Chapter 5

Regression analysis

5.1 Linear regression analysis

5.1.1 Introduction – an example

The methods that have been examined so far are mostly descriptive and/or exploratory. All the methods covered have been widely used in archaeological applications. Regression methods – the focus of this chapter – have also been widely used (Baxter, 2003: 50–65). Such methods require a model to be formuated for the data, representing an important departure from most of what has gone before.

At its simplest, and stripped of context, the starting point in treatments in introductory texts is that of finding a 'best-fitting' straight line through a 'cloud' of points displayed in a bi-variate scatterplot. This is mathematically easy but, from the point of view of the average end-user, detailed knowledge of the mathematics is unnecessary. Other than in taught courses based on texts that illustrate hand calculations it is doubtful that anyone does anything other than use software to obtain results. This, once a data file is created, can be accomplished almost immediately. This frees the user to concentrate on the more interesting and challenging problems of model formulation and model interpretation.

These matters are discussed below and involve more use of mathematical notation than has hitherto been the case. To fix some initial ideas, an example is first presented. The data used are those of Table B.5, named `pmedwine` in `R` (from Robertson 1976), and are for six variables descriptive of the morphology of 49 post-medieval sealed wine bottles of known date.

It is clear that morphology changes over time and we shall suppose that interest lies in developing a model that can predict the date of undated bottles from their morphology. We shall further suppose that a simple (linear) regression model with just one variable as a *predictor* is sought. The full data set is examined in more detail in Example 1 of Section 5.2 where it is clear that body height, `BH` is likely

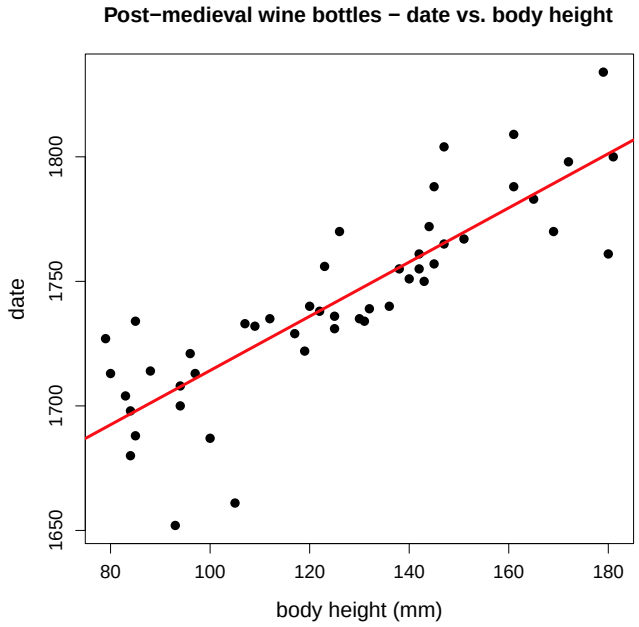

Figure 5.1: *A linear regression fit superimposed on a plot of date against body height for the data of Table B.5.*

to be the best single linear predictor of date.

Figure 5.1 is a plot of date against body height with a linear fit superimposed. It is straightforward to produce this. Omitting presentational arguments, and assuming that variables `BH` and `date` have been previously created, use

```
plot(BH, date)
fit <- lm(date ~ BH)
abline(fit)
```

where `lm` is the linear modelling function that fits the model required and saves the result, in this example, in the object `fit`. The `abline` function adds the fitted line to the plot. These are discussed in more detail in Section 5.4.

Once this is done, we would minimally like to know what the fitted line is and how well it fits the data. Execute the `summary` function, using `summary(fit)` to get the following (deleting some of the output).

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.605e+03  1.282e+01  125.18  < 2e-16 ***
BH          1.088e+00  9.977e-02   10.91  1.8e-14 ***
```

