14

LINEAR REGRESSION

OBJECTIVES

After reading this chapter, you should be able to:

1. Identify if least squares regression is an appropriate analytical tool for meeting your objectives given the characteristics of your data.

2. Construct a linear model to meet your objectives, including control of confounding and identification of interaction.

3. Interpret the regression coefficients from both a technical and causal perspective.

4. Convert nominal, ordinal or continuous predictor variables into regular or hierarchical variables and interpret the resulting coefficients correctly.

5. Assess the model for linearity between continuous predictors and the outcome, and for homoscedasticity, and normality of residuals. You should also be able to identify appropriate transformations of the outcome or predictor variables to help ensure that the model meets these assumptions.

6. Detect and assess individual observations as potential outliers, leverage observations and/or influential observations.

7. Identify study designs whose data require a time-series approach to analysis.

14.1 INTRODUCTION

Up to this point, most of our examples in which we relate an outcome to an exposure, have been based on qualitative outcome variables—that is, variables that are categorical or dichotomous. Here we describe linear regression which is suitable for modelling the outcome when it is measured on a continuous, or near-continuous scale. Examples of these would include weight gain, milk production, somatic cell counts, and in some circumstances, disease frequency at the herd level. Recent work has also shown that linear regression can be used to model incidence risk differences (Cheung, 2007). One example to demonstrate the use of linear models is Abu-Zidan & Rao (2003) in which multiple regression was used to identify factors related to the severity of injury in falls from horses. For a readable introduction to linear regression, see Marill (2004a) and Marill (2004b).

In regression analysis, the relationship between the outcome and the predictors is asymmetric in that we think the value of one variable (the **outcome**) is caused by (or we wish to predict it by) the value or state of another variable (the **predictor**(s)). (**Note** The outcome and predictor variables are sometimes referred to as dependent and independent variables respectively.). We will refer to the predictor variable(s) of primary interest as the **exposure variable**(s) and other predictors as **extraneous variables**. The predictor variables can be measured on a continuous, categorical or dichotomous scale.

14.2 REGRESSION ANALYSIS

When only one predictor variable is used, the model is called a **simple regression model**. The term 'model' is used to denote the formal statistical formula, or equation, that describes the relationship we believe exists between the predictor and the outcome. For example, the model

$$Y = \beta_0 + \beta_1 X_1 + \varepsilon \qquad \textit{Eq 14.1}$$

is a statistical way of describing how the value of the outcome variable Y changes across population groups formed by the values of the predictor variable X_1. More formally it says that the mean value of the outcome Y for any value of the predictor variable is determined using a starting point, β_0, when X_1 has the value 0, and for each unit increase in X_1 the outcome Y changes by β_1 units. β_0 is usually referred to as the **constant** or the **intercept** term whereas β_1 is usually referred to as the **regression coefficient**. The ε component is called the **error** and reflects the fact that the relationship between X_1 and Y is not exact. The errors are assumed to be normally and independently distributed ($\varepsilon \sim N(0,\sigma^2)$). We estimate these errors by **residuals**; these are the difference between the observed (actual) value of the observation and the value predicted by the model for a given value of X_1.

The βs represent population parameters which we estimate based on the observed data and our model. We will refer to predictor variables as Xs. In general, we will denote the number of observations as n. Thus, our predicted values are:

$$\hat{Y}_i = \beta_0 + \beta_1 X_{1i}, \quad i = 1, \ldots, n \qquad \textit{Eq 14.2}$$

where $\hat{Y}_i$ is the predicted value of the outcome for the i^{th} observation at the observed value of the predictor X_{1i}. (**Note** While it is common practise to use a '^' to designate predicted values $\hat{Y}_i$ or estimated coefficients $\hat{\beta}$, we will generally omit the '^' because whether we are referring to observed data and true population parameters or predicted values and estimates of

parameters is generally obvious from the context. Similarly, in Eq 14.2, specific observations are denoted by the subscript *i*; but in most instances, for simplicity we will omit reference to specific observations.)

Bear in mind that in using *X*-variables to predict *Y* in a regression model there is no necessary underlying assumption of causation; we might just be estimating predictive associations. Nonetheless, we often use terms such as '*X* affects *Y*', or the 'effect of *X* on *Y* is...' when interpreting the results of our models. For clarity we will always try and indicate if we are making 'causal' assumptions.

Almost without exception, the regression models used by epidemiologists will contain more than one predictor variable. These belong to the family known as multiple regression models, or **multivariable** models (**Note** that **multivariate** indicates 2 or more outcome variables; multivariable denotes more than one predictor). With 2 predictor variables, the regression model could be written as:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

which suggests that we can predict the value of the outcome *Y* knowing the baseline (intercept or constant) β_0 and the values of the 2 predictor variables (*ie* the *X*s). The parameters β_1 and β_2 describe the direction and magnitude of the association of X_1 and X_2 with *Y*. More generally, there can be as many *X*-variables as needed (the number of predictors is often denoted with *k*). A major difference from the simple regression model is that in the above multivariable model, β_1 is an estimate of the effect of X_1 on *Y* after controlling for the effects of X_2, and β_2 is the estimated effect of X_2 on *Y* after controlling for the effects of X_1. Expressed another way, β_1 is an estimate of the effect of X_1 on *Y* among individuals that have the same value of X_2. As in simple regression, the model suggests that we cannot predict *Y* exactly, so the random error term (ε) takes this into account. Thus, our prediction equation is:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

where *Y* is the predicted value of the outcome for specific values of the 2 predictors X_1 and X_2. In this equation, β_1 describes the number of units change in *Y* as X_1 changes by one unit, X_2 being held constant, whereas β_2 describes the number of units change in *Y* as X_2 changes by one unit, X_1 being held constant.

In observational studies, incorporating more than one predictor almost always leads to a more complete understanding of how the outcome varies and it also decreases the chance that the regression coefficients for exposures of interest are biased by confounding (extraneous) variables. Assuming that we have not included intervening variables (see Chapter 13), or effects of the outcome in our model, the βs are not biased (confounded) by any variable included in the regression equation, but can be biased if confounding variables are omitted from the equation. From a causal perspective, if intervening variables are included, the coefficients do not estimate the causal effect (see Section 14.7). Unfortunately, one can never be sure that there are not other variables that were omitted from the model that also affect *Y* and are related to one or more of the *X*s. These *X*-variables could be unknown, not thought (at least initially) to be important, or (as it often happens) not practical/possible to measure. In other circumstances, we might have numerous potential confounders and need to decide on the important ones to include. As noted in Chapter 15, a major trade-off in model-building is to avoid omitting necessary variables which could confound the relationship described by the βs, while not including variables of little importance in the equation, as this will increase the number of βs

estimated and may lead to poor performance of the equation on future datasets. Also, having to measure unnecessary variables increases the cost of future work.

In order to assist with the principles of describing multiple regression, we will develop examples from a dataset concerning the impact of diseases on reproductive performance in dairy cows in Australia. However, the initial outcome variable will be milk production in the first 120 days of lactation (-milk120-) and we will regress it on -parity- to ascertain if parity changes are associated with changes in milk production. Subsequently, we will focus on the effects of reproductive events and diseases on the length of the interval from the time at which the owner starts breeding cows (end of the 'wait period') to conception (-wpc-). These data were obtained from a longitudinal study in 4 states in Australia, but they are a subset from the original database and the analyses in this chapter are limited to data from 7 herds with particularly high rates of reproductive diseases. The names of the variables used in this Chapter, and their descriptions are shown in Table 14.1; further details are in Chapter 31. The diseases are listed in order of their average time to occurrence (*eg* dystocia occurs before retained fetal membranes which occurs before vaginal discharge).

Table 14.1 Selected variables from the dataset -daisy2-

Variable	Scale of measurement	Description
herd	nominal	herd number
cow	nominal	cow number (unique)
herd_size	continuous	herd size
calv_dt	date	date of calving
mwp	continuous	minimum wait period for herd
parity	continuous	lactation number
milk120	continuous	litres of milk in first 120 days of lactation
wpc	continuous	interval from wait period to conception
twin	dichotomous	twins born
dyst	dichotomous	dystocia at calving
rp	dichotomous	retained placenta at calving
vag_disch	dichotomous	vaginal discharge observed

14.3 HYPOTHESIS TESTING AND EFFECT ESTIMATION

14.3.1 The ANOVA table

The idea behind using regression is that we believe that information in the *X*-variables can be used to predict the value of *Y*. Now, if we have collected the data, we know the observed *Y*-values and we can describe the distribution of *Y* using the mean, variance and other statistics. Relevant statistics for -milk120- were: median and mean (average) were both 3,215 litres, the standard deviation was 698 l and the range was 1,110 to 5,630 l.

With no further information, the best estimate of the value of *Y* for a particular subject would be an estimate of central tendency such as the median or mean value (here they are equal so we

know the distribution of -milk120- is symmetric). However, if the *X*-variable contains information about the *Y*-variable, we should be able to do a better job of predicting the value of *Y* for a given individual (cow) than if we did not have that information. The formal way this is approached in regression is to ascertain how much of the sums of squares (SS) of *Y* (the numerator of the variance of *Y*) we can explain/predict with knowledge of the *X*-variable(s).

Table 14.2 Analysis of variance showing decomposition of sums of squares in regression model with k predictor variables

Source of variation	Sum of squares	Degrees of freedom	Mean square	F-test
Model (or regression)	$\text{SSM}=\sum_{i=1}^{n}(\hat{Y}-\overline{Y})^2$	dfM = k	MSM = SSM/dfM	MSM/MSE
Error (or residual)	$\text{SSE}=\sum_{i=1}^{n}(Y_i-\hat{Y}_i)^2$	dfE = n-(k+1)	MSE = SSE/dfE	
Total	$\text{SST}=\sum_{i=1}^{n}(Y_i-\overline{Y})^2$	dfT = n-1	MST = SST/dfT	

In the formulae in the table, $\overline{Y}$ is the mean of the *Y*s, and *k* is the number of predictor variables in the model (not counting the intercept). When the SS are divided by their degrees of freedom (df), the result is a **mean square**, here denoted as MSM (model), MSE (error) and MST (total)—in other settings we might call these variances, but the jargon in regression is to call them mean squares. Formally, this decomposition of the total sum of squares (SST) is shown in the second column of Table 14.2 (*ie* SST=SSM + SSE; also, dfT=dfM + dfE). For our example, parity will be the *X*-variable of interest. The MSE is our estimate of the error variance and therefore also denoted as σ^2. Furthermore, σ, the square root of σ^2, is called the **root MSE**, or the **standard error of prediction** (see Example 14.1).

