theorem it then follows that a confidence interval has the general form shown in Equation (9.6), where the estimate is $\hat{\mu}_{y|x}$ and the estimated standard error of $\hat{\mu}_{y|x}$ is

$$s_{\hat{\mu}_{y|x}} = s_r \sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}} .$$

Consequently, a 95% confidence interval for the mean response at a new observation x is given by

$$(\hat{\beta}_0 + \hat{\beta}_1 x) \pm t_{(1-\alpha/2)}^* \times s_r \sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}} , \quad (9.7)$$

where $t_{(1-\alpha/2)}^*$ is the $1 - \alpha/2$ quantile from the t -distribution with $n - 2$ degrees of freedom. As stated above, the estimated standard error, and hence the confidence interval, depends on

Error variability (σ_E): Estimated by s_r — the larger the error variability, the greater the variability in the estimates and the greater the width of the confidence interval.

Sample size (n): Larger samples result in less variability in the estimated line, and hence the mean of the dependent variable at a new value of x ;

Distance from the mean ($x - \bar{x}$): The further away from the mean of previously observed x -values, the greater the uncertainty about where the mean response actually lies. Consequently, small changes in the slope will have a large effect in the mean response the further away from the mean the new observation lies.

Note:

For many practical purposes, an adequate, approximate 95% confidence interval is given by

$$\hat{\mu}_{y|x} \pm 2 \times \frac{s_r}{\sqrt{n}} .$$

This very rough interval is often adequate for ball-park estimates, but the exact interval (9.7) is preferable for general reporting.

While this interval can be calculated by hand,¹⁰ it is generally preferable to use a software package for this. In fact, R can easily perform the calculations as follows.

```
> predict(rec.lm, newdata=data.frame(income=c(69500)),
+       interval="confidence")

      fit      lwr      upr
[1,] 4462.593 3954.982 4970.204
```

The three components of the output, namely `fit`, `lwr` and `upr`, denote the mean response,¹¹ the lower and upper 95% confidence bounds, respectively. Consequently, a family earning \$69,500 per year is expected to spend, on average, between \$3,955 and \$4,970 on recreation annually.

By default, `predict()` uses a confidence level of 95%, but other levels can be specified via the `level` argument, *e.g.* a 90% confidence interval is found as follows.

```
> predict(rec.lm, newdata=data.frame(income=c(69500)),
+       interval="confidence", level=0.9)

      fit      lwr      upr
[1,] 4462.593 4043.62 4881.566
```

In addition to calculating a single confidence interval for a single new observation, confidence bounds for the regression line can be formed and plotted as follows. The results are shown in Figure 9.3.

```
> conf.data <- data.frame(income=1000*seq(18,75,by=1))
> conf.data <- cbind(conf.data, predict(rec.lm, newdata=conf.data,
+       interval="confidence"))
> plot(Tourism$income, Tourism$expend,
+       ylab="Annual Recreation\nExpenditure ($)",
+       xlab="Annual Income ($)", pch=20, col="blue")
> lines(conf.data$income, conf.data$fit)
> lines(conf.data$income, conf.data$lwr, col="red")
> lines(conf.data$income, conf.data$upr, col="red")
```

It is worthwhile to emphasize that this interval is a confidence interval for the mean response. So, if there were a large number of families with the same annual income of \$69,500, then the 95% confidence interval would cover the true mean annual expenditure 95% of the time. This however does not say anything about how much an individual family may be spending — this is addressed by the prediction interval.

¹⁰Notice that $\sum_{i=1}^n (x_i - \bar{x})^2 = (n-1)s_x$

¹¹Note that the mean is slightly different to the hand calculations above, which is due to rounding.