```
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 20.46 on 47 degrees of freedom
Multiple R-squared: 0.7168
F-statistic:   119 on 1 and 47 DF,  p-value: 1.801e-14
```

Under `Coefficients`, the estimated model for predicting date is given as

$$1605 + 1.088 \texttt{ BH}$$

with a goodness-of-fit of $R^2 = 71.7\%$, which is the `Multiple R-squared` expressed in percentage terms and rounded to one decimal place (See Section C.2.1 for more on R^2 and Sections 12.2.2 and 12.3.2 for the interpretation of p-values and the `F-statistic` that appear in the output.). This anticipates the fuller discussion provided in Sections 5.1.3, but in this context would usually be regarded as 'reasonable'. The main point about introducing the idea here is to show how easily such basic information is extracted.

With simple linear regression, other than provision of the line of best fit, inspection of the graph is often as or more informative than the basic output. For example:

- It is clear that a reasonable, though not perfect, linear fit will be obtained.
- It is evident that at the lower (less than 105) and higher (greater than 155) body heights prediction is less good than at the intermediate heights; that is, the variation about the fitted line is greater.
- The two earliest dates are badly over-predicted by the model and stand out as potential *outliers*.

Most of this is not evident from the numerical analysis to date. It can be taken further to address some of the issues involved. This is dealt with in Section 5.1.3 after a discussion of models, terminology and notation.

5.1.2 Regression models and notation

Begin with n observations on a *dependent* variable, y, and a single *independent* variable, x. A plot of y against x will suggest whether or not there is a relationship between y and x, its nature, and whether or not any of the data are unusual. Typically there will be some deviation from an exact mathematical relationship, attributable to what is conceived of as random *error* or *variation*. The most familiar model, the *simple linear regression model*, can be written

$$y = \alpha + \beta x + \varepsilon \tag{5.1}$$

where α and β are unknown *parameters* and ε is an unobserved *error* term[1].

The model specified is an *additive* one. By this it is meant that the model is *linear in the parameters* and the error term added as shown. The parameters, α and β, are the intercept and gradient (or slope) of the line. The intercept is the value of y when $x = 0$. The gradient is the change in y that occurs when there is a unit change in x, and is dependent on the units of measurement.

The simple linear model looks restrictive, but it can be extended in various ways, for example, to a two-variable regression model of the form

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \varepsilon \tag{5.2}$$

One special case is when $x_2 = x^2$, a quadratic term, so that model (5.2) becomes

$$y = \alpha + \beta_1 x + \beta_2 x^2 + \varepsilon \tag{5.3}$$

where the systematic component is a non-linear quadratic function that can be used to model this kind of non-linear pattern. The regression model is linear because of the linearity in the parameters. These are examples of *multiple regression* models. It can obviously be extended by adding cubic, quadratic terms and so on.

Another extension is when the data are split into two groups, and interest lies in seeing if and how relationships vary between groups. Define a variable z to have the value 0 for one group and 1 for the second group, an example of what is called a *dummy* or *indicator* variable. The model

$$y = \alpha + \beta_1 x + \beta_2 z + \varepsilon \tag{5.4}$$

has the effect of fitting two parallel lines through the data for the two groups. If a further term $(xz) = x \times z$ is defined the model

$$y = \alpha + \beta_1 x + \beta_2 z + \beta_3 (xz) + \varepsilon \tag{5.5}$$

simultaneously fits separate regressions for the two groups. The use of these models is illustrated, with further discussion in Example 3 of Section 5.2.

[1]The expression $\alpha + \beta x$ is the mathematical expression for a perfect straight line and is the *systematic* component of the model, in contrast to the *random* error; real data almost never follow such a line and to express this the error term is added to model the scatter about the line. The terms *dependent*, *independent* and *error* are hallowed by usage, and used here, but the terminology has been queried. The use of *dependent* might be taken to imply that the relationship is a causal one. This is sometimes the case, but if regression is used for description or prediction, for example, there is no implication of causality, and terms such as *regressor* and *regressand* have been used as an alternative. The term *predictor* has also been used above for the independent variable. Similarly, the 'error' need not be an error (of measurement or model mis-specification, for example) but may represent natural random variation about the dependent variable. The more neutral term, *disturbance*, is sometimes preferred.

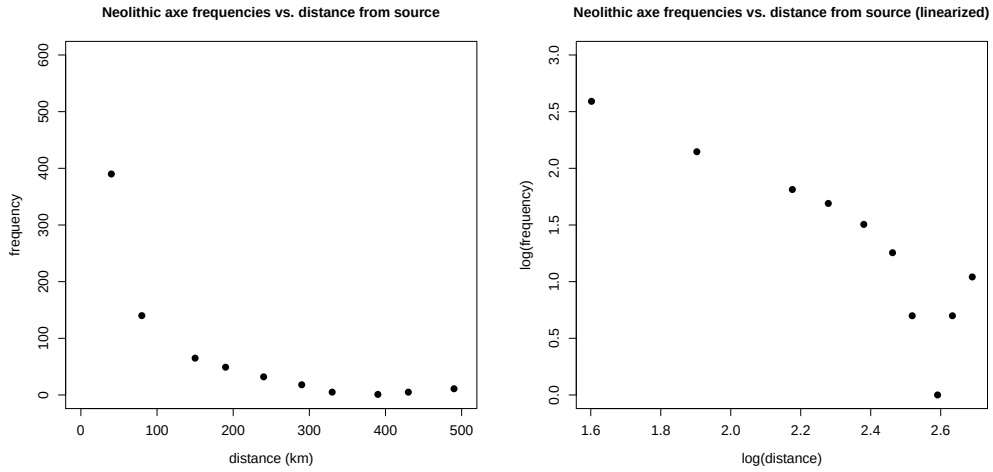

Figure 5.2: *The stone axe distance-decay data of Table 5.1 before and after a log-log transformation.*

To motivate further discussion two small data sets are shown in Tables 5.1 and 5.2. These use data of a similar kind, both measuring the frequency of artifact types found at different distances from a source of production or distribution center. Plots of frequency against distance typically show a *distance-decay* pattern with frequency declining to zero at some distance from the source. Linear models are inappropriate as they will result in negative predictions at some point. Two simple distance-decay models are explored below.

Table 5.1 is based on Cummins (1980) and shows the frequency of Neolithic stone axes at different distances from a distribution center. The data have been reconstructed from Figure 7 in Cummins, which is on a log-log scale.

Frequency	390	140	65	49	32	18	11	5	5	1
Distance	40	80	150	190	240	290	490	330	430	390

Table 5.1: *Frequency of neolithic stone axes at different distances (km) from a distribution center (Cummins, 1980).*

A linear regression model (5.1) will be fitted after data transformation where y is the logarithm of frequency and x is the logarithm of distance (Section 5.2, Example 2). Figure 5.2 shows plots of the data before and after the log-log transformation, and before undertaking the regression.

Although Cummins does not discuss this, the implicitly assumed model for the untransformed data has the form

$$y' = \alpha' x'^{\beta} \varepsilon' \tag{5.6}$$

where $y = \log y'$, $x = \log x'$, $\alpha = \log \alpha'$ and $\varepsilon = \log \varepsilon'$ show the correspondence with the notation of model (5.1). This is called a *power-law* model and is an example of a multiplicative model where ε' is a multiplicative error term. It is also an example of a *linearizable* model.

The appearance of the plot for the untransformed data suggests that a smooth model of distance-decay is reasonable. After the 'linearization' it is noticeable that there are departures from linearity at the longer logged distances. There is a clear outlier that may cause problems in fitting a linear model to the transformed data. This is discussed in detail in the continuation of the example in Section 5.2.

As a second and similar example that poses different problems in analysis, data from Morris (1994) are used. These are based on Figure 2A of that paper and are given in Table 5.2. The table shows the frequency of Late Iron Age pottery found at different distances from a production source. Morris did not examine the data in the way it is to be treated here.

Distance	4	18	21	22	23	27	30	34	36	43	52	62
Frequency	80	61	41	17	18	8	6	43	2	3	3	1

Table 5.2: *Frequency of Middle-Late Iron Age pottery at different distances (km) from a source (Morris 1994).*

In contrast to the power-law model, an *exponential decay* model will be used which has the form

$$y' = \alpha' \exp x^{\beta} \varepsilon' \tag{5.7}$$

which is linearizable after a (natural) logarithmic transformation where, in the simple linear regression model (5.1), $y = \log y'$, $\alpha = \log \alpha'$ and $\varepsilon = \log \varepsilon'$. As with the data from Cummins (1980) the data before and after transformation are shown in Figure 5.3.

The most obvious feature of this is a clear outlier in both plots (that Morris does not discuss). It would be legitimate from a model-fitting perspective to omit this from the outset and seek an explanation for it, but it is retained in some later analyses for illustrative purposes. The transformation to 'linearity' is not especially impressive, and this will be explored further in the continuation of the example in Section 5.2.

Finally, note that models of the kind, with a non-linear systematic component and additive error, such as

$$y = \alpha' \exp x^{\beta} + \varepsilon \tag{5.8}$$

cannot be simply linearized.[2]

[2]For those unfamiliar with logarithms, terms of the form ab can be transformed as $\log ab = \log a + \log b$ but this is not possible for $a + b$. The systematic component *can* be linearized, and this is an example of a *generalized linear model*, which is beyond the scope of the present notes.

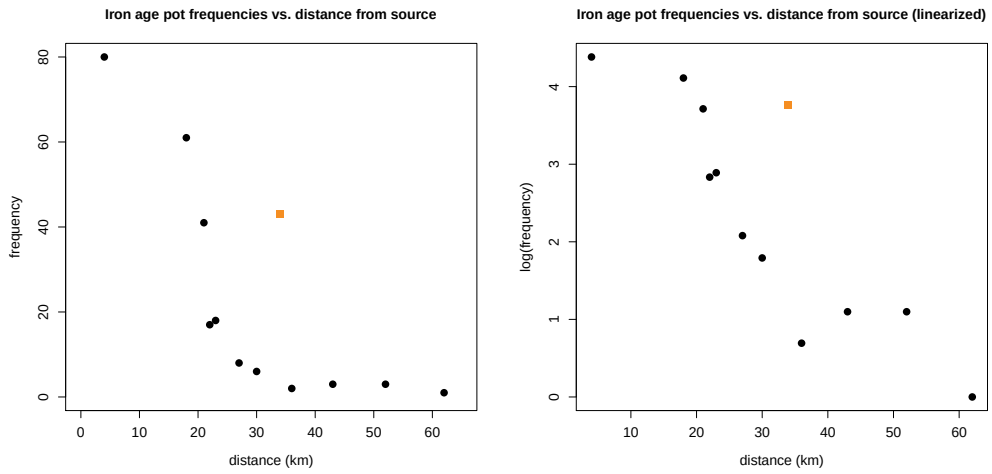

Figure 5.3: *The pottery distance-decay data of Table 5.2 before and after a log-transformation of the frequency. An obvious outlier is highlighted.*

5.1.3 Model checking

More notation and terminology

The general aim in regression analysis is to say something useful about the unknown systematic component in the model with, in simple linear regression, the focus often being on β. This requires *estimation* of the parameters that, in the light of the unknown errors, is 'sensible'. This can be done in more than one way.

It is important to distinguish between the 'theoretical' model, in which the parameters and error term are unknown, and the fitted model. The latter can be written as

$$\hat{y} = \hat{\alpha} + \hat{\beta}x \tag{5.9}$$

where $\hat{\alpha}$ and $\hat{\beta}$ are the estimated parameters and $\hat{y}$ is the fitted/predicted value of y using these estimates. With this in place $\hat{\varepsilon}$ defined as