The sums of squares are partitioned by choosing values of the βs that minimise the SSE (or MSE); hence, the name 'least squares regression'. There is an explicit formula for doing this, which, in general, involves matrix algebra, but for the simple linear regression model, the βs can be determined using:

$$\beta_0=\overline{Y}-\beta_1\overline{X}_1 \quad \text{and} \quad \beta_1=\sum(X_{1i}-\overline{X}_1)(Y_i-\overline{Y})/\text{SSX}_1 \quad (\text{with } \text{SSX}_1=\sum(X_{1i}-\overline{X}_1)^2) \qquad \textit{Eq 14.3}$$

For a small dataset, these computations could be done by calculator, but in practise we always use computer software.

14.3.2 Assessing the significance of a linear regression model

To assess whether or not the predictors in the model (collectively) have a statistically significant relationship with the outcome, we use the *F*-test from the analysis of variance (ANOVA) table. The null hypothesis is H_0: $\beta_1=\beta_2=...\beta_k=0$ (*ie* all regression coefficients except the intercept are zero). The alternative hypothesis is that this is not true—that is, at least one (but not necessarily all) of the βs is non-zero. The distribution of the *F*-statistic is an *F*-distribution with the numerator degrees of freedom equal to dfM and the denominator degrees of freedom equal to dfE (as given in Table 14.2). In Example 14.1, the *F*-value 262.3 is highly significant ($p<0.001$) indicating that the *X*-variable(s) in the model (-parity- in this instance) explains some of the variation in -milk120-. One feature of the ANOVA table we should always pay attention to is the number of observations included in the model. In multivariable models with missing data, this number can decrease considerably.

Example 14.1 A simple linear regression model of -milk120- on -parity-
data = daisy2

A linear regression model with -milk120- as the outcome and -parity- as the sole predictor was fit to the 7 herd subset of -daisy2- data. The top left of the table below shows the decomposition of the sums of squares, the top right gives details about the regression model.

Source	SS	df	MS
Model	109234227	1	109234227
Residual	638905966	1534	416496.7
Total	748140192	1535	487387.7

Number of obs = 1536
F(1, 1534) = 262.27
Prob > F = 0.0000
R-squared = 0.1460
Adj R-squared = 0.1455
Root MSE = 645.37

Note that the variance (MS) of -milk120- is 487387.7 and this is somewhat larger than the MS residual suggesting that parity does explain some of the variation in -milk120-. The root MSE has the same scale as -milk120- (*ie* litres) and because -parity- is associated with -milk120- it is smaller than the standard deviation (698 l) reported above.

The regression coefficients from the model are shown below.

milk120	Coef	SE	t	P>t	95% CI	
parity	178.347	11.013	16.190	0.000	156.746	199.948
constant	2727.080	34.340	79.410	0.000	2659.722	2794.438

The coefficient for parity suggests that, for each unit increase in parity, -milk120- increases by 178.3 l. Given the SE (11.0) of this statistic, the *t*-statistic (16.2) is significant at the 5% level so we can assume, at this point, that -parity- has an association with (or effect on) -milk120-. This is consistent with the 95% confidence interval values which does not include 0 (the no-effect level). The CI suggests that a reasonable range for the effect of a unit change in parity is between 157 l and 200 l.

We usually do not test the intercept, but it is essential for interpretation of this model as it represents the value of the outcome (-milk120- in this instance) when the values of all *X*-variables in the model have the value 0. Of course cattle of -parity-=0 do not have a real -milk120- value, so subsequently, we will describe how to scale the predictor variable(s) so that the intercept has a sensible interpretation (Section 14.4.1).

Some care is necessary when interpreting the model *F*-statistic as its meaning changes with the method of model-building (Livingstone and Salt, 2005). The *F*-test probably has only a straightforward meaning when the *X*s are manipulated treatments in a controlled experiment, and all comparisons are appropriately planned *a priori*. In observational studies, the *F*-statistic is influenced by the number of variables available for entry, their correlations with each other, the number actually selected for inclusion in the model, and the total number of subjects (sampling units). Most variable selection methods (Chapter 15) choose variables in a manner that tends to maximise *F*; hence the observed *F* overestimates the actual significance of the model. On the other hand, if useless variables are forced into the model with the hope of controlling all confounding, the *F*-statistic might be biased downwards. Sometimes with highly correlated variables in the model, the *F*-test might be significant yet the test of the individual coefficients might suggest that none of them differ significantly from zero (see Section 14.5).

14.3.3 Testing the significance of a regression coefficient

A *t*-test with n-(k+1) degrees of freedom (dfE) is used to evaluate the significance of any of the regression coefficients (*eg* the j^{th} coefficient). The usual null hypothesis is H_0: β_j=0 but any value of β* other than 0 can be used in H_0: β_j=β* depending on the context. The *t*-test formula is:

$$t = \frac{\beta_j - \beta^*}{SE(\beta_j)} \qquad \textit{Eq 14.4}$$

where SE(β_j) is the standard error (SE) of the estimated coefficient. This SE is always computed as the root MSE times a constant that depends on the formula for the estimated coefficient and the values of the X-variables in the model. Except for the simplest situations, it is not easily computable; however, it is always given in the computer output from the estimation of the model. For a model with only one predictor (X_1), the SE of the regression coefficient is:

$$SE(\beta_1) = \sqrt{MSE/SSX_1} \qquad \textit{Eq 14.5}$$

As the formula indicates, both the variance of X_1 and the MSE affect the standard error. In Example 14.1, the *t* value of 16.2 has a P-value of <0.001 so we would reject the null hypothesis that the true regression coefficient has the value β_1=0 which would indicate no association of -parity- with -milk120-. Fig. 14.1 shows the trend of increasing milk production with increasing parity.

Similar to the *F*-statistic, the inference to be made based on the P-value associated with the calculated *t*-statistic is often difficult to assess in non-experimental studies. In experiments, the *X*s are manipulated treatments, or blocking factors, and the observed *t*-value can be referred to tables (of the *t*-distribution) for a P-value (observed level of significance). The same is probably true if the variable being tested in an observational study was of *a priori* interest (*eg* if the observational study was conducted to determine the effect of a specific X on Y, given control of a set of other variables). However, if a variable selection programme was used to sort through a list of variables, selecting those with large *t*-values in the absence of a specific *a priori* hypothesis, then the actual level of significance is higher than the nominal level of significance (usually termed α) that you specify for a variable to enter/stay in the equation. Nonetheless, using the P-value as a guideline is a convenient and accepted way of identifying potentially useful predictors of the outcome.

14.3.4 Estimates and intervals for prediction

Calculating the point estimate for predictions in regression is straightforward. The complex component is determining the appropriate variance associated with the estimate, because there are 2 types of variation in play. One source of variation results from the estimation of the parameters of the regression equation (*ie* this is the usual SE). The other is the variation associated with a new observation (*ie* the variation about the regression equation for the mean). The prediction (confidence) interval for a new observation involves both of these sources of variation.

For example, in a simple linear regression model, the predicted value for a population of individuals with X_1=x* has a SE (designated SE_{mean}; sometimes called the prediction error) of:

$$\mathrm{SE}_{\mathrm{mean}}(Y|x^*)=\sigma\sqrt{\frac{1}{n}+\frac{(x^*-\bar{X}_1)^2}{\mathrm{SSX}_1}} \qquad \textit{Eq 14.6}$$

which can be interpreted as the variation associated with the expectation (*ie* mean) of a large number of new observations, with the particular value x^* chosen for prediction. Using the data in Example 14.1, for a parity 5 cow (the mean parity is 2.73), the predicted outcome is 3,618.8 l with a prediction SE of 29.9 l.

The standard error for a new single observation (designated $\mathrm{SE}_{\mathrm{obs}}$; sometimes called the forecast SE) with predictor value x^* is increased because we must account for the additional σ^2 because the individual predicted value is unlikely to equal its expectation (*ie* unlikely to exactly equal the average value for all individuals with $X=x^*$):

$$\mathrm{SE}_{\mathrm{obs}}(Y|x^*)=\sigma\sqrt{1+\frac{1}{n}+\frac{(x^*-\bar{X}_1)^2}{\mathrm{SSX}_1}} \qquad \textit{Eq 14.7}$$

Using the data in Example 14.1, for a specific parity 5 cow, the predicted outcome is 3618.8 l with a forecast SE of 646.1 l. Two points can be made here. First, the variation associated with predicting the mean outcome is much less, and prediction intervals much more narrow than those for a specific subject. Second, the further that x^* is from the mean value of X_1, the greater the variability in the prediction. The 95% confidence limits for the predictions are found using:

$$95\%\ \ \mathrm{CI}=Y\pm t_{.05}(SE) \qquad \textit{Eq 14.8}$$

where the *t*-statistic has the dfE and SE is either $\mathrm{SE}_{\mathrm{mean}}$ or $\mathrm{SE}_{\mathrm{obs}}$ (as noted above).

The association between -parity- and -milk120- in dairy cows as determined by a linear regression of -milk120- on -parity-, with prediction intervals for the mean and for a new observation are shown in Fig. 14.1.

14.3.5 Interpreting R^2 and adjusted R^2

R^2 describes the amount of variance in the outcome variable that is 'explained' or 'accounted for' by the predictor variables and usually is called the **coefficient of determination** (in Example 14.1, this is 14.6%). Given that more than 85% of the variation in -milk120- is unexplained, this suggests that we cannot predict milk production very precisely if we only know the parity of a cow. Perhaps additional variables can add to the explained proportion (a rationale for a multivariable model). One formula for R^2 is R^2=SSM/SST=1-(SSE/SST). It also is the squared correlation coefficient between the predicted and observed *Y*-values. The contribution of a specific variable to R^2 is one way of measuring the relative importance of that variable in the final model. Several indices of importance based on this approach have been evaluated (Chao *et al*, 2008).