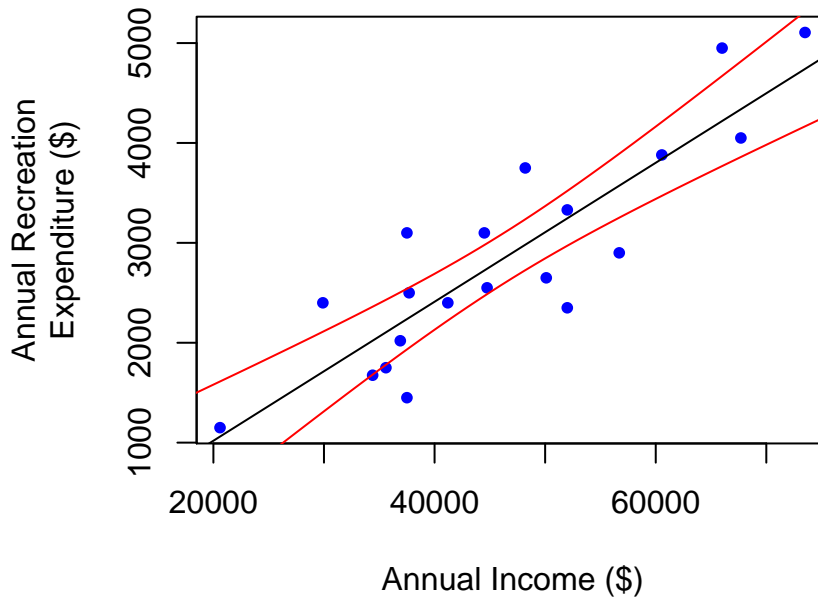


Figure 9.13: Scatter plot of annual expenditure on recreation versus annual income with least squares regression line and 95% confidence bands added.

Prediction Intervals for a New Observation

Unlike the confidence interval for the mean, the prediction interval for a new observation attempts to give information about the set of likely values of the dependent variable, given a new observation. In terms of the example, what are the likely values of recreation expenditure for a family with annual income of \$69,500?

As for confidence intervals, the prediction interval is centred about some mean point. This mean is given by

$$\hat{y}_x = -372.65 + 0.070 \times 69500 = 4492.35 .$$

Note that the value for \hat{y}_x is the same as the value for $\hat{\mu}_{y|x}$ used to denote the predicted mean response. This is no coincidence, since both use the estimated straight line to predict where the centre of the interval should be. The difference in notation between $\hat{\mu}_{y|x}$ and

\hat{y}_x is used to indicate that the confidence interval is used to make inferences about the mean of y given x , while the prediction interval is used to make inferences about the actual value of y given x .

The prediction interval is given by

$$(\hat{\beta}_0 + \hat{\beta}_1 x) \pm t_{(1-\alpha/2)}^* \times s_r \sqrt{1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}} , \quad (9.8)$$

where $t_{(1-\alpha/2)}^*$ is the $1 - \alpha/2$ quantile from the t -distribution with $n - 2$ degrees of freedom. Note that the prediction interval, like the confidence interval, depends on

Error variability σ_E : Estimated by s_r . The extra '1' under the square root sign is included to allow for the variability in the dependent variable in addition to the variability introduced through estimating the mean response at x .

Sample size (n) : Larger samples result in less variability in the estimated line, and hence the mean at a new x .

Distance from the mean ($x - \bar{x}$) : As for the confidence interval, small changes in the estimate of the slope can have a big impact in the predicted value when the new observation is far away from \bar{x} .

Note:

A rough ball-park estimate for the 95% prediction interval can be obtained using the formula

$$\hat{y} \pm 2s_r$$

The `predict()` function can also be used to obtain a prediction interval for a new observation. This is done by specifying the argument `interval="prediction"`.

```
> predict(rec.lm, newdata=data.frame(income=c(69500)),
+         interval="prediction")

           fit      lwr      upr
[1,] 4462.593 3225.092 5700.094
```

Consequently, 95% of the time, a family with annual income of \$69,500 spends between \$3,225 and \$5,700 on recreation annually. Note how much wider this interval is compared to the corresponding 95% confidence interval.

Similar to the scatter plot with confidence bands, a scatter plot with prediction bands can be generated (Figure 9.3), showing the likely recreation expenditures for the range of annual incomes.

```
> pred.data <- data.frame(income=1000*seq(18,75,by=1))
> pred.data <- cbind(pred.data, predict(rec.lm, newdata=pred.data,
+                                     interval="prediction"))
> plot(Tourism$income, Tourism$expend,
+      ylab="Annual Recreation\nExpenditure ($)",
+      xlab="Annual Income ($)", pch=20, col="blue")
> lines(pred.data$income, pred.data$fit)
> lines(pred.data$income, pred.data$lwr, col="red")
> lines(pred.data$income, pred.data$upr, col="red")
```

Notice how much wider these prediction bounds are — here all observations fall within them.