$$\hat{\varepsilon} = y - \hat{y}. \tag{5.10}$$

is the estimated error or *residual*. The latter term is used to distinguish the observable estimated errors from the unknown true errors[3].

Parameters must be estimated. The default, often the only one in the texts and software used by archaeologists, is the *method of least squares*. This involves determining the estimates to minimize the sum of squared residuals. These estimates

[3]The use of Greek letters for unknown parameters, surmounted by a circumflex or 'hat' for their estimates, is a common convention; other conventions can be used.

have 'optimal' properties under certain error assumptions – most importantly that they have a normal distribution with constant variance and are independent. It is not, however, necessary for linear regression to be useful for the errors to be normally distributed, as is sometimes incorrectly asserted, so long as they are reasonably 'well-behaved'.

It is convenient here to introduce the idea of *correlation*. The data are to be regarded as a sample from a population, and the correlation coefficient in the population, ρ, is estimated as r (see Appendix C, and texts such as Shennan (1997: 140) for the mechanics of calculation). The correlation is a measure of the strength of the *linear* relationship between x and y, with 1 showing a perfect positive linear relationship and -1 a perfect negative one. In the context of simple linear regression the square of r, perhaps confusingly called R^2, often expressed as a percentage, is used as a measure of the 'success' of the regression, with values close to 1 'good' and close to zero 'bad'. It is called the *coefficient of determination* with the interpretation that it is the amount of variation in y 'explained' by variation in x. One reason for the notational distinction is that the definition of R^2 extends to linear models with more than one independent variable, and with the same interpretation.

Diagnostic statistics

It is desirable to check the validity of the error assumptions in assessing the fit of the model. While R^2 provides a global meaures of fit it can be more helpful to identify unusual observations or patterns in the data that compromise the adequacy of the model. A plethora of statistics have been developed for this purpose; many are just variations on similar themes and only some have garnered reasonably widespread use. The reader should be warned that for simple linear reqression the use of these statistics can be superfluous, since what they tell you can be obvious from graphical examination. They do, however, extend to more complex models where potential problems may be much less evident.

The error terms are unobservable but their properties are 'mimicked' by the observable residuals. Cases can be unusual becauses the value of the dependent variable is an *outlier*, having a 'large' residual. They can also be unusual because the value of the independent variable is 'extreme'; such cases are said to have high *leverage*. Outliers generally need some attention; cases with high leverage can actually be beneficial in fitting a model, but this depends on their other characteristics. Neither need affect the fitted model much; cases whose removal has an undue effect on the parameter estimates are said to have a large *influence*.

Dealing with leverage first, often denoted by h_i or m_i and lying between $1/n$ and 1, it measures how distant a case is from the centroid of the 'point' cloud of independent variables. For simple linear regression where the number of indepen-

dent variables $p = 1$, points of high leverage will stand out at the extremes of the scatter. Mathematically it is sensible to base a formal measure of the leverage of case i on the distance of the case from the mean, $(x_i - \bar{x})$, and the mathematics leads to a measure based on the square of this scaled to have a maximum of 1. For $p = 2$ recourse is needed to matrix algebra, but the measure obtained does a similar job. Rules-of-thumb exist for deciding what is an extreme leverage, but an index plot (of h_i against i) is often most useful.

Cases with high leverage are sometimes called *outliers*, but the term is best reserved for cases with large residuals. These can be defined in various ways. The *raw* or *ordinary* residuals, $\hat{\varepsilon}$, form a sensible starting point for detecting outliers, but are scale dependent. Rescaling is achieved by dividing $\hat{\varepsilon}$ by an estimate of its standard deviation s, where s^2 is an estimate of the assumed common error variance σ^2.

Let s^2 be this estimate using all the data, with $s^2_{(i)}$ an estimate omitting the ith case (which will differ for each i). Terminology is confusing. What have been called standardized residuals can be defined as $\hat{\varepsilon}/s$, and were what was available in older software. We follow the usage of Venables and Ripley (2002: 151) and define *standardized* residuals as

$$r_i = \hat{\varepsilon}/s\sqrt{(1 - h_i)}.$$

The term *studentized* residuals will be used for

$$t_i = \hat{\varepsilon}/s_{(i)}\sqrt{(1 - h_i)}.$$

See Cook and Weisberg (1982: 20) for the mathematical relationship between r_i and t_i. The r_i have also been called *internally studentized* residuals; the t_i have variously rejoiced in the names *studentized*, *externally studentized*, *jacknife* or *deleted-t* residuals, the last being used in the `MINITAB` package.

Although residuals mimic the properties of the errors they do not have exactly the same properties. In particular their standard errors depend on h_i in the manner indicated. If a case is an outlier it will inflate the estimate of s and $s_{(i)}$ will be somewhat smaller, so $t_i > r_i$. The general idea is that, for large enough samples, the distribution of the scaled residuals should mimic the assumed (usually) normal distribution of the errors. This means that, keeping the numbers simple, values in excess of 2 (in absolute value) can be regarded as 'unusual', and values in excess of 2.5 or 2.6 as 'very unusual'. For small samples the t-distribution can be used to define such 'rules-of-thumb'. For specialized situations more exact theory exists, but in practice it is generally more useful to inspect a plot of the scaled residuals against the fitted values rather than relying on rules-of-thumb.

An illustrative example for the post-medieval bottle body heights and date, following the regression illustrated in Figure 5.1, is provided in Figure 5.4, where

standardized and studentized residuals are contrasted with reference lines at ±2.5 shown (only the negative value being relevant here).

Figure 5.4: *Residuals from the regression fit of Figure 5.1.*

The two kinds of residual are largely coincident in their values and the standardized residuals have been 'jittered' to avoid overwriting the studentized residuals (see Section 5.4). The exceptions to this are two cases at the bottom left of the plot where the studentized residuals are larger than the standardized residuals, though both residuals suggest the cases are outliers. These are just the two bottles with the earliest dates, which are predicted to have somewhat later dates than the known values.

Of measures of influence available Cook's statistic

$$d_i = \frac{1}{p} r_i^2 \frac{h_i}{1 - h_i}$$

is the most commonly used. It is a function of residual and leverage statistics and large values will often have large values of one or both statistics, though this is not inevitable (remember h_i is bounded above by 1). Conversely, large values of one statistic will not necessarily give rise to a large d_i if the other is small. The statistic has an interpretation as the distance between parameter estimates with and without case i, or, equivalently, the distance between predicted values. An index plot is useful for making judgments about what is 'large'.

Figure 5.5 shows index plots of the leverage statistic h_i, and Cook's statistic for the regression of body height against date.

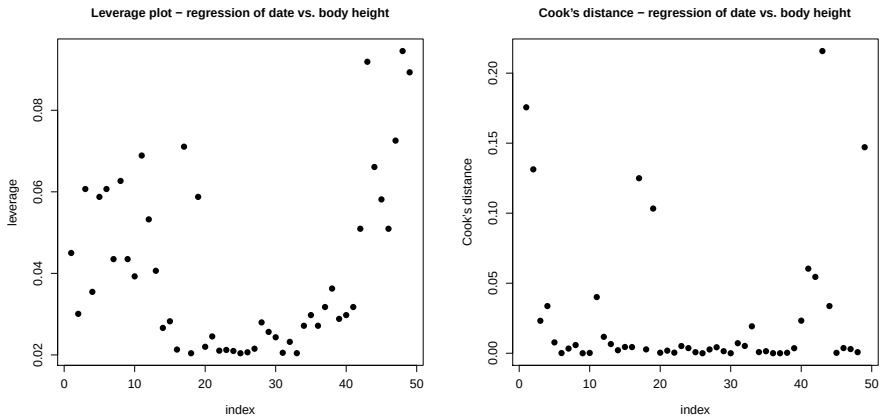

Figure 5.5: *Diagnostic index plots from the regression of* `Date` *against* `BH`. *Cook's distance to the right and leverage, h_i, to the left.*

The plots don't give too much cause for concern. The shape for h_i is to be expected as the index corresponds to a natural ordering (by date) with the extremes at either end. The plot for d_i has some values larger than others (some have to be), but nothing 'shouts out' as seriously extreme. Omitting the first two cases that were suggested as outlying in Figure 5.4 increases R^2 by about 5% without introducing further noticeable problems.

Other than the last case, the gap between the second and third earliest dates, of 19 years, is noticeably larger than for other adjacent pairs (ordered by date). The last case has nothing else unusual about it. Reporting results omitting the first two cases, on the basis that they are early and untypical, is a sensible option. A plot of the residuals against *Date*, which is sensible though not shown here, suggests even more starkly that the first two cases are untypical.

5.1.4 Inference

Intentionally, not too much has been said about traditional methods of statistical inference at this point. There are differing points of view about its value for archaeological data analysis; mine is that its importance has been exaggerated, partly for historical reasons (see Chapter 12). Some engagement with ideas is needed with model-based methods, however, if only to interpret output provided by software.

As far as regression goes, two concepts need to be distinguished, that of *statistical significance* of the regression fit, and *goodness-of-fit* which has been covered

in the foregoing discussion. For $p = 1$ the hypothesis that $\beta = 0$ (i.e. there is no linear relationship) is notionally of interest but formally testing this, using significance tests, is often a waste of time, since it will either be obvious that the regression is significant, or that it is too poor to be of substantive interest. Some acquaintance with *p-values* is desirable and discussed in context below and in Chapter 12. Repeating the analysis previously reported, with some editing