Unfortunately, R^2 always increases as variables are added to a regression model which makes R^2 useless for variable selection. However, R^2 can be adjusted for the number of variables in the equation (*k*), and this adjusted value will tend to decline if the variables added contain little additional information about the outcome. The formula for the adjusted R^2 is: adjusted R^2=1-(MSE/MST).

In multivariable models, the adjusted R^2 is slightly lower than the R^2. The adjusted R^2 is useful

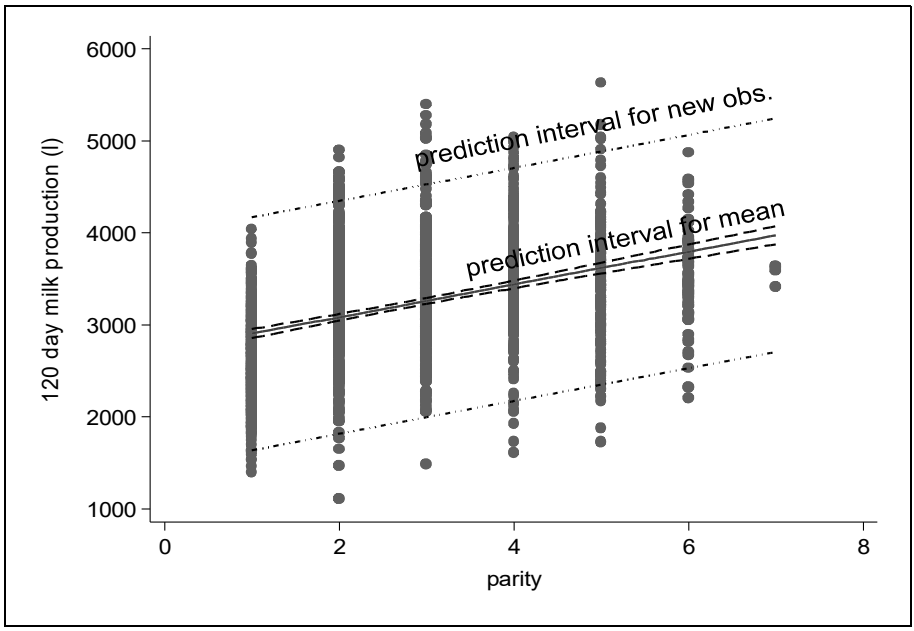

Fig. 14.1 Prediction (confidence) intervals for mean & new observation

for comparing the relative predictive abilities of models, with different numbers of variables in them. For example, if one model has 7 variables and another equation 3, the R^2 for the model with 7 might exceed that for the model with 3 (and it always will if the smaller model is a submodel of the larger one), but its adjusted R^2 might be less. The adjusted R^2 is sometimes used as a basis for selecting potentially good models, but this approach is not without its drawbacks (see Section 15.8.1).

When assessing R^2 we should be aware that non-random sampling can have a pronounced effect on its value. For example, if you select subjects on the basis of extreme X-values, as in a cohort study, you might artificially increase the R^2. It would be okay to use regression to estimate the effect of X on Y, but the R^2 *per se* would be of little value. In a similar manner, if the X-values are limited to a narrow range, the R^2 might be very low. It is perhaps useful to point out that if subjects are sampled based on their Y-values, we cannot use linear regression to assess the effect of selected X-variables on Y.

Before moving on to multivariable models, we include Example 14.2—a regression model with a dichotomous predictor; namely one of our key exposure variables -dyst-.

14.3.6 Assessing the significance of groups of predictor variables

Often it is necessary to simultaneously evaluate the significance of a group of X-variables, as opposed to just one variable. For example, this approach should be used when a set of indicator variables has been created from a nominal variable (Section 14.4.2), or if it is desired to add or remove more than one variable at a time (*eg* a set of variables relating to housing or feeding practices) from the model.

In order to assess the impact of the set of variables, we note the change in the error (residual) sum of squares (SSE) before and after entering (or deleting) the set of variables. (Alternatively,

Example 14.2 A simple regression with a dichotomous predictor
data = daisy2

A simple linear regression model of 120-day milk production (-milk120-) with dystocia (-dyst-) as the only predictor was fit.

Source	SS	df	MS
Model	2227805.7	1	2227805.7
Residual	745912387	1534	486253.2
Total	748140192	1535	487387.7

Number of obs = 1536
F(1, 1534) = 4.58
Prob > F = 0.0325
R-squared = 0.0030
Adj R-squared = 0.0023
Root MSE = 697.32

Note that in this model, the *X*-variable(s) in the model is deemed to be significantly (P=0.032) associated with -milk120-. Nonetheless, the variable -dyst- explains very little (0.3%) of the variation in -milk120-.

milk120	Coef	SE	t	P>t	95% CI	
dyst	-160.493	74.981	-2.140	0.032	-307.569	-13.418
constant	3224.708	18.350	175.730	0.000	3188.714	3260.703

The regression coefficient for -dyst- is -160.5 indicating that as -dyst- increases by 1 unit the -milk120- decreases by 160.5 l (recall that the coding for -dyst- is 0 if dystocia did not occur and 1 if dystocia did occur; so an increase of 1 unit is the difference in outcome between 2 cows, one of which had -dyst- and the other did not). The P-value indicates that the apparent effect of 160.5 l is significantly different from 0 so 'chance' is not a likely explanation for the association. However, given the low R^2 we should temper our interpretation until we have looked at the effect of -dyst- in combination with other variables.

one might use the model sum of squares, as indicated below.) That is, note SSE_{full} with the variable-set of interest in the model (called the 'full model'), then remove the set of variables (*eg* X_j and X_k) and note the SSE_{red} (for the 'reduced model'). If variables X_j and X_k are important, then $SSE_{full} \ll SSE_{red}$ (and $SSM_{full} \gg SSM_{red}$).

The increase in SSE (or reduction in SSM) by deleting the variables from the model is divided by the number of variables in the set (which equals dfE_{red} - dfE_{full}) to give us the MS from these variables. Dividing this MS by the MSE_{full} provides an *F*-test of the significance of X_j and X_k conditional on the other variables in the model. In summary, the formula to assess a set of variables is:

$$F_{group} = \frac{\left(\frac{SSE_{red} - SSE_{full}}{df\,E_{red} - df\,E_{full}}\right)}{MSE_{full}} \sim F\left([df\,E_{red} - df\,E_{full}], df\,E_{full}\right) \text{ under } H_0 \qquad \textit{Eq 14.9}$$

where the null hypothesis (H_0) is that the reduced model gives an adequate description of the data, and large values of the *F*-test are considered as evidence against H_0. As noted, the numerator of the formula might alternatively be calculated from differences of SS- and df-values for the model (instead of error); as SSM_{full}-SSM_{red}=SSE_{red}-SSE_{full}, it gives the same result. Most software contains specific procedures to automate this process. Example 14.3 shows the calculation of an *F*-test for 4 reproductive events that were added to the simple linear model of Example 14.1.

Example 14.3 Testing the significance of multiple variables
data = daisy2

In this example, we added the 4 reproductive events/disease predictor variables (-twin-, -dyst-, -rp-, -vag_disch-) to our model containing -parity- to evaluate the effects of reproductive events on -milk120-. In the full model, only -vag_disch- is significant on its own (data not shown), but we want to check the overall significance of the reproductive events. The ANOVA table from the full model is shown below:

Source	SS	df	MS
Model	112932560	5	22586512
Residual	635207633	1530	415168.4
Total	748140192	1535	487387.7

Number of obs = 1536
F(5, 1530) = 54.40
Prob > F = 0.0000
R-squared = 0.1510
Adj R-squared = 0.1482
Root MSE = 644.34

In the simple model with only -parity- as a predictor, we had SSE_{red}=638905966 with 1,534 df. Hence the F-test is:

$$F = \frac{(638905966 - 635207633)/(1534 - 1530)}{415168.4} = 2.23$$

This F-statistic is borderline significant with 4 and 1,530 df (P=0.064) so, we are left unsure whether, collectively, the reproductive parameters have an effect on -milk120-.

14.4 NATURE OF THE *X*-VARIABLES

The X-variables can be continuous or categorical with the latter being either nominal (meaning that the variable's values constitute 'levels' (or categories) with no meaningful numerical representation) or ordinal (in which case the values represent ordered levels of the variable, *eg* high, medium, low). Examples of nominal variables include: farm identification, categories representing different ways of feeding colostrum, categories representing different breeds of dog *etc*. Nominal and ordinal variables with more than 2 levels should not be used as predictors in their numerical form, they need to be converted to indicator variables (see Section 14.4.2). This is because the corresponding βs would be meaningless (*eg* because herd 4 is not twice something in herd 2, or breed 5 is not 3 units more than breed 2 *etc*), and would not achieve the desired effect (*eg* of removing herd-to-herd variation when examining the relationship between disease and production in cattle, or noting the weight of different breeds of dog).

However, a nominal predictor with only 2 levels (a dichotomous variable) can be used directly, especially when it is coded as 1 or 0 (*eg* the variables -rp- and -dyst-; see Table 14.1). Such variables often serve as answers to questions about present/absent, alive/dead, sick/well *etc*. The regression coefficient represents the difference in the outcome between the 2 levels (*ie* level 1 minus level 0) representing the disease status.

For categorical (nominal or ordinal) variables with multiple levels, we use **indicator variables** (also called **dummy variables**) to code the information into a set of dichotomous variables. See Sections 14.4.2 for a discussion of **regular indicator variables** that can be used for both nominal and ordinal variables, and Section 14.4.3 for **hierarchical indicator variables** applicable only to ordinal or quantitative variables. However, we first examine how to improve the interpretation of regression parameters.

14.4.1 Scaling variables to improve the interpretation of the regression parameter(s)

Often, the predictor variables have a limited range of possible, or sensible values. For example, many cannot be interpreted, sensibly, at the value 0 (*ie* if age, parity, weight, or days to breeding were predictor variables, they have no meaningful interpretation at the value 0). Yet, the intercept reflects the value of the outcome when the predictor has the value 0. Thus, it is often useful to **scale** these variables by subtracting the lowest possible sensible value from each observed value before entering the variable into the model. Then, the intercept coefficient β_0 will be the value of the outcome at this lowest possible value of the original X-variable(s) instead of at zero. As an example, only parity 1 or greater cows can have a -milk120- value, so we could scale parity by subtracting 1; thus a scaled parity (-parity1-) value of 0 is a parity 1 cow. In other situations, for example -age-, the minimum values could be 2 years (age of first calving) so we would subtract this value from -age- to scale it. Scaling has no effect on the regression coefficient or its SE, but it does change the value of the intercept (constant) (see Example 14.4). The scaling can also be done by subtracting values other than the lowest possible value, for example a centre value (mean or median) of the distribution of X.