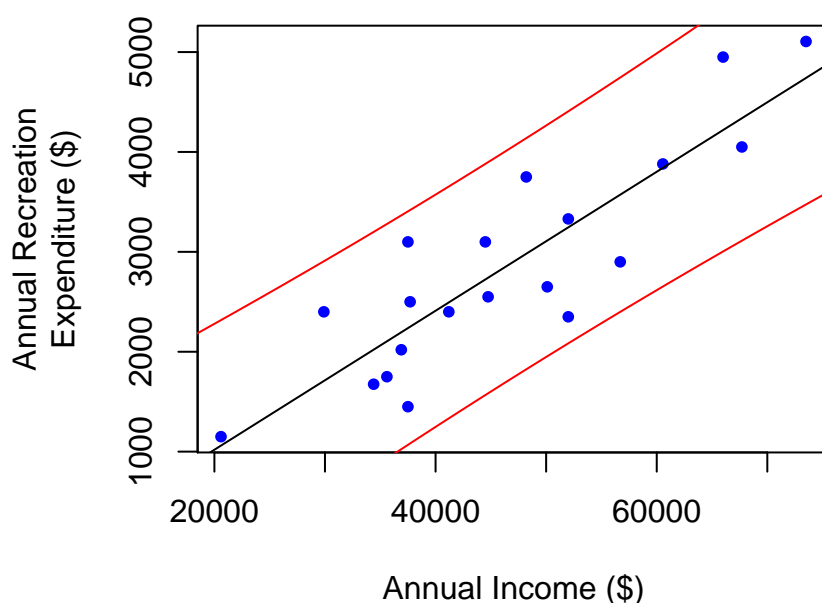


Figure 9.14: Scatter plot of annual expenditure on recreation versus annual income with least squares regression line and 95% prediction bands added.

9.4 Transformations

Fitting of a standard, simple least squares regression model was presented in Section 9.2. Of course, the example used until now has had the feature that the least squares regression model fitted to the data met all the model assumptions. But what can and should you do when things don't go quite to plan, *i.e.* when the assumptions are not met?

Well, let's look at the assumptions in turn and decide on how they can be tackled in order to get a valid model.

Dealing with Violated Assumptions

Recall from Model (9.1) and from the discussion on diagnostics that there are four assumptions that need to hold for the least squares regression model to be valid.

1. The E_i are statistically independent of each other;
2. The E_i have constant variance, σ_E^2 , for all values of x_i ;
3. The E_i are normally distributed with mean 0;
4. The means of the dependent variable y fall on a straight line for all values of the independent variable x .

How to address an assumption which is not met is discussed in the following sections. It is difficult to give advice on which order to address multiple violated assumptions — this will often depend on

the individual circumstances. Nevertheless, independence of errors will generally be the first assumption which needs to be met. Next, constant variance and normality will need to be address — often these two assumptions may be related. Last, but by no means least, is the assumption that a straight line is appropriate. This assumption may at times also be related to the assumption of normality and constant variance.

Independence of Errors

As indicated in Section 9.2, the independence of the observations, and hence of the errors, can only be checked by knowing how the data were collected. If there is reason to believe that the errors are not independent, there is little that can be done using elementary methods — it's time to see a professional statistician for a serious attempt to rescue your data analysis.¹²

One common reason for lack of independence is due to data which are collected over time. In these cases you may be dealing with auto-correlated data and methods for dealing with auto-correlated observations are collectively known as time series methods. They are beyond the scope of this introductory chapter on regression.

Constant Variance of Errors

Lack of constant variability can be related to a lack of normally distributed errors, but not necessarily so, *e.g.* Figure 9.2. One way of dealing with non-constant variability is to weight the influence observations — observations which are considered to be more variable are given less weight in fitting the model than observations which are less variable. This however makes the interpretation of the output more difficult and is beyond the scope of this book, and a professional statistician should be consulted.

An alternative, and possibly preferable, approach is to transform the dependent variable. This approach will be discussed in more detail below.

Normality of Errors

As discussed in Section 9.2 a lack of normally distributed errors becomes evident through a Normal Quantile-Quantile plot which does not show points which fall along a straight line.

The main problem that arises from the lack of normality lies in the calculation of estimate of the variability of the random errors, σ_E .

¹²Ideally, you should have consulted with a professional long before the data were collected, namely at the planning stages. This can often save a lot of heartache later on.

This also affects the calculation of the residual sums of squares. Consequently, hypotheses tests, confidence and prediction intervals will also be affected, resulting in inappropriate conclusions.

The general remedy for this type of problem is to use **generalized linear models** which assume distributions other than the Normal distribution. However, these methods are also beyond the scope of this book and a professional statistician needs to be consulted.

Having said this, in some instances, lack of normality and constance of variance can be related, and both can be addressed simultaneously using a transformation (see example below).