```
           Estimate SE  t-value  Pr(>|t|)
Intercept   1.605   13    125   0.0000 ***
BH          1.088  .01   10.9   0.0000 ***

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

Multiple R-squared: 0.7168
F-statistic:   119 on 1 and 47 DF,  p-value: 0.0000
```

allows us to infer from the very small p-values that, with the t-distribution as a reference, the hypotheses that $\alpha = 0$ and $\beta = 0$ are unsustainable. The F-statistic does the same job as the t-statistic for β (`BH`) in this instance but this is not the case if there are two independent variables or more. The conclusions are obvious from graphical inspection alone.

For $p > 1$ things are more complicated and interesting[4]. The significance of the regression is often obvious, but it is not always clear which variables are important, so testing the importance of subsets of variables is of interest. An illustration is provided in Example 3 in the next section.

5.2 Examples

Example 1 – Post-Medieval wine bottle dimensions

A *pairs plot*, also sometimes called a *scatterplot matrix*, of the post-medieval wine bottle data is shown in Figure 5.6 omitting `Type`, and `Height` which is the sum of neck height, `NH`, and body height, `BH`. This was obtained using the `pairs` function and was part of the preliminary data analysis that informed the choice of body height as the independent variable in the analyses that are the subject of Figures 5.1, 5.4 and 5.5.

It seems clear that `BH` will be the best single linear predictor of date, and a reasonable fit can be anticipated. The relationship of `Base` to `Date` is interesting; the pattern is distinctly non-linear but suggests a positive relationship for earlier dates and a negative one for later dates. The issues are pursued in Section 5.3.

[4] p as notation for the number of independent variable and p-value need to be distinguished.

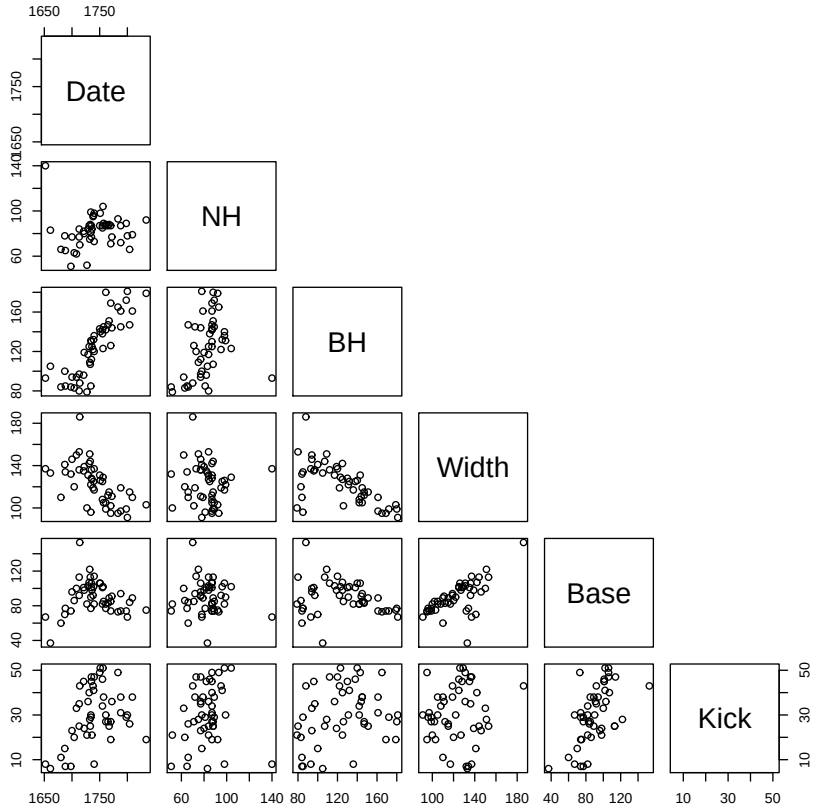

Figure 5.6: *Post-medieval wine bottles - a pairs plot for selected variables.*

Example 2 – Linearizable models

This continues the analyses of the data from Cummins (1980) and Morris (1994), begun in Figures 5.2 and 5.3. Specifically, a variety of fitted models are added to the plots previously presented, with commentary.

Cummins (1980) preferred log-log model, chosen presumably on the basis of visual interpretation, was to use just the first five observations. This was interpreted as a change in 'regime' at the associated distance and is shown as the dotted blue line in Figure 5.7. The dashed red line shows the fit using all the data.

Apart from the small number of observations used to fit the preferred model the interpretation is questionable on statistical grounds. Transforming the model using all the data back to the original scale results in a fit in the left-hand plot that is virtually indistinguishable from the data. On the log-scale, with all the data, the tenth observation for which the frequency is 1 is highlighted, with $t_i = -4.1$. If 1 is changed to 2 then $t_i = -2.1$. That is, the evidence for a 'boundary effect' resides

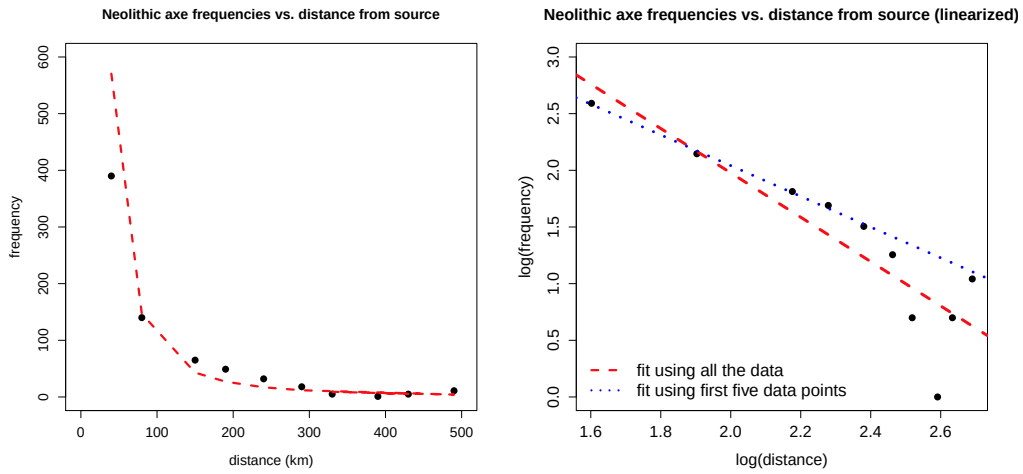

Figure 5.7: *The plot to the shows linearized model fits for the Cummins (1980) data, one using all the data and the other the first five observations. The plot to the left transforms the model fit using all the data back to the original scale. See the text for a full discussion.*

with a single observation, and its importance is almost certainly attributable to the small frequency and use of a log-transformation. The estimates of β differ by 0.24 with standard errors of 0.38 and 0.22. The estimates are not independent, but this is convincing evidence that they do not differ significantly as the standard errors would need to be much smaller to suggest a significant difference. The conclusion has to be that there is little statistical evidence for a boundary effect, with the attendant lesson that relying on visual interpretation in the presence of small samples is unsafe.

Turning to Morris (1994), following the initial presentation of the data in Figure 5.3, the right-hand plot of Figure 5.8 shows the fits for the exponential model, linearized as in model 5.7, with and without the outlier. Case 8 has $|t_i| = 2.79$. Omitting the outlier changes the fit from 73% to 84%, but has virtually no influence on $\hat{\beta}$, which can been seen visually. This is because the omitted case has very low leverage as it is centrally placed along the x-axis.

The exponential model was used for illustration. The linear transformation does not do a good job of 'straightening out' the plot, suggesting that a power-law model might have been better. If, omitting the outlier, this and the exponential model are fitted, the left-hand plot suggests the power-law model does a better job of fitting the data other than the first observation. Interestingly, for the linearized power-law model, case 1 has a larger value of $|t_i|$ than case 8, 2.46 compared to 2.36.

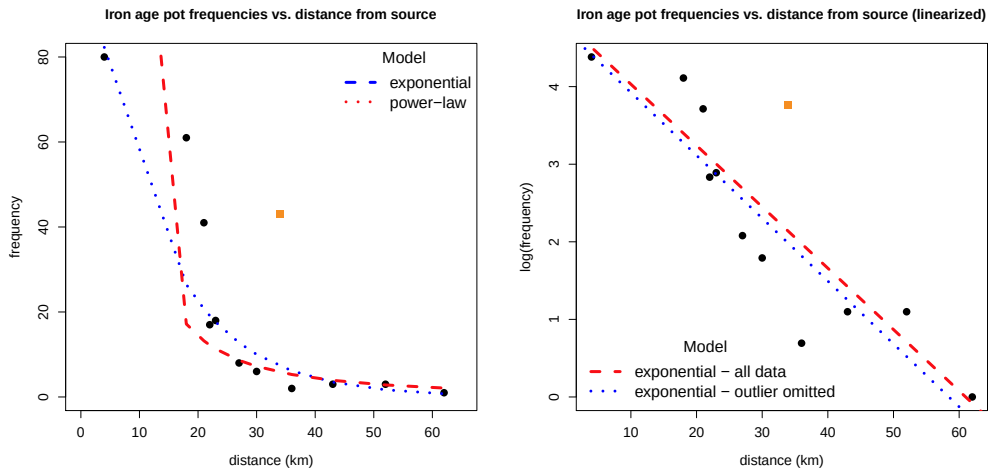