Another use of scaling is when the X-variable is measured with much greater accuracy than needed (*eg* regressing -milk120- in litres on -herd_size- in our example). In its original form, even if herd_size has a large 'effect' on -milk120-, its coefficient might be very small reflecting the change in -milk120- for each additional cow in the herd. This problem can be circumvented by dividing the value of X by a suitable constant (*eg* 100). Here, a unit change in the scaled herd size (-hs100-) reflects the change in milk production from adding 100 cows to a herd. As another example, if one was predicting badger numbers using the area of pasture, if the latter was measured in square metres, it might be more practical to divide this by 100^2 so that the X-variable is now measuring hectares. Now the coefficient would reflect the change in the number of badgers as pasture is increased by one hectare.

14.4.2 Coding regular indicator variables

Indicator variables (also called **dummy variables**) are created variables whose values have no direct physical relationship to the characteristic being described. For example, suppose there is

Example 14.4 Scaling predictor variables
data = daisy2

Here we scale -parity- by subtracting 1 from the actual parity, so our new variable is: parity1=parity-1.

milk120	Coef	SE	t	P>t	95% CI	
parity1	178.347	11.013	16.190	0.000	156.746	199.948
constant	2905.427	25.235	115.140	0.000	2855.928	2954.925

The effect of an increase of one parity in the scaled variable is the same as in the unscaled variable (Example 14.1; 178 l). In the original scale, -milk120- was predicted to be 2,727.1 l for a 0-parity cow; here it is 2,905.4 l for a parity 1 cow (parity1=0). In general, the coefficient for an appropriately scaled variable will be a sensible number that is easy to interpret and explain.

a variable called -herdnum- that identifies what herd the animals in your study came from. Further, suppose there are 3 herds coded as 1, 2, or 3 (or A, B, C) and we wish to control for 'herd effect' when examining the potential effect of calfhood disease(s) on growth rate in calves. To do this, we create 2 regular indicator (sometimes called **disjoint**) variables (X_1 and X_2) as logical answers to the following questions: Is this calf from herd 1?; if yes, then X_1=1 else X_1=0. For the next indicator variable we ask: is this calf from herd 2?; if yes, then X_2=1 else X_2=0. With respect to these variables, the following values would be present in the dataset:

herdnum	X_1	X_2
1	1	0
2	0	1
3	0	0

Thus, herd 3 is identified as the herd with both indicator variables equal to 0, and will be the **referent** (or comparison level or **reference category**) for assessing the effect of herds 1 and 2 on the outcome. So, in general, to code j levels of a nominal variable, j-1 indicator variables are required, and the j^{th} herd takes the value 0 for all the indicators (see Example 14.5). As the third herd has become the referent level (when all the indicator variables are in the equation), β_1 (the coefficient of X_1) estimates the difference in the outcome between herds 1 and 3, whereas β_2 estimates the difference in the outcome between herds 2 and 3.

One of the levels of the nominal variable will be the referent, so there is merit in considering which level it should be. In terms of the information provided to the model, it does not matter, but careful consideration can enhance the interpretation of the coefficients. In essence, considerations about biological interpretation and the precision of estimates in each level of the nominal variable should be weighed in choosing a referent (*eg* if body temperature is recorded as below normal, normal or above normal, it might make sense to use 'normal' as the referent value). In addition, the referent should have a sufficiently large sample size so that the contrasts (comparisons with the referent) have reasonable precision. Sometimes the level of the nominal variable that has an 'average' response (*eg* close to the mean of the dependent variable) is the desired referent; however, this can lead to a situation where no design variables are significant, as the extreme categories might differ from each other but not from the outcome in the middle (mean) indicator. (**Note** The significance of the indicator variables as a set (Section 14.3.6) is unaffected by the choice of reference category.) In other instances, the choice of the referent can be arbitrary, as for example when the indicators are herd indicators and the herd effects are not of primary interest, but they must be controlled to prevent confounding. Example 14.5 shows the creation of a set of indicator variables for method of colostrum feeding.

Most software programs have automated procedures to create indicator variables, and the coding can be more flexible than shown here. By default, some use the first category of the nominal variable as the referent, others use the last category as the referent. Most allow the user to set the referent using the contextual considerations just mentioned. In Example 14.6, we use regular dummy variables to code for herd when predicting the association of parity (in its continuous form) with milk production.

As noted earlier, all indicator variables (of each nominal variable) usually should be entered or excluded from the model as a set using the F-test in Section 14.3.6. Once the set has been deemed important in a statistical sense or from the perspective of confounding control, it sometimes is desirable to allow only some (*eg* the statistically significant or the 'important'

Example 14.5 Coding indicator (regular dummy) variables

We will demonstrate forming regular (*ie* disjoint) indicator variables from a nominal variable. For example, when conducting a study in which one predictor is method of colostrum feeding we might have coded the answers in the variable -colfeed- as 1=suckling, 2=nipple pail, 3=open pail, and 4=intubation. Let's assume that 'nipple pail' is a sensible referent and has sufficient sample size. The coding of the 3 disjoint variables could be completed by writing logical code to answer the following:

If colfeed=1	then suckle=1	else suckle=0
If colfeed=3	then openpail=1	else openpail=0
If colfeed=4	then tube=1	else tube=0

The effect and significance of each new variable (-suckle-, -openpail- and -tube-) would be in relation to nipple-pail feeding. Whether or not the information in the original variable -colfeed- added significantly to the model should be assessed by an *F*-test as shown in Example 14.3.

indicators) to remain in the model. Removal of unnecessary indicators can aid the development of a more parsimonious model but should be done with caution. The decision about removing some of the indicators can be assisted by testing the equality of selected indicator coefficients. (**Note** To select indicators in a statistically correct sense, multiple comparison procedures which make the P-value for significant differences smaller, must be applied—see Section 11.9.1.) One must be aware that removal of some indicators changes the interpretation of the coefficients for the remaining indicators. For example, when using indicator variables for herd (as above), if only indicator X_1 is in the model, the referent will be the weighted average of the outcome in herds 2 and 3 and the coefficient associated with X_1 will represent the difference in response

Example 14.6 Using and interpreting regular indicator variables in linear regression
data = daisy2

A model for -milk120- was fit with parity, and 6 indicator variables for herds as the predictors.

milk120	Coef.	SE	t	P>t	95% CI	
parity1	201.807	9.291	21.72	0.000	183.583	220.031
herd=2	-117.701	49.079	-2.40	0.017	-213.969	-21.433
herd=3	-678.784	45.371	-14.96	0.000	-767.780	-589.789
herd=4	-380.858	50.494	-7.54	0.000	-479.904	-281.813
herd=5	-563.714	59.074	-9.54	0.000	-679.589	-447.839
herd=106	357.972	47.621	7.52	0.000	264.563	451.381
herd=119	62.074	53.391	1.16	0.245	-42.653	166.800
constant	3047.921	35.931	84.83	0.000	2977.441	3118.401

In this instance, the referent herd is herd 1; the -milk120- in herd 1 for parity 1 cows (parity1=0) is 3,047.9 (the intercept). The coefficient for each herd reflects the difference in -milk120- between each herd and herd 1. We already know that collectively these variables explain a significant proportion of the variance of -milk120-. Individually, all are significantly different from herd 1, except for herd 119 with a P-value of 0.25.

between herd 1 and this average. Any effects from indicators not included in the model are present in the constant term.

14.4.3 Coding hierarchical indicator variables

If the predictor variables are ordinal in type (*eg* reflect relative changes in an underlying characteristic, *eg* severity of milk fever), it is sometimes difficult to associate the levels of severity with specific numerical values that would make it meaningful to use the variable as a continuous predictor. As an example, when coding a variable representing severity (*eg* using 1, 2, or 3 to represent stages 1, 2 or 3 milk fever), there might be concern when using these codes as a continuous predictor (*eg* is the biological effect of the difference between stage 1 and stage 2 milk fever the same as between stage 2 and stage 3?). It is always possible to use regular indicator variables, but they do not reflect the ordering of levels. Therefore, the use of **hierarchical** (or **incremental**) indicator variables is often the preferred approach, in order to maintain the ordering inherent in the original variable. This approach can also be used to recode a continuous variable based on using appropriate cutpoints.

Hierarchical indicator variables contrast the outcome in the categorised version of the original variable against the level just preceding it (assuming all hierarchical variables are in the model). As with regular indicator variables, it is possible to just include a subset of the indicators. One such situation occurs if we are interested in identifying cutpoints of an ordinal or continuous variable where the relationship with the outcome changes. In this setting, we can select the most significant incremental variable(s) for entry. The corresponding coefficient contrasts the outcome in each level of the categorised X-variable to the outcome in the levels below it (Walter *et al*, 1987). Other codings are available, but are beyond the scope of this book—see http://www.ats.ucla.edu/stat/stata/webbooks/reg/chapter5/statareg5.ht m.

In Example 14.7, we coded parity levels using disjoint (*ie* dummy) and hierarchical indicators. The disjoint coefficients reflect the difference in -milk120- in each parity relative to the baseline parity 1 cows; the referent is the lowest level (unless you change it). With the hierarchical indicators, say for parity 5, the regression coefficient reflects the difference in -milk120- between parity 5 cows and parity 4 cows. For some reason, parity 5 cows did not have the increased milk production that was experienced by cows in the the other lactations (a general trend toward increased production with age).

14.4.4 Errors in the X-variables

In the regression model, the X-variables are 'fixed' (*ie* constant), and assumed to be measured without error. In reality, they might be fixed because they are set by the experimenter in a controlled trial (*eg* treatment or dose) or because they represent values that *are* constant (*eg* site or year). However, when the X-variables are measured quantities (*eg* in observational studies), these measurements might contain error: either a natural variation related to the measurements, or error in the sense of misrecordings. The consequence of this error is that relations between the outcome and the observed X-values are not the same as those with the true X-values. The regression model estimates the relationship between the observed X-values and the outcome, and this is the appropriate relationship for purposes of prediction. However, when attempting to describe a causal relationship between the X-variables and the outcome, it is desirable to have the true values of the X-variables.