Straight line is appropriate

As discussed in Section 9.2 patterns of curvature in the residual plot indicate that the straight line is not appropriate. The patterns of curvature can be accommodated in a multiple or polynomial regression model, but not in a linear regression model. **Multiple regression** models include more than one independent variable — a special case is the **polynomial regression**. Polynomial regression models include, in addition to the linear term x , higher order terms, such as a quadratic term (x^2), cubic term (x^3), *etc.* to allow for patterns of curvature.

Both extensions of the linear regression model are addressed in [Chapter ???](#).

However, a possible alternative can sometimes be found through transforming the dependent and/or independent variables.

Finding an Appropriate Transformation

As discussed in the previous sections, transformations may be useful in a number of circumstances. But, once a situation has been identified which suggest that a transformation may be useful, how do you decide *what transformation* should be used? Often, there are a number of options which give, more or less, equivalent results, but the subject matter may very well dictate which of these is to be preferred. For example, \log_{10} , \log_2 and $\ln = \log_e$ will likely work equally well, as they are all logarithmic functions. However, only \log_{10} is meaningful to microbiologists, while computer scientists may prefer \log_2 , yet statisticians tend to favour the natural log (due to mathematical reasons).

The following is a non-exhaustive list of suggested transformations which may prove useful:

Count data: Try a square root transformation, especially when the counts are low, *e.g.* number of phone complaints on a particular day or number of breakages per shipment.

Areas: Try a square root transformation, since areas have units such as ‘m²’ or ‘cm²’. The square root transformation will have the effect of linearizing, *i.e.* the length of a square with the same area.

Volumes: Try a cube root transformation, since volumes have units of ‘m³’ or ‘cm³’, even when they are recorded in ‘l’ or ‘ml’. Similar to the transformation of areas, the cube root transformation standardized the measurement to the length of the edge of a cube with the same volume.

Ratios: Try the inverse or reciprocal transformation. The result is still a ratio, but the numerator and denominator have been switched, *e.g.* transform fuel economy measure in *litres per 100 km* into *km per litre*, or even *miles per gallon*.

Microbiological count data: Use a log₁₀ transformation as this will also make sense to the microbiologists. In particular, a 1 log₁₀ reduction is equivalent to a 99% reduction in microbial counts, 2 log₁₀ reduction is equivalent to a 99.9% reduction in microbial counts, *etc.*

Financial data: Similar to microbiological data, log₁₀ is may be preferable for ease of interpretation.

Computing data: Many computing measurements, such as storage space or transmission rates, include powers of 2. Consequently, it will be natural to use a log₂ transformation.

Other: If all else fails or if there is no clear logical transformation, try the natural logarithm ln.

When both the dependent and independent variables may need transformation, *e.g.* when linearity is an issue, ?) suggested transforming the x and y values based on the shape of the bulge which can be observed in the scatter plot of the data (Figure 9.4). So, for example, if the bulge is pointing down and toward the left, a log¹³ or square root transformation for either x or y or both should be tried. Note that for the purpose of interpretation it is generally preferable to use the same transformation for both variables wherever possible.

For example, consider the microbiological concentration data in Table 9.2. The data were collected during a storage trial of a food product. The independent variable is time (in days) and the dependent variable is the Total Viable Count (TVC), recorded as coliform forming units (cfu) per gram. A scatter plot of the data is given in Figure 9.4.

¹³This could be any of the logarithmic transformations.

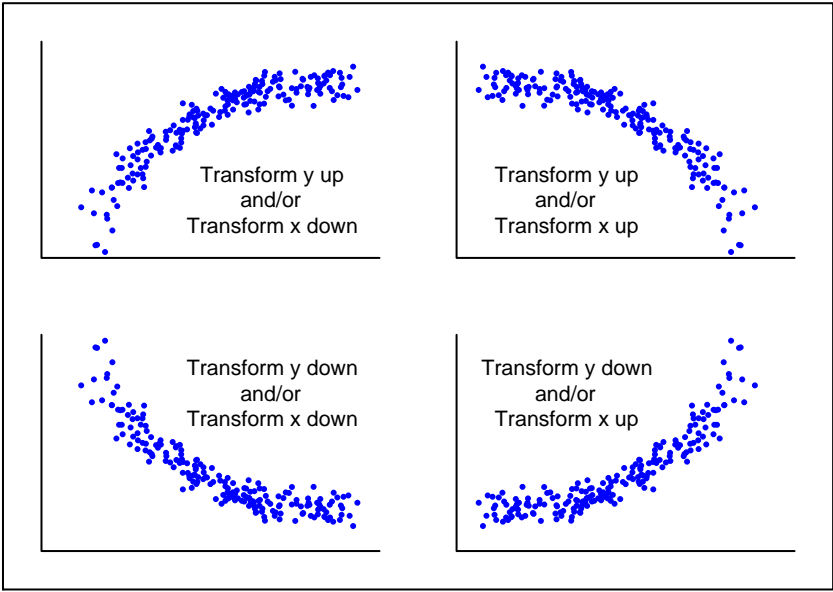


Figure 9.15: ?) suggested linearising transformations based on the shape of the *bulge* observed in the data. *Transforming up* involves squaring, cubing, etc., while *transforming down* involves taking logarithms, square roots, etc.