Figure 5.8: *The plot to the right shows linearized model fits for the Morris (1994) data, both omitting and including the outlier. The plot to the left transforms the model fits for the exponential decay and power-law models, omitting the outlier, back to the original scale. See the text for a full discussion.*

There are, perhaps, two lessons here. One is that a large residual with low leverage need not affect the fitted model much. The other is that the two models can give rise to noticeably different results, theory not always providing a guide to choice .

Example 3 – Stone 'circle' dimensions

Barnatt and Moir (1984) present and analyze data from Thom (1967) on the dimensions of stone 'circles'. These are not usually exact circles, some deviations from true circularity being greater than others, The difference between maximum and minimum diameters is used as a measure of deviation. Figure 5.9 reproduces Figure 3 of Barnatt and Moir using 69 circles with diameters less than 160 ft and deviations less than 20 ft (Table B.6). A distinction is made between northern circles (open triangles) and southern circles (closed triangles). Separate regressions are fitted to each subset, that for the northern data having the steeper slope.

Barnett and Moir (1984: 210) draw several conclusions from this. They claim, presumably on their interpretation of the visual evidence, that the regression lines are distinct, and that in southern England circles tend to be constructed more accurately. These conclusions need to be qualified. It is questionable whether linear regression should be used at all and this is pursued in Section 5.3.

To begin, however, the analysis shown in Figure 5.9 is examined more closely. Model (5.5) is used, defining a dummy variable $z = 1$ for southern circles and 0

otherwise. Bearing in mind that z can only take these values it can be seen that for northern circles the intercept and slope are (α, β_1) and for southern circles $(\alpha + \beta_2, \beta_1 + \beta_3)$. A test of the hypothesis that $\beta_3 = 0$ tests whether regressions have the same slope. If this is not rejected, drop (xz) from the model and refit it; a test of the hypothesis that $\beta_2 = 0$ then indicates whether the separate intercept is needed or not. As opposed to this sequential testing, a simultaneous test of the hypothesis $\beta_2 = \beta_3 = 0$ using ANOVA (analysis of variance) methods is also possible (Section 12.3.4).

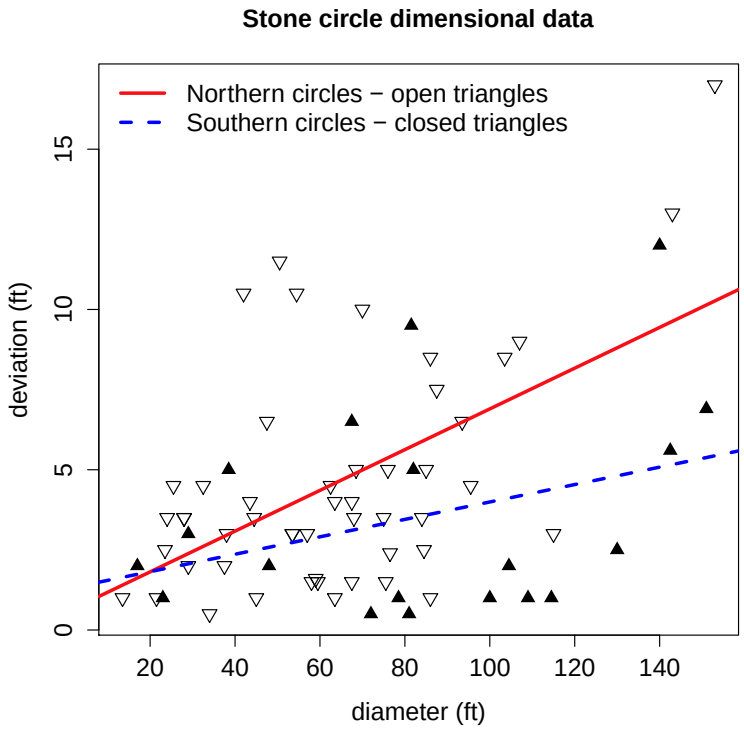

Figure 5.9: *The plots are based on data for 69 stones circles. Circles are divided into northern (open triangles) and southern circles (closed triangles) with simple linear regressions fitted separately to the two regions. See the text for a full discussion.*

Whichever approach is adopted there is no real evidence at the levels usually used to suggest, on the basis of the methodology in the paper, that regressions for the two regions differ significantly. This is shown by the tests just described (details not presented); it can be seen more informally as follows. Barnatt and Moir (1984) fit their regressions separately (i.e. they use two simple linear regressions, rather than the interaction model used here). This gives a difference in slope estimates of

0.036. The standard error of $\hat{\beta}$ for the southern data is 0.018 so that the northern estimate lies within two standard errors of the southern estimate ignoring error in the latter. It would need to be larger to show a significant difference (neither intercept estimate differs significantly from 0, as is to be hoped for). If allowance is also made for the standard error of the slope estimate for the northern data of 0.014 it becomes even clearer that the regressions are not significantly different.

That this conclusion differs markedly from that in the paper is because no allowance was made there for the uncertainty of estimation. This is relatively large because of the noticeable variability in the data. The example is pursued using non-parametric regression methods, in the next section.

5.3 Non-parametric regression

The models used so far have the form $y = f(x) + \varepsilon$ where $f(x)$ is a function linear in the parameters. Linearizable models that reduce to this form have been illustrated. It is possible to define models with rather simple functions $f(x)$ that cannot be linearized so that non-linear least squares estimation, for example, is required. Such models use explicitly defined forms of $f(x)$ usually dependent on just a few parameters.

An alternative approach, rather than assuming a parametric model, is to allow the data itself to determine a form for $f(x)$. This gives rise to the idea of *non-parametric regression* or *scatterplot smoothing*. These methods have been used much less than linear methods in archaeology; Baxter (2003: 63–65) lists some examples available at that date; this section provides a discussion of some possibilities, with application.