Example 14.7 Indicator vs hierarchical coding of variables
data = daisy2

The effect of parity on -milk120- was estimated by using ordinary (disjoint) indicator and hierarchical dummy variables in a linear regression model

variable	indicator coding	hierarchical coding
parity=2	708.213	708.213
parity=3	789.843	81.630
parity=4	848.514	58.670
parity=5	787.609	-60.905
parity=6	878.161	90.552
parity=7	925.955	47.794
constant	2639.645	3345.116

Special models exist for taking error in the *X*-variables into account, so-called **measurement error models**, but they are beyond the scope of this book (Fuller, 2006). Nonetheless, many software programs support the use of regression calibration (see Section 12.8) which is useful for adjusting for measurement errors. Murad and Freedman (2007) have extended this to the situation when an interaction term between 2 covariates, each measured with error is being assessed. Austin and Hoch (2004) describe methods to adjust the regression when one or more X-variables are censored. However, as indicated in Chapter 12, if the magnitude of the measurement error is small relative to the range of the *X*-values in the model, we need not be unduly worried when using the ordinary regression model. Ignoring measurement errors generally tends to bias the parameters towards the null (*ie* effects will be (numerically) smaller than if the completely accurate information was present). On the other hand, if the errors are large relative to the range of *X*-values serious consideration of the need for validation studies (see Chapter 10) is in order.

14.5 DETECTING HIGHLY CORRELATED (COLLINEAR) VARIABLES

Despite the fact that multiple regression is used to adjust for correlations among predictor variables in the model, if the variables are too highly correlated then a number of problems arise. Before discussing these, recall that in a multivariable regression model the estimated effect of each variable generally depends on the other variables in the model. On one hand, this is the advantage of a multivariable analysis—that the effect of a variable is studied while taking into account the correlations between that variable and others in the model and their effects on *Y* thereby avoiding duplication of effects. On the other, this means that the effect of any variable might change when other variables are added to, or removed from, the model. If, for a particular variable, such changes are large (*eg* involving a shift of sign), its interpretation becomes difficult. Only in the special case that all the *X*-variables are uncorrelated are the effects of different variables estimated completely independently of each other. Thus, the first problem arising from highly correlated (or collinear) predictors is that estimated effects (*ie* the regression coefficients) will depend strongly on the other predictors present in the model. As a consequence, it might be difficult to statistically select the 'important' predictors from a larger

group of predictors. Both of these concerns are less serious when the purpose of the analysis is prediction than when interpretation of causal effects is the objective. If we express this problem in a more technical manner, the standard errors of regression coefficients become very large in a highly collinear model (Section 14.5.1), and hence we become less certain of the likely magnitude of the association (*ie* of the true value of β).

In a multivariable model, one *X*-variable should not be a perfect mapping of another *X*-variable or be perfectly predictable by a combination of the other *X*-variables in the regression model. However, even before the correlations become 'perfect', as a general rule, if 2 (or more) variables are highly correlated ($|\rho| > 0.8$–0.9), it will be difficult to select between (among) them for inclusion in the regression equation. When 2 variables are highly and positively correlated, the resulting coefficients (βs) will be highly and negatively correlated. In extreme situations, none of the coefficients of the highly correlated variables will be declared significantly different from zero, despite the fact that the *F*-test of the model might indicate that the variable(s) collectively contributes significantly to the model.

Extreme values of odds ratios (*eg* 8-10 of more) can be used to detect collinearity between dichotomous variables, and extreme correlation coefficients (>0.7-0.8) for continuous variables. In linear models, a convenient way to detect either collinearity or multicollinearity is through the use of the variance inflation factor (Section 14.5.1). Pitard and Viel (1997) describe more formal methods for detecting collinearity and provide some solutions when using regression models.

One way of eliminating collinearity problems is through considered exclusion of one of the collinear variables, or by making a new combination of the variables on substantive grounds. In extreme situations specialised regression approaches, such as ridge regression, might be needed.

Most software provide indicators about possible collinearity using a **variance inflation factor** (Section 14.5.1) or its reciprocal **tolerance**. Unfortunately, the methods we use for including interaction terms (Section 14.6) and power terms (Section 14.9.3) in models sometimes leads to a high collinearity between the variables. Thus, we describe a general method for circumventing high correlations between the latter constructed variables, known as **centring** (Section 14.5.2). Before doing that we will discuss the problem of collinearity in terms of variance inflation.

14.5.1 Variance inflation factors

The effect of entering a new variable into the model, on the variance of the coefficients for variables currently in the model can be assessed with a statistic known as the **variance inflation factor** (*VIF*). The formula for *VIF* is:

$$VIF = \frac{1}{1 - R_X^2} \qquad \textit{Eq 14.10}$$

where R^2_X is the coefficient of determination from regressing the variable that is about to enter the model on the other variables in the model. As this coefficient gets larger (as it does if it is collinear) so does the *VIF*. We illustrate the importance of the *VIF* in a simple linear regression model, in which the variance of the regression coefficient β_1 for X_1 is from Eq 14.5.

$$\text{var}^{(1)}(\beta_1) = \frac{\text{MSE}^{(1)}}{\text{SSX}_1} \qquad \textit{Eq 14.11}$$

where the superscript (1) refers to the simple linear regression model. When we place X_2 in the model, if it is correlated with X_1, 3 things will happen:

1. the coefficient β_1 will change because we account for the correlation of X_1 with X_2,
2. the residual sum of squares (and in most cases also the $\text{MSE}^{(2)}$) will become smaller because X_1 and X_2 together can predict Y better than X_1 on its own, and
3. the standard error of β_1 might increase by an amount roughly equal to $\sqrt{VIF}$; specifically, var(β_1) in the combined model (2) with both X_1 and X_2 is:

$$\text{var}^{(2)}(\beta_1) = \frac{\text{MSE}^{(2)}}{\text{SSX}_1} \times \frac{1}{(1 - R_2^2)} \qquad \textit{Eq 14.12}$$

where R^2_2 is the coefficient of determination from a regressing X_2 on X_1. Thus, the standard error of β_1 increases unless the reduction in $\text{MSE}^{(2)}$ from $\text{MSE}^{(1)}$ by adding X_2 more than offsets the increase due to the *VIF*. Adding variable X_2 also can cause the variance of β_1 to decrease if X_2 is a good predictor of the outcome and X_1 and X_2 are nearly (or totally) independent of each other in which case $\sqrt{VIF}$ is approximately 1.

The role of the *VIF* in multiple regression models is similar to this. A (conservative) rule of thumb for interpreting *VIF*s is that values above 10 indicate serious collinearity. As discussed above, this does not necessarily mean that the model is useless or that one is obliged to remove one or more *X*-variables from the model; it should, however, always be taken as a warning for the interpretation of regression coefficients and the increase in their standard errors.

14.5.2 Centring variables to reduce collinearity

Centring a continuous variable is performed by subtracting the mean value (or some other central value) from each observed *X*-value, similarly to the scaling discussed in Section 14.4.1. Centring a variable prior to creating a power term (or an interaction term between 2 continuous variables) reduces the correlation between the variables to a low level (provided the variables are symmetrically distributed about their mean). If the distribution is not symmetric, then larger (or smaller) values than the mean might need to be subtracted. It should be stressed that centring only affects correlations between variables constructed from each other, and it does not change the predictions or the fit of the model, only the values and interpretation of the regression coefficients. See Example 14.8 for a discussion of *VIF*s and centring.

14.6 DETECTING AND MODELLING INTERACTION

In Chapter 1 we developed the view that, given the component cause model, we might expect to see interaction when 2 factors act synergistically or antagonistically. Whereas, within limits, this might be true, the significance of an interaction term need not indicate anything about the causal model; it might merely describe the nature of the relationship being modelled. In the previous Section, the model contained only **main effects** of the *X*s; hence it assumes that the association of X_1 to Y is the same at all levels of X_2 and the association of X_2 to Y is the same at all levels of X_1. A test of this assumption (whether or not the effect of one variable depends on

Example 14.8 The use of centring to avoid collinearity problems
data = daisy2

Our outcome of interest is now the time from the end of the 'wait period' (*ie* the time at which a producer will start breeding his/her cows) to conception (-wpc-; median 53, mean=68.8, SD 51.6, and range 1—298 days). One of the potential confounders of the effects of diseases and -milk120- on -wpc- is herd size. In order to prevent bias from herd size, we need to include it in our model. However, the relationship between herd size and -wpc- appeared non-linear (see Chapter 15 for methods of assessment), so a quadratic model with -hs100- (herd size scaled by dividing by 100) and -hs100_sq- (scaled herd size squared) was built.

wpc	Coef	SE	t	P>t	95% CI	
hs100	-29.516	15.044	-1.960	0.050	-59.024	-0.008
hs100_sq	9.744	3.105	3.140	0.002	3.654	15.835
constant	77.748	17.475	4.450	0.000	43.472	112.024

Both terms are statistically significant. However, the correlation between -hs100- and -hs100_sq- is 0.99 with a resulting *VIF* of 54. We can see the impact of this by noting that the SE of the herd size coefficient increased by over 7 times (from 2.1 (simple linear model not shown) to 15) when the quadratic term was added.

In order to deal with this problem of collinearity, we can centre the herd size variable by subtracting its mean (for -hs100- this is 2.5) to create the centred variable -hs100_ct-, and then, we create the squared centred herd size -hs100_ctsq-. The summary of this model is shown below:

wpc	Coef	SE	t	P>t	95% CI	
hs100_ct	19.400	2.158	8.990	0.000	15.167	23.633
hs100_ctsq	9.744	3.105	3.140	0.002	3.654	15.835
constant	65.052	1.742	37.340	0.000	61.635	68.470

First, we should note that the coefficients and SEs for the quadratic terms are exactly the same in the 2 models, but the coefficient for the linear term has changed. The 2 models also have identical R^2 (0.049) and MSE (2,535.3). Second, we note that the SE of the linear component -hs100_ct- is approximately back to what it was when only the linear term (-hs100-) was in the model. Centring has reduced the correlation between -hs100_ct- and -hs100_ctsq- to -0.32 and the *VIF* is now reduced to 1.11. Because herd size was scaled, the constant in this model represents the predicted -wpc- in a 250-cow herd (-hs100_ct-=0).

the level of another variable(s)) is to examine if an 'interaction term' adds significantly to the regression model.