Time (days)	TVC (cfu/g)	Day	TVC
1	242	6	155,474
1	13	6	5697
2	128	7	716,068
2	113	7	200,870
3	105,174	8	121,623
3	351	8	13,364,791
4	6,988	9	2,051,185
4	169,079	9	2,158,450
5	3332	10	5,944,514
5	856,176	10	273,034,489

Table 9.2: Data on Total Viable Counts (TVC) of a food product over time. The data can be found in [micro.csv](#).

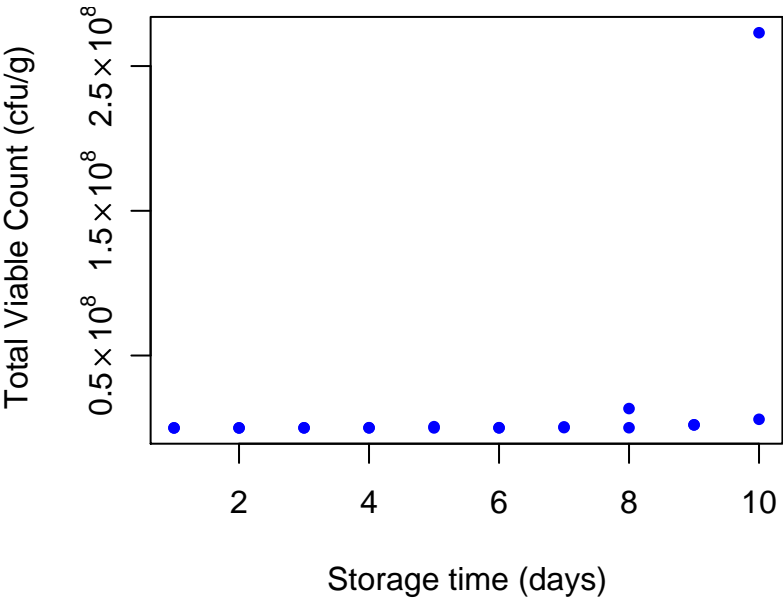


Figure 9.16: Scatter plot of microbiological concentration data collected during a storage trial of a food product.

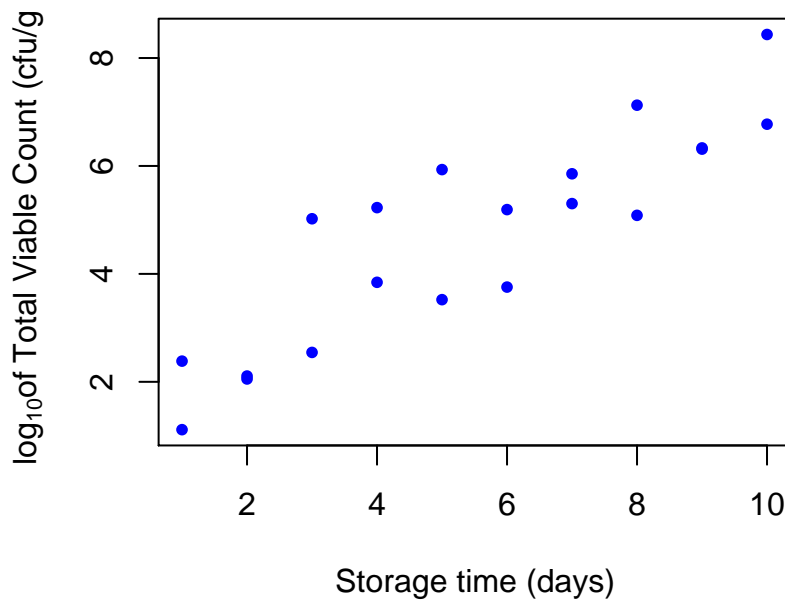


Figure 9.17: Scatter plot of \log_{10} transformed TVC data collected during a storage trial of a food product.