The methodology can be viewed as conceptually simple and mathematically complicated but, with the aid of `R`, relatively straightforward to implement. This qualifies the methodology as 'simple' in the sense defined in Section 1.1 but caveats need to be entered. There are a lot of approaches that have been developed – Venables and Ripley (2002: 229) illustrate six. Choices within each approach need to be made that determine the degree of smoothing so that many different estimates are possible and, with much variation in the data, selection and interpretation of an estimate is not straightforward. Accounts of non-parametric smoothing geared to `R` include Venables and Ripley (2002: 228–232) and Faraway (2006: 211–230); Simonoff (1996: 134–214) provides a general presentation. Ambitions here do not extend beyond providing an intuitive account of the method of *loess* smoothing, with examples. Faraway (2006: 228) suggests that the loess smoother is a good all-purpose smoother.

Example 4 – Non-parametric regression of stone 'circle' dimensions

The idea is illustrated in Figure 5.10. For the circle data and the two regions separately a smooth, using the defaults in the `loess.smooth` function, has been fitted. This is discussed in more detail in Section 5.4. The departure from linearity is sufficient to suggest that the original fitting of linear models is not appropriate (a conclusion that might be reached by simply looking at the pattern of scatter involved in the two plots).

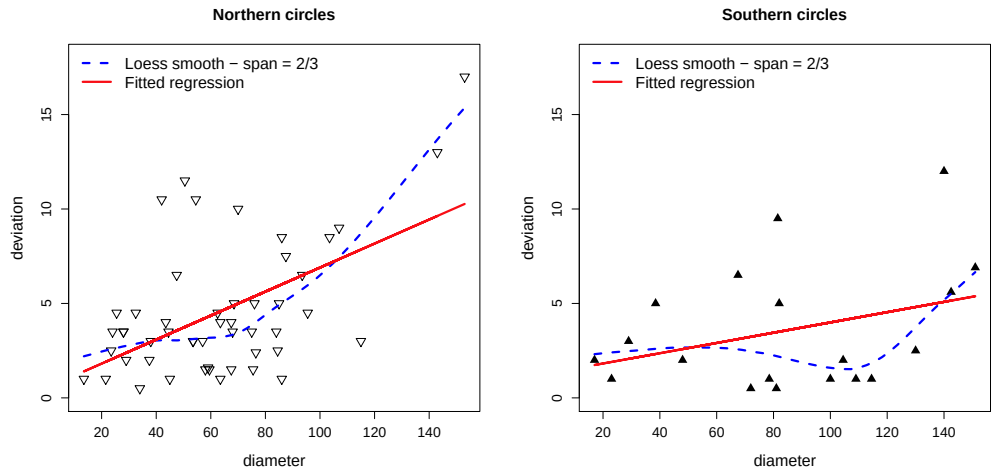

Figure 5.10: *The plots are for the northern and southern circles fitted separately and showing loess smooths as well as the fitted regression.*

The loess method works by defining a *neighborhood* of points (or *window*) about each value of x and fitting a linear or quadratic regression model within each neighborhood to predict the smoothed values at x. Cleveland (1979) provides a detailed technical account. A complex iterative weighted regression procedure is used for fitting (Baxter 2003: 65). Loosely speaking, a complicated kind of average is computed for each neighborhood, and the resultant points are 'joined up' to get the smooth. The appearance and smoothness of an estimate is dictated by the size of the neighborhood, determined by the `span` in the `loess.smooth` function, and precise fitting procedure. Section 5.4 discusses other arguments available.

Example 5 – Varying neigborhood size in non-parametric regression

Figure 5.11 uses the stone circle data for the southern region. The main purpose is to illustrate the variation that can arise with different levels of smoothing; defaults in the `loess.smooth` function have been used, other than that the span is varied.

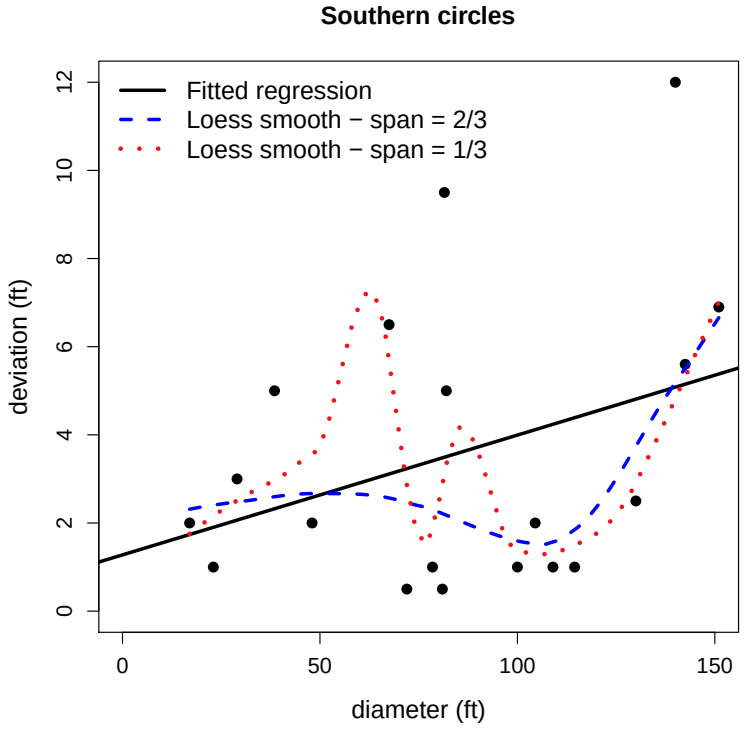

Figure 5.11: *Non-parametric regressions for the southern circle data showing the effect of varying the span.*

The solid line shows the linear regression fit. The dashed line shows the fit based on the loess smoother with a span of 2/3; the dotted line uses a span of 1.3. The larger values of a span produce larger neighbourhoods and hence greater smoothing. The choice of the degree of smoothing has a limited effect for larger diameters of more than 100 ft; for smaller diameters the results are highly sensitive to the choice. Both spans suggest that simplifying to a liner model is inappropriate.

Example 6 – Non-parametric regressions of the post-medieval wine bottle data

For a final example, and to make some slightly different points, we return to the post-medieval wine bottle data, the pairs plot for which was shown in Figure 5.6. With *body height* as the independent variable, smoothed fits are presented using four different dependent variables.[5]

In the analyes the `loess.smooth` function with defaults is used. This explains

[5]Other than for date, treating body height as the 'independent' variable is a convenience. The interest lies more in description than prediction.

why some obvious outliers have little effect on the smooths since their effect is downweighted (see the notes for Figure 5.10 in Section 5.4). Varying the level of smoothing between 1/3 and 2/3 has little effect on the fitted models.

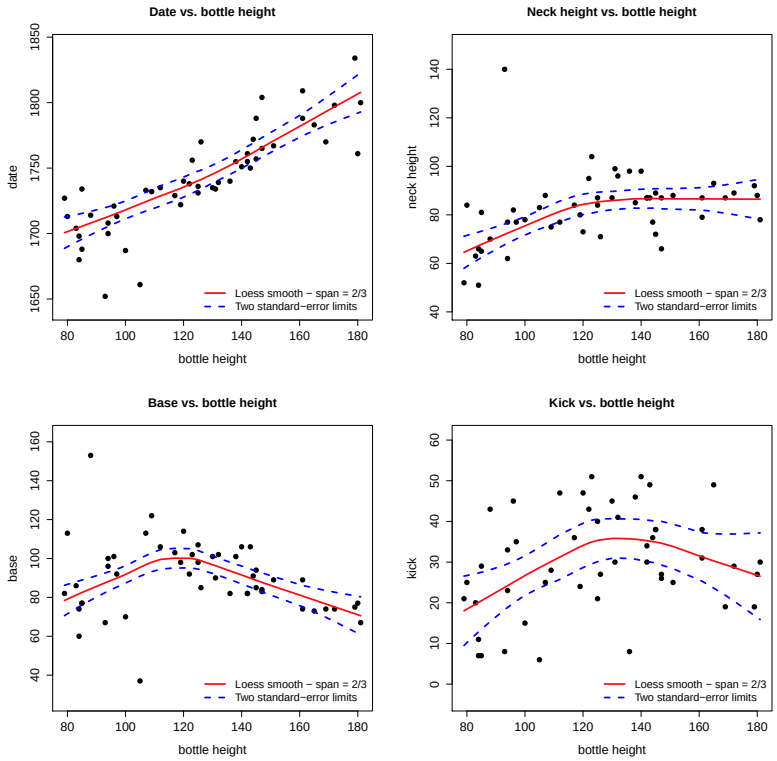

Figure 5.12: *Non-parametric regressions for the post-medieval wine bottle data with approximate two standard-error limits. See the text for an explanation.*

For *date* as the dependent variable there is little evidence of any serious departure from linearity, justifying the use of a linear fit; for other variables there is clear evidence of non-linearity with two 'regimes' either side of a body height of about 120 mm. For *date* as the dependent variable the two outliers previously noted depart from the general linear pattern; if the `family` and/or `degree` arguments are varied in the `loess.smooth` function these result in a departure from linearity at the lower end of the smooth. In summary, and notwithstanding the fact that variability is appreciable in most examples as the standard-error limits show, there is generally a fairly stable and simple underlying non-linear pattern, differentiated by bottles with small (less than 120 mm) and large body heights.

5.4 R notes

Figures 5.1, 5.4, 5.5, and 5.6

Code for Figure 5.1 is in the text where an object `fit` was created that is needed for other plots. The preliminaries are