In the situation where X-variables are not indicator variables, the interaction term is formed by the product X_1*X_2 which can be tested in the following model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \varepsilon$$

by assessing if β_3=0 (see Example 14.9). If interaction is absent (β_3 is deemed to be not different from 0), the main effects (or 'additive') model is deemed to describe the effects adequately. It is not necessary to centre the variables X_1 and X_2 to see if an interaction term is needed, because β_3 and its standard error will be unaffected by centring. However, if the interaction is needed, centring might be useful because it allows us to interpret β_1 and β_2 as linear effects when the interaction cancels (*eg* β_1 applies to the situation when (the centred version of) X_2 is zero).

Higher order interactions can be investigated by extending this process to an interaction term that is the product of 3 (or more) variables (see Chapter 15).

Interactions involving categorical variables (with more than 2 levels) are modelled by including products between all indicator variables needed in the main effects model. For example, the interaction between a 3-level and a 4-level categorical variable requires (3-1)*(4-1)=6 product variables. These 6 variables should be tested and explored as a group (Section 14.3.6). In many multivariable analyses, the number of possibilities for interaction is large and there is no single correct way to assess if interaction is present. Section 15.7 discusses some options for deciding which interaction terms to include when building a multivariable model. However, in general we suggest that, unless the potential number of interactions is small, interactions be limited to those of biological relevance and that 3- and 4-way interactions only be investigated when there are good, biologically sound, reasons for doing so. Example 14.9 demonstrates interaction between 2 dichotomous variables, and Example 14.10 between a dichotomous and a continuous predictor. There are no statistically significant interactions between continuous variables in -daisy2-, but Example 14.11 shows a non-significant interaction for demonstration purposes only.

14.7 CAUSAL INTERPRETATION OF A MULTIVARIABLE LINEAR MODEL

So far in this chapter we have focused on the technical interpretation of regression coefficients. Example 14.12 is presented to focus on the development of, and causal interpretation of, a multivariable linear model designed to assess the effects of 3 diseases (-dyst-, -rp-, and -vag_disch-) on the time to conception. When making causal inferences, care is needed to ensure that only the appropriate variables are included in the analysis (see Section 13.3). In this regard, a causal diagram is very helpful (see Fig. 14.4).

Based on this diagram, and our objective which is to ascertain the effects of the 3 diseases on -wpc-, we will not include -milk120- or days to first service in our model because they are intervening variables between these diseases and -wpc-.

Before presenting the model, a few comments are in order. Herd effects can be potent confounders so we need to control for herd; we could do that by including a set of 6 indicator variables in the model. However, this would preclude having any herd level predictors in the model (they are perfectly collinear with the herd indicator variables). One aspect of the herd effect is herd size and when we examined its relationship to -wpc-, we noted that it was curvilinear so we created a quadratic term after centring herd size. Because we only have 7 herds, and because we want to include a continuous variable in our model for pedagogical purposes, we chose to include herd size instead of herd (using dummies). (**Note** Because herd size is a herd-level predictor and we are not taking the clustering of cows within herds into account, we might expect the SE for the effect of herd size to be underestimated. See Chapters 20-24 for methods for dealing with clustered data.) Month of calving could also have an impact on -wpc- and when we examined this we noted that cows that calved in February through July had a shorter time to conception; hence, we created a new variable called -aut_calv- (*ie* autumn calving) to denote these cows. Finally, although the variable parity was always non-significant we forced it into the causal model based on previous evidence that, in general, higher parity cows do not have as good reproductive performance as lower parity cows and that parity is related to the 3 diseases of interest (so it could/should be a confounder). Other comments on model-building are contained in Chapter 15, but here we include all potential confounders, then

Example 14.9 Interaction between 2 dichotomous variables
data = daisy2

A model was built with only -rp- and -vag_disch- as predictors of -wpc-.

wpc	Coef	SE	t	P>t	95% CI	
rp	10.240	4.541	2.260	0.024	1.333	19.146
vag_disch	9.067	5.982	1.520	0.130	-2.666	20.800
constant	67.358	1.381	48.770	0.000	64.649	70.067

In this model, -rp- has a significant association with -wpc-, but -vag_disch- does not. In order to assess if the effect of one variable depends on the level of the other variable we form an interaction term (a product of the 2 variables) and add it to the model.

wpc	Coef	SE	t	P>t	95% CI	
rp	6.340	4.914	1.290	0.197	-3.300	15.979
vag_disch	0.543	7.265	0.070	0.940	-13.708	14.794
rp * vag_disch	26.349	12.774	2.060	0.039	1.293	51.404
constant	67.669	1.388	48.760	0.000	64.946	70.391

Note that since both -rp- and -vag_disch- are dichotomous and coded 0 for 'no' and 1 for 'yes', the interaction term has the value 1 only when both diseases are present. In this sense, if it is significant, it says that we need to adjust (using β_3) the predicted outcome when both diseases are present to better reflect what was observed. Otherwise the combined effect of the 2 is just the sum of their individual effects.

In daisy2, it is apparent that the effect of -rp- depends on the presence or absence of -vag_disch- as the coefficient is significant statistically. Note also that, although the main effect terms of -rp- and -vag_disch- are not significant, we leave them in the model for interpretation.

- When neither disease is present, the predicted outcome is 67.67 days
- When only -rp- is present, the predicted outcome is 67.67+6.34=74.0 days
- When only -vag_disch- is present, the predicted outcome is 67.67+0.54=68.21 days
- When both diseases are present, the predicted outcome is 67.67+6.34+0.54+26.35=100.9 days
- The *VIF* from adding the interaction is small (1.8), so centring of variables is not needed

What this model implies is that neither disease by itself has much of an impact on -wpc- but when both are present, -wpc- is delayed by about 33.2 days (relative to neither disease being present).

added our disease variables of interest (these were our key exposure variables). Interactions between the diseases and other variables (except for herd size where we assumed no interaction was present) were selected based on prior biological knowledge. Had we used statistical significance as our method of detecting interactions, we would have discovered an interaction between -vag_disch- and -dyst- that defied (at least to us) explanation. We are also aware that had we selected a subset of variables from a larger 'pool' we should alter the *F*-statistic critical value for significance (Livingstone & Salt, 2005).

14.8 EVALUATING THE LEAST SQUARES MODEL

Valid regression analyses are based on a set of assumptions, and once our initial model is built we need to evaluate whether the model meets these (we say initial because after checking

Example 14.10 Interaction between a dichotomous and a continuous variable
data = daisy2

We began this analysis by scaling -milk120- by dividing it by 1,000 (denoted as -milk120k-); otherwise, the effect of a one kg change in -milk120- was very small. We then regress -wpc- on -dyst- (dystocia), -milk120k- and their interaction (denoted as -dyst_milk-) to see if the effect of -milk120k- on -wpc- depends on whether or not the cow has -dyst- (or conversely, if the effect of -dyst- depends on the level of milk production. (**Note** Neither -dyst- nor -milk120k- were statistically significant when fit without an interaction term.)

wpc	Coef	SE	t	P>t	95% CI	
dyst	-85.488	29.601	-2.890	0.004	-143.551	-27.426
milk120k	-3.447	1.929	-1.790	0.074	-7.229	0.336
dyst_milk	29.142	9.468	3.080	0.002	10.571	47.714
constant	79.838	6.365	12.540	0.000	67.352	92.323

The interaction term is clearly significant. Dystocia appears to shorten the -wpc- which is a surprising result, but this coefficient represents the effect of -dyst- when -milk120k- is zero (which is not possible). Increasing milk production appears to have a small negative effect (-3.4 days per 1,000 kg of milk) in cows without dystocia (P=0.074) but potentially a larger positive effect in cows with dystocia. In a situation such as this, a graph is more likely to make the interaction effects apparent. This is easily accomplished by obtaining the predicted -wpc- from the model and graphing it against the continuous predictor (-milk120-) in cows with and without dystocia (Fig. 14.2).

Here we can see the difference in effect of -milk120k- between when -dyst- is present (the solid sloped line) and when it is absent (the dashed nearly horizontal line). The graph indicates that increasing levels of milk production is detrimental (longer -wpc-) in cows that start the lactation with dystocia, but not in cows without dystocia. If interaction was absent, the regression lines on -milk120k- in cows with and without -dyst- would be parallel.

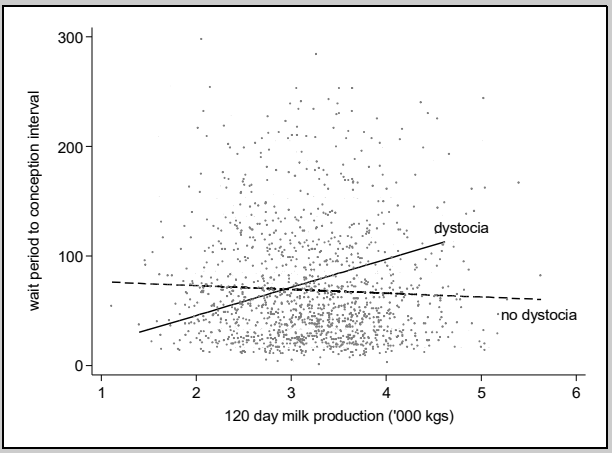

Fig. 14.2 Interaction between -dyst- and -milk120k-

whether the model meets the key assumptions we might have to alter it). We will use the model shown in Example 14.12 for the purposes of this evaluation.