From Figure 9.4 it should be obvious that a linear relationship is not appropriate for this data — it's hard to tell what's going on since a single observation seems to dominate the choice of axes on the plot. Since the data are microbiological in nature a \log_{10} transformation for TVC may be appropriate. A scatter plot of the log-transformed data is shown in Figure 9.4. This plot shows that there is a reasonable strong linear relationship — $\log_{10}(\text{TVC})$ increases as storage time increases. The correlation is found to be $r = 0.883$. Clearly, the transformation appears to have remedied the problem and a transformation of Time does not appear necessary — this may need to be revised once a linear regression model has been fitted and the assumptions have been assessed.

Fitting the model

Once a suitable transformation has been found, the process of fitting the linear regression model is identical to that discussed in Section 9.2. For the TVC data, the fitting process is as follows.

```
> micro.lm <- lm(log10(TVC) ~ Day, data=Micro)
> anova(micro.lm)
```

Analysis of Variance Table

```
Response: log10(TVC)
      Df Sum Sq Mean Sq F value    Pr(>F)
Day      1 56.896   56.896   63.553 0.0000002576 ***
Residuals 18 16.115    0.895
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The ANOVA output indicates that there is a significant linear association between the $\log_{10}(\text{TVC})$ and Time. The next step is

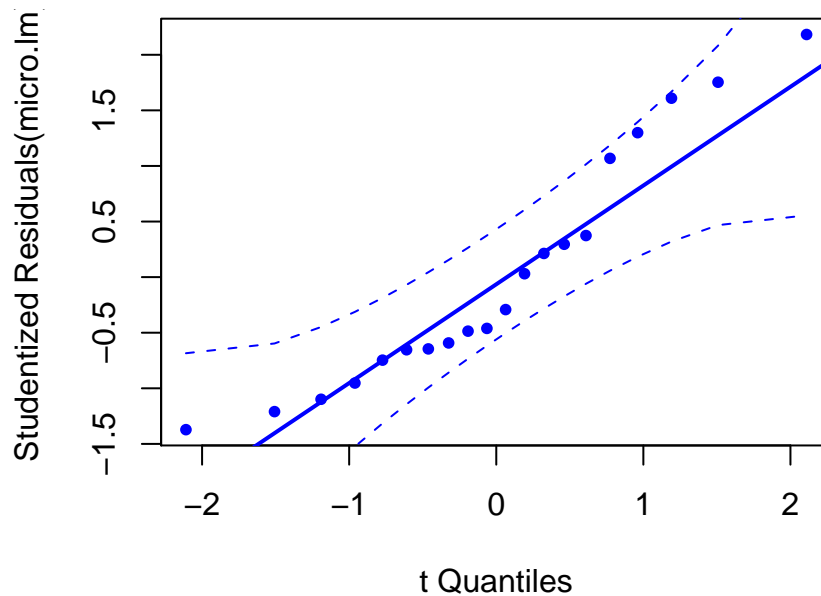


Figure 9.18: Normal quantile-quantile plot of the residuals from the linear regression fitted to the Micro data.

to assess the various assumptions and check for leverage points and outliers.

Model Diagnostics

The model is assessed using the same diagnostics as discussed in Section 9.2.

Firstly, from the trial design it can be assumed that the observations are independent since each observation was based on a separate unit of food and all units were stored under identical conditions.¹⁴

The Normal Q-Q plot shown in Figure 9.4 shows that all points fall within the bands and hence that the assumption of normality appears to hold. This is further confirmed when running a Shapiro-Wilks test (output not shown).

The assumptions of constant variance and linear relationship can be assessed with the studentised residual plot shown in Figure 9.4. Since there appears to be a fairly constant band of residuals, with no patterns being obvious, it can be concluded that these two assumptions also hold.

From the studentised residual plot it can also be seen that there are no outliers. In addition, an assessment of the hat values shows

¹⁴The fact that all units were stored in the same fridge has no effect here. However, care needs to be taken if this trial was replicated in two or more fridges, since observations which come from units stored in the same fridge may be more alike than units stored in different fridges. The fridges in this case are known as blocks (see [Chapter on Experimental design – Reference](#)) and their effect needs to be properly accounted for. A professional statistician should be consulted prior to embarking on such complicated trials.

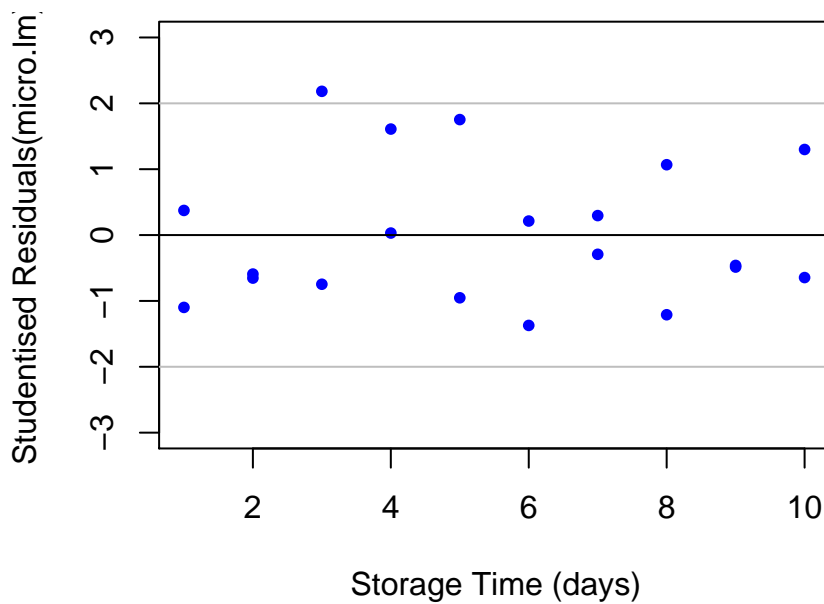


Figure 9.19: Studentised residual plot for the linear regression model fitted to the Micro data.

that there are no leverage points.

Consequently, since all assumptions appear reasonable, the parameter estimates can now be examined and used for prediction.

Interpretating the Output

The estimates of the model parameters are also found as previously discussed.