```
date <- pmedwine$Date
BH <- pmedwine$BH
index <- 1:dim(pmedwine)[1] # create list of row numbers
fit <- lm(date ~ BH)
```

The residual plot, 5.4, is obtained with

```
library(MASS)
win.graph()
plot(fitted(fit), studres(fit))
points(jitter(fitted(fit),amount = 0), jitter(stdres(fit)))
abline(h = 0); abline(h = 2.5); abline(h = -2.5)
```

where labeling and legend commands etc. are omitted. The `MASS` package is needed to obtain the standardized (`sddres`) and studentized (`studres`) residuals. The `points` function operates in the same way as the `lines` function and adds points, with the specified coordinates, to the graph previously created. The usual color, character expansion etc. arguments are available. The `jitter` function is used to displace points slightly to avoid overwriting; see `?jitter` for details of control.

The plots put together for Figure 5.5 are obtained with

```
plot(index, lm.influence(fit)$hat)     # leverage statistic
plot(index, cooks.distance(fit))
```

The `lm.influence` function using `lm.influence(fit)$hat` extracts the *leverage* statistic. The `?lm.influence` query helps direct you to a lot of other diagnostic statistics, not illustrated here, obtained via the `influence.measures` function. The `cooks.distance` function extracts, as the name suggests, Cook's statistic.

For Figure 5.6

```
pairs(pmedwine[ , -c(1,3)])
```

removes the first and third variables not used in the plot.

Figures 5.2, 5.7, 5.3 and 5.8

Data from Tables 5.1 and 5.2 are in the files `cummins` and `morris`. Presentational arguments and the legends are omitted. As written, the appropriate models, discussed in the text, are added to the initial plot; if only the former is needed use the first lines after the `win.graph()` directives.