The key assumptions of the model are:

- **independence**—the values of the dependent variable are statistically independent from one another (*ie* the -wpc- value of one cow does not depend on the -wpc- value of other cows in the dataset). Usually we do not worry about independence unless the context is such that the assumption is likely to be broken. For example, the structure of the data might signal a lack of independence when there are multiple observations on a single animal, or on multiple animals within herds (as we have in -daisy2-). Methods for dealing with clustered data are presented in Chapters 20-23. A specific type of clustering (serial

Example 14.11 Interaction between two continuous variables
data = daisy2

Here we regress -wpc- on -parity-, -milk120k- and their interaction (denoted as -p_m-) to see if the effect of -milk120k- on -wpc- depends on the parity of the cow (or conversely if the effect of parity depends on the level of -milk120k-).

wpc	Coef	SE	t	P>t	95% CI	
parity	4.890	4.438	1.100	0.271	-3.815	13.595
milk120k	-1.666	4.213	-0.400	0.692	-9.929	6.597
p_m	-0.876	1.364	-0.640	0.520	-3.551	1.798
constant	69.022	12.777	5.400	0.000	43.961	94.083

It turns out that the interaction term is no where near significant statistically; however, we show the plot (Fig. 14.3) to demonstrate this. The lines reflect the expected decrease in -wpc- as -milk120k- increases. Because the interaction term is not significant we expect the lines (representing different levels of parity) to be parallel which, for practical purposes, they are.

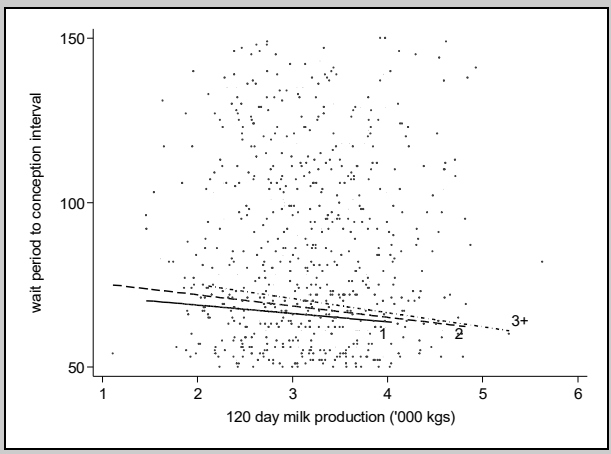

Fig. 14.3 Interaction between -parity- and -milk120k-

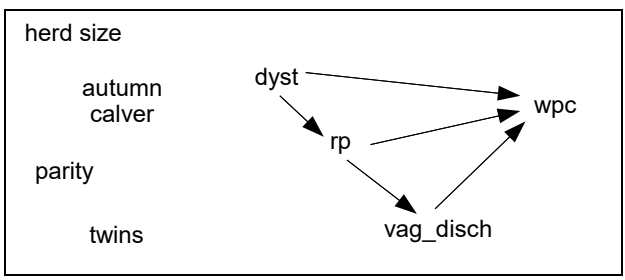

Fig. 14.4 Causal diagram of 3 diseases and 4 confounders 'wait-period-to-conception interval' in dairy cows

Note Variables to the left are assumed to be correlated and have potential effects on all variables to the right (causal arrows omitted for clarity of presentation)

Example 14.12 An initial causal model of the impact of reproductive diseases on -wpc-
data = daisy2

A model was fit based on the causal diagram shown in Fig. 14.4.

Source	SS	df	MS	
Model	296062.7	9	32895.9	Number of obs = 1574
Residual	3892027.9	1564	2488.5	F(9, 1564) = 13.22
Total	4188090.6	1573	2662.5	Prob > F = 0.0000
				R-squared = 0.0707
				Adj R-squared = 0.0653
				Root MSE = 49.885

wpc	Coef	SE	t	P>t	95% CI	
aut_calv	-8.264	2.538	-3.26	0.001	-13.242	-3.286
hs100_ct	19.857	2.163	9.18	0.000	15.614	24.101
hs100_ctsq	11.138	3.111	3.58	0.000	5.036	17.241
parity_sc	1.137	0.858	1.32	0.185	-0.546	2.821
twin	20.683	9.845	2.10	0.036	1.372	39.994
dyst	11.700	5.463	2.14	0.032	0.986	22.415
rp	5.987	4.812	1.24	0.214	-3.452	15.425
vag_disch	1.228	7.161	0.17	0.864	-12.819	15.275
rp*vag_disch	22.852	12.516	1.83	0.068	-1.698	47.402
constant	64.330	2.634	24.42	0.000	59.164	69.497

Subject to this model meeting the major assumptions of linear regression (Section 14.9) and a case-by-case analysis of the residuals (Section 14.10), we offer the following interpretation:

The model is highly significant (F=13.22 $p<0.001$) although it only explains 7% of the overall variation in -wpc-. The SE of prediction (49.9 days) is only slightly smaller than the original crude SD of 51 days.

Cows calving in the autumn had a decreased -wpc- by 8d relative to those that calved at other times. The effect of herd size was to increase the time to conception in a curvilinear manner with greater effects as herd_size increased. Scaled parity (-parity_sc-) was left in the model in the belief that it should have been an important confounder. Cows with twins took 21d longer to conceive than cows with single calves. Having controlled for the effects of these variables, and assuming there was no additional uncontrolled confounding we can make causal inferences on the effects of the 3 diseases.

- Cows with -dyst- had a significant delay in -wpc- of 11.7 days
- Cows with -rp- but not -vag_disch- did not have a significant delay in -wpc-
- Cows with -vag_disch- but not -rp- did not have a significant delay in -wpc-
- Cows with both -rp- and -vag_disch- had a significant delay in -wpc- of approximately 23+6+1=30 days (The estimate in Example 14.9 is slightly different because the earlier model had fewer variables included).

correlation) is likely to occur when assessing if the daily average temperature affects the daily milk production of cattle during a period of time (*eg* summer). Repeated data that are collected at equal time intervals over an extended period such as this are called time-series and specific methods are required to adjust for the fact that the value of the outcome on one day is likely highly correlated with the value on the previous day; hence, the errors are correlated and not independent (see Section 14.11).

- **homoscedasticity**—the variance of the outcome is the same at all levels of the predictor variables (*ie* the variance in -wpc- in cows that have a -parity- of 3 should be the same as the variance for those that have a -parity- of 5 *etc*) and within all combinations of the values of the predictor variables. If this is true, then the MSE will be constant. This is an important assumption, perhaps more so than having a normal distribution of residuals.
- **normal distribution**—the residuals should be normally distributed at all levels of the predictors, or at all combinations of predictors in the model (*ie* residual values for cows that did not have an -rp- should be normally distributed as they should for cows that had an -rp-). We often try to get a quick assessment of this before starting the regression analysis by assessing the normality of the distribution of the outcome. The residual errors from very non-normally distributed outcomes are unlikely to be 'normalised' by regression on the predictor variables unless the R^2 of the model is very high. On the other hand, as a simple example, if a strong dichotomous predictor for the outcome exists, then the raw distribution of the outcome will show as bimodal and therefore non-normal, but the residuals from the model might be normally distributed.
- **linearity**—because the relationship between the outcome and continuous or ordinal predictors (modelled as continuous) is described by a single coefficient, this assumes that the association is a straight-line relationship (*ie* a 1-unit increase in -parity- from 2 to 3 affects -wpc- by the same amount as a 1-unit increase from 5 to 6 kg). There is no assumption involved for dichotomous variables as 2 points can always be connected by a straight line.

Each of the last 3 assumptions will now be discussed in more detail, and we can learn much about them by examining residuals, often using graphical methods, although formal tests are also available. (**Note** Whether or not the observations are independent is usually known from the structure of the data and will not be discussed further in this section (see Chapters 20-23 for further discussion of this issue.)) At this point, we would note that ensuring our model meets the 3 major assumptions (homoscedasticity, normality, linearity) is very important, and alterations to meet one of these assumptions can influence the validity of the other 2 assumptions. In order to expedite model-building, we suggest a cursory examination of these major assumptions early in the model-building process. If any of the major assumptions are obviously violated at this stage, we would suggest instituting whatever changes are necessary to 'improve' the fit before serious model-building. We have ignored that principle to date in order to keep the model 'simple' and explain the basic features of linear regression. Once we are satisfied that these 3 major assumptions have been met, we should pursue a more detailed search for specific observations that might be outliers, leverage points, and/or influential points. Because of the importance of residuals in these assessments, we begin by describing different types of residual.

14.8.1 Residuals

The **raw residual** (r_i) is the difference between the observed and predicted value for the i^{th}

observation and has the same units as the outcome variable,

$$r_i = Y_i - \hat{Y}_i \qquad \textit{Eq 14.13}$$

where the subscript i denotes the particular observation on subject 'i' from the 'n' study subjects. The raw residual r_i is our 'estimate' of the error for observation i, by subtracting its predicted mean from the observed value.

The mean of all residuals is zero, and the variance of each residual is:

$$\text{var}(r_i) = \sigma^2(1 - h_i) \qquad \textit{Eq 14.14}$$

where h_i is the weight of the i^{th} observation in determining r_i. The h_i is called the **leverage** of that observation and indicates the potential for this observation to have a major impact on the model. In a simple regression model, h_i has the following formula:

$$h_i = \frac{1}{n} + \frac{(X_{1i} - \overline{X}_1)^2}{\text{SSX}_1} \qquad \textit{Eq 14.15}$$

indicating that as the value of the predictor gets farther from its mean, the leverage of the observation increases. Note that this 'potential' impact depends only on the predictor, not on the value of the outcome. Leverage has a more obvious meaning when the predictor is measured on the continuous scale. We return to the subject of leverage in Section 14.10.2.

The raw residuals can be scaled by dividing them by their SE. If all observations are used to estimate σ^2 this produces what are called **standardised (std) residuals** (these are also called internally studentised residuals):

$$r_{si} = \frac{r_i}{\sigma\sqrt{1 - h_i}} \qquad \textit{Eq 14.16}$$

The reference distribution for standardised residuals is a t with (dfE), so for sample sizes with n>30, based on the Gaussian distribution, there should be only about 5% of values outside of the interval (-2, 2). The major advantage of standardised residuals relative to raw residuals is that we have this absolute scale for what constitutes a large residual.

The raw and standardised residuals are computed from the prediction for the i^{th} observation from the regression equation based on all observations. That is, the observation itself contributes to the prediction. An influential observation might not show a large residual because of its impact on the prediction. To 'truly' examine whether the i^{th} observation is in agreement with the model, we should compare it with the prediction based on the other n-1 observations. Such (standardised) residuals are called **studentised (stu)** residuals or externally studentised residuals (others denote them as **deletion** residuals, or **jackknife** residuals):

$$r_{ti} = \frac{r_{-i}}{\sigma_{-i}\sqrt{1 - h_i}} \qquad \textit{Eq 14.17}$$

where the '-i' notation indicates that observation i is not included in the prediction or the model's variance. These residuals are distributed as a t-distribution (with dfE-1; Table 14.2), assuming the model is correct.