```
> summary(micro.lm)

Call:
lm(formula = log10(TVC) ~ Day, data = Micro)

Residuals:
    Min       1Q   Median       3Q      Max
-1.2342 -0.6117 -0.3458  0.4879  1.7937

Coefficients:
            Estimate Std. Error t value    Pr(>|t|)
(Intercept)  1.46659    0.45705   3.209    0.00487 **
Day           0.58722    0.07366   7.972 0.000000258 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9462 on 18 degrees of freedom
Multiple R-Squared:  0.7793,    Adjusted R-squared:  0.767
F-statistic: 63.55 on 1 and 18 DF,  p-value: 0.0000002576
```

From the model output it can be seen that the estimate of the intercept is 1.467, indicating that the average concentration of the product on day zero, the day of production, is 1.467 (\log_{10} cfu/g). Similarly, the average increase in microbial concentration per day of storage is 0.587 (\log_{10} (cfu/g) per day). Confidence intervals

are also found as before, *e.g.* the 95% confidence intervals for both parameters are found as follows.

```
> confint(micro.lm)

                2.5 %      97.5 %
(Intercept) 0.5063706 2.4268128
Day          0.4324625 0.7419697
```

Note: log-log transformations

A particularly important combination of transformations is the log-log transformation, *i.e.* when both the dependent and independent variables are log transformed.¹⁵

This transformation is frequently observed in economics, where the term **elasticity** is used to denote the slope of the regression line on the log-log scale and can be interpreted as the percentage change in the dependent variable associated with a 1% change in the independent variable (on the original scale).

For example, consider the price of a particular product. As the price is lowered, more product units tend to get sold, while as the price increases, less units tend to get sold. Often price and number of units sold are related linearly after log transforming both variables. An estimate of the slope of -5.1 indicates that a 1% increase in price is associated with a 5.1% *decrease* in the number of units sold.

Prediction

Care must be taken when trying to predict from a model which used transformed variables, which is why a separate section has been dedicated to this subject. The complication comes not when making the predictions on the transformed scale, but when those predictions are to be transformed back to the original scale for interpretation.

For example, the predicted microbial concentration on day zero is given by

$$\hat{y} = 1.467 + 0.587 \times 0 = 1.467 ,$$

which of course is just the value of the intercept. In addition, the 95% prediction interval is found as discussed in Section 9.3.