```
cummins.plots <- function(){
win.graph()
plot(Frequency ~ Distance, data = cummins) # Figure 5.2

pred <- lm(log(Frequency) ~ log(Distance),
data = cummins)$fitted.values

lines(cummins$Distance, exp(pred))         # Add for Figure 5.7

win.graph()
# Figure 5.2
plot(log10(cummins$Distance), log10(cummins$Frequency))

# Add for Figure 5.7
abline(lm(log10(cummins$Frequency) ~ log10(cummins$Distance)))
abline(lm(log10(cummins$Frequency[1:5]) ~ log10(cummins$Distance[1:5])))
}
cummins.plots()
```

The above is for the data from Cummins (1980). Logarithms to base 10 in the second plot emulate Cummins (1980). The code for the data from Morris (1994) is similar, allowing for the difference in the treatment of outliers and the models fitted. The `exponential` function, `exp`, transforms the fit for the linearized models back to the original scale.

```
morris.plots <- function() {
Col <- rep("black", 12); Col[8] <- "darkorange"
Sym <- rep(16,12); Sym[8] <- 15  # Case 8 is the outlier

win.graph()
# Use first line for Figure 5.3
plot(Frequency ~ Distance, data = morris, pch = Sym, col = Col)
# Add for Figure 5.8
# Exponential decay omit outlier
ze <- lm(log(Frequency) ~ Distance, data = morris[-8,])
lines(morris$Distance[-8], exp(ze$fitted.values))

# power-law model omitting outlier
```

```
zp <- lm(log(Frequency) ~ log(Distance), data = morris[-8,])
lines(morris$Distance[-8], exp(zp$fitted.values))

win.graph()
# Use first line for Figure 5.3
plot(log(Frequency) ~ Distance, data = morris, pch = Sym, col = Col)
abline(lm(log(Frequency) ~ Distance, data = morris))      # All data
# Outlier omitted
abline(lm(log(Frequency) ~ Distance, data = morris[-8,]))
}
morris.plots()
```

Figure 5.9

The original data are split into northern and southern data for the stone circles, called `circlesN` and `circlesS`. Most of the presentational arguments are omitted.

```
circles.plots <- function() {
N <- circlesN; S <- circlesS
regN <- lm(N$Deviation ~ N$Diameter)
regS <- lm(S$Deviation ~ S$Diameter)
plot(N$Diameter, N$Deviation, pch = 6)
points(S$Diameter, S$Deviation, pch = 17)
abline(regN)
abline(regS)
}
circles.plots()
```

In general the `xlim` and `ylim` arguments, not shown here, may need setting so that when the `points` function is invoked all the points are included on the plot previously created.

Figure 5.10

In the following function presentational arguments, other than `pch` and `main` which are specified in the call to the function, have been omitted. Its use is illustrated for the Northern circles `circlesN`; use `circlesS` for a plot of the Southern circles with the appropriate modification of `Pch` and `Main`. Legend specifications have been omitted from the function but can be supplied either within the function or after it is invoked. For both plots `ylim = c(0,18)` was used to ensure comparability.

```
region.plots <- function(Data, Pch = 16, Main = ""){
reg <- lm(Data$Deviation ~ Data$Diameter)$fitted.values
```

```
plot(Data$Diameter, Data$Deviation, pch = Pch, main = Main)
lines(loess.smooth(Data$Diameter, Data$Deviation))
lines(Data$Diameter, reg)
}
region.plots(circlesN, 6, Main = "Northern circles")
```

Loess smoothing is described briefly in the text; the `loess.smooth` function is a convenience for adding the fit produced by the `loess` function to a plot; the latter function is more flexible for some purposes. Users need to be aware that the defaults in the two implementations vary. The default `span`, for example, is 2/3 for `loess.smooth` and 3/4 for `loess`. The `degree` and `family` arguments control other aspects of a smooth with, respectively, options `1, 2` and `"s", "g"`. The first option in each case is the default for `loess.smooth`; the second options are the defaults for `loess`.

The method works by fitting a (weighted) regression to points within a neighborhood, using the fit to predict a representative point for the neighborhood. A local linear fit is applied if `degree = 1`, otherwise `degree = 2` produces a quadratic fit. If `family = "s"` is used a robust regression fit that downweights the influence of outliers is used; otherwise `family = "g"` assuming Gaussian errors (i.e. normally distributed) is applied. If spans are specified to be the same then the defaults in `loess.smooth` will produce a smoother fit than the `loess` defaults – which may not be what is sought.

Although the defaults for `loess.smooth` have been used in the example, for illustration and the message is clear enough, the choices can have a noticeable effect on the smooth and are worth experimenting with. This is pursued, with regard to the span, in the next example.

Figure 5.11

Presentational arguments other than `span`, and the legend, are omitted.

```
varyspan <- function() {
S <- circlesS
regS <- lm(S$Deviation ~ S$Diameter)

win.graph()
plot(S$Diameter, S$Deviation)
abline(regS)
lines(loess.smooth(S$Diameter, S$Deviation, span = 2/3))
lines(loess.smooth(S$Diameter, S$Deviation, span = 1/3))
}
varyspan()
```

Figure 5.12

Define `BH, Date, NH, Base, Kick` using `BH <- pmedwine$BH` etc. Some presentational arguments and the legend are omitted from the code.

```
bottlesloess <- function(x, y, Xlab = "", Ylab = "", Main = "")
{
pred <- predict(loess(y ~ x, span = 2/3, family = "s", degree = 1),
se = TRUE)
upper <- pred$fit + 2*pred$se
lower <- pred$fit - 2*pred$se

plot(x, y, xlab = Xlab, ylab = Ylab, main = Main)
lines(loess.smooth(x, y, span = 2/3))
lines(loess.smooth(x, upper, span = 2/3))
lines(loess.smooth(x, lower, span = 2/3))
}

win.graph()
bottlesloess(BH, date, "bottle height", "date", Main = "Date vs.
bottle height")
```

The `predict` function needs to be applied to a fit from the `loess` function to obtain (approximate) standard errors using the `se = TRUE` argument. Arguments to `loess` produce the same fit as the `loess.smooth` default; `predict` does not work with `loess.smooth`, hence the need for `loess` if standard errors are needed. In the call to the function replace `Date` and labeling arguments with those for `NH` etc. to get the other plots.