```
> predict(micro.lm, newdata=data.frame(Day=c(0)),
+         interval="prediction")

      fit      lwr      upr
[1,] 1.466592 -0.741022 3.674205
```

Up to this point, things are just as previously discussed, *i.e.* the predicted \log_{10} TVC concentration on the day of production is 1.467 \log_{10} cfu/g, and 95% of the time the actual concentration is expected to lie between -0.741 and 3.674 \log_{10} cfu/g.

In general though, people tend to be interested in the predictions on the original scale rather than the transformed scale.¹⁶ So, to get an estimate of actual concentration on day zero the transformation on the dependent variable must be reversed, *i.e.* the anti-log must be taken, resulting in a prediction of $10^{1.467} = 29$ cfu/g. However, this value is not the average on the original scale, but the median. The reason for this is that under a one-to-one transformation, such as log transformation, *quantiles* are preserved (?, ?). Consequently, the estimate of the median on the transformed scale, which equals the estimate of the mean due to the assumption of normality, transforms back to the median on the original scale. In addition, the 95% prediction bounds (both are quantiles/percentiles), as calculated on the transformed scale, transform back to 95% prediction bounds on the original scale.

In summary, the estimated median concentration after production is 29 cfu/g, with 95% of production units expected to contain between $10^{-0.741} = 0.2$ and $10^{3.674} = 4721$ cfu/g.

9.5 Summary

Least squares regression models are one of the most common statistical models. In this chapter you have learned that when the relationship between two quantitative variables is of interest, then the general pattern of the relationship can be observed using a scatter plot. When the pattern in the scatter plot looks linear then a linear regression model can be fitted via the method of least squares.

In addition, you've seen how to diagnose problems in the fit of the straight line and how transformations can be used to address some of the possible short-comings.

The output from a linear least squares regression model has been discussed and interpretation and inferential procedures for the intercept and slope have been demonstrated.

Finally, how the estimated regression line can be used to make predictions has been illustrated.

¹⁶Microbiology may be one of the few exceptions, since microbiologists are quite adept at dealing with log-ed data — most people however are not.

Exercises

1. Show that the point (\bar{x}, \bar{y}) is a pivot point, *i.e.* that the least squares regression line always goes through this point.
2. In many situations, an estimate of the intercept at $x = 0$ is of little interest, especially when the x values are from zero. In order to make the output from the regression model more meaningful and informative, a mean-corrected model is fitted. This type of model uses $x_i^* = x_i - \bar{x}$ instead of x_i as the independent variable.¹⁷ This is done as follows with the recreation data.
 - Calculate the x_i^* for all i , *i.e.*

```
> Tourism$income2 <- Tourism$income -
+   mean(Tourism$income)
```
 - Now fit the linear least squares regression model with the new variable as the independent variable.
 - Look at the estimates of the model. How do they differ from the those calculated earlier?
 - How would you interpret the intercept now?
3. Using the recreation data, divide the dependent and independent variables by 1000 — this changes the units for dollars to thousands of dollars. Refit the model using the new scaled variables.
 - a) What are the estimates of the intercept and slope now? What are their units? What is the interpretation of them?
 - b) What effect did re-scaling have on the model diagnostics?
4. Could really do with an exercise where students have to find a transformation.

¹⁷Adding and subtracting a constant in this way has not effect on the model itself. The estimate of the slope will be unaffected, while the intercept will be modified to correspond to $x_i^* = 0$.

References

Glossary

Bad leverage see *Leverage*

Coefficient of determination Applicable in regression problems and gives the percentage of variation in the dependent variable that is explained by the independent variable.

Correlation coefficient A quantitative measure which gives the strength and direction of the linear relationship between two variables.

Dependent Variable The variable to be predicted or explained by the independent variable in the regression model.

Explanatory Variable see *Independent Variable*.

Good leverage see *Leverage*

Hat value A numerical measure used to assess the leverage of a point (see *Leverage*).

Independent Variable A variable which is used to explain (systematic) changes in the dependent variable. The values of the independent variable are generally considered to be fixed values, possibly through experimental design.

Leverage Denotes the influence an observation has on the estimates of the coefficients in a least squares regression model. Observations with high leverage should be investigated to see whether they are also outliers (bad leverage) or not (good leverage).

Method of Least Squares A method for estimating the coefficients of the regression model. It works on the basis of minimizing the sum of squared residuals.

Regression Coefficients The simple regression model has two coefficients, the intercept β_0 and the slope β_1 . The estimates are denoted by $\hat{\beta}_0$ and $\hat{\beta}_1$, respectively.

Residual The difference between the actual value of the dependent variable and the value predicted by a statistical model. The residual will be negative (positive) when the prediction is greater (smaller) than the observed value.

Response variable see *Dependent Variable*.

Scatter plot A two dimensional plot that shows the relationship between two quantitative variables. The dependent variable is plotted on the vertical axis and the independent variable is plotted on the horizontal axis.

Simple Linear Regression A model that has one independent and one dependent variable. The simple linear regression attempts to explain the variation in the dependent variable by fitting a

straight line to the data.

Spurious Correlation A correlation that exist between two otherwise unrelated variables.