To summarise, standardised residuals might yield a large value if:

- the observation is an outlier in the response (Y) variable (*ie* r_i is large), or

- the observation is an outlier in the predictor variable(s) (*ie* h_i is large).

Studentised residuals might be large if either of the above is true, or if the observation strongly affects the fit of the model (*ie* the model changes considerably when the observation is removed).

We now proceed to use data on the residuals to assess the overall fit of the model. Although we separate the study of homoscedasticity from normality, in practice one should look at both, as well as linearity before deciding on modifications (*eg* transformations) to the variables.

14.9 EVALUATING THE MAJOR ASSUMPTIONS

In general, evaluating the model assumptions relies heavily on graphical methods, although a large battery of statistical tests exists for evaluating different assumptions. However, we recommend the tests to be used only as a supplement to the graphical methods, and that caution should be exercised when tests and graphics lead to different conclusions.

14.9.1 Homoscedasticity

A constant variance of residuals is an important assumption in linear regression. Without equality of variance (a situation called **heteroscedasticity**), the significance tests are at best only approximate because the standard error is too small for some values and too large for others. One can examine the homoscedasticity assumption, by plotting the standardised residuals against the predicted values. If the variance is constant across the range of predicted *Y*-values, then a scatter of points resembling a horizontal band will result. If the variance is not constant, a pattern such as fanning (increased variance with larger predicted values), or coning (decreased variance with larger predicted values) might result. These patterns suggest that the dependent variable might need to be transformed (or a weighted regression used). It might also be useful to plot standardised residuals against individual (continuous) predictors and look for similar patterns, and to compare the residual variances in the groups formed by levels of categorical variables.

A number of statistical tests for heteroscedasticity exist and a commonly used one is Breusch-Pagan test (also known as the Cook-Weisberg test) (Cook & Weisberg, 1983). The null hypothesis is homoscedasticity so a significant ($P<0.05$) test results indicates the presence of heteroscedasticity. An evaluation of heteroscedasticity is presented in Example 14.13.

14.9.2 Normality of residuals

To examine for normality, one can plot the residuals in the form of a histogram (Example 14.14). An alternative, and more sensitive display, is a normal probability plot (sometimes called Q-Q (quantile-quantile) plot) for the residuals. If the residuals are normally distributed, the resulting plot will be (approximately) a straight line at 45° to the horizontal (see right side figure in Example 14.14). If the residuals are skewed to the right, the normal plot will curve below the 45° line (the curve is convex), whereas, if the residuals are left skewed, the normal plot will curve above the 45° line (the curve is concave). If the residuals are too peaked (platykurtic), the normal plot will be sigmoid curved. Whether such departures from normality are most easily visualised in the normal plot or the histogram is largely a matter of taste. As an

Example 14.13 Evaluation of homoscedasticity (equal variances) with -wpc- as the outcome
data = daisy2

A scatterplot of standardised residuals vs predicted values based on the model presented in Example 14.12 was generated.

An approximately equal-width band of points suggests the model meets the assumption of equal variances. Visually, it is difficult to detect any pattern to the residuals. However, the Cook-Weisberg test for hetero-scedasticity yields a χ^2-statistic of 20.58 with 1 df. This very significant result (P<0.001) indicates a non-constant variance. Computing the SD of the residuals in ranges of predicted values with cutpoints of 40, 60, 80 and 100, suggests that the variance is smaller at low and high values of predicted values than in the middle 2 categories (data not shown).

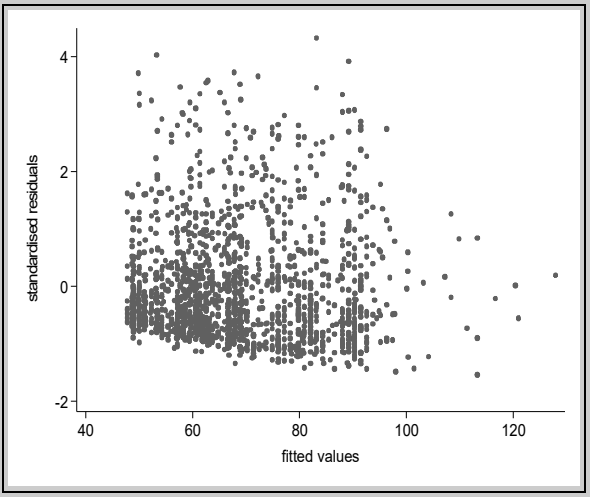

Fig. 14.5 Scatterplot of standardised residuals vs fitted values

aid for the interpretation, the skewness and kurtosis of the standardised residuals can also be computed.

Many statistical tests for normality are available, with one of the commonly used tests being the Shapiro-Wilk test. The null hypothesis is that the distribution is normal, so a significant P-value (<0.05) is an indication of non-normality. However, our experience is that with large sample sizes, this test often yields a significant result when only mild departure from normality is evident. Consequently, we often rely more heavily on visual assessment (especially of a Q-Q plot.

14.9.3 Correcting error distribution problems: transformations of the outcome

There are a number of possible transformations of the outcome variable, but only the more frequently used ones are mentioned here. Most software programs provide a variety of easily accessed transformations so that we can readily try different approaches. The selection of the correct transformation also is aided by knowledge of what has worked in similar situations in the past, although formal assessment of the appropriate transformation remains useful (Afifi *et al*, 2007). Some general rules are:

- if the variance of the residuals increases mildly (*ie* proportional to the mean) with the outcome, a square-root transform of *Y* may prove useful ,
- if the 'fanning' is stronger (proportional to the mean squared) a logarithmic transformation of *Y* often works,
- if the 'fanning' decreases with the outcome and the relationship of *X* and *Y* is nearly linear, a reciprocal transformation of *Y* could prove helpful,

Example 14.14 Evaluating normality of residuals with -wpc- as the outcome
data = daisy2

The histogram on the left and the Q-Q plot (which displays the quantiles of the residuals versus the quantiles of the normal probability distribution) indicate moderate non-normality. Here we have a dish-shaped (convex) Q-Q plot consistent with the right skewed distribution of residuals.

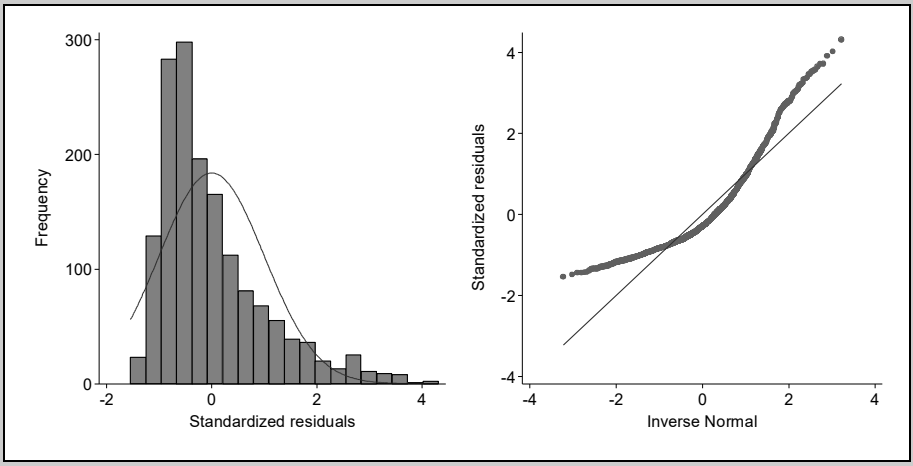

Fig. 14.6 Histogram and Q-Q plot of standardised residuals

Further evidence of a lack of normality can be obtained from a test for a normal distribution. The Shapiro-Wilk's statistic has a value of W=0.88 (small values of W are critical for a normal distribution) with p<0.001, indicating non-normality. The residuals are clearly not normally distributed so we need to consider improving this aspect of our model, as is discussed in subsequent examples and summarised in Section 14.9.3. Following transformation using -wpc_sqrt- the graphs appeared 'visually better'; however, W=0.96, and the Shapiro-Wilk's test remained highly significant.

- if Y is a proportion (p) (or more generally, an outcome in a bounded interval but without a binomial denominator) the variance-stabilising transformation for proportions is: $\arcsin(p^{1/2})$.

Sometimes a more formal approach to identifying the optimal transformation is needed. In this regard, if we are concerned about a lack of normality, there is a family of transformations called **Box-Cox transformations**. The intent here is to determine the power transformation Y^{λ} (except for λ=0, see below) which will make the distribution of the errors as close to an independent Gaussian sample as possible. The Box-Cox analysis, available in most software, computes the value of λ which best 'normalises' the errors using an iterative maximum-likelihood procedure. These transforms can only be used on positive numbers (*ie* >0), but they can be applied to the outcome variable, the predictor(s) or both. Some examples of Box-Cox transformations (where Y^* is the transformed value of Y) are:

- if $\lambda = 1$, we use $Y^* = Y$
- if $\lambda = 1/2$, we use $Y^* = \sqrt{Y}$ (square root of Y)
- if $\lambda = 0$, we use $Y^* = \ln Y$,
- if $\lambda = -1$, we use $Y^* = -1/Y$.

Usually it is sufficient to round the estimated λ to the nearest 1/4 unit (*ie* λ=0.45 would be λ=1/2), or to pick a 'nice' value within the 95% confidence interval for λ. In the model from

Example 14.12, λ=0.11 which is close to 0, indicating that a log transformation might be appropriate. As noted above, we used this in the next formulation of our model. **Note** If there are 0 values to be transformed, you should add a small number (usually the lowest observed value of *Y* in your data) to *Y* before making the log transform (Afifi *et al*, 2007).

Note that the analysis for homoscedasticity and normality should be based on the residuals (from an appropriate linear model), not on the distribution of the outcome itself. It should also be noted that Box-Cox is only one (but commonly used) type of transformation; there is no guarantee that the optimal λ works well (only that it is the best among the power transforms), and many other transformations might be relevant. For example, if the distributional problem with the residuals is mainly one of skewness, an alternative transform is of the form $Y^*=\ln(Y-c)$, where *c* is a value to be selected to help correct the skewness. An advantage of this transform is that it is not constrained to transforming only positive numbers; but $Y-c$ must be positive.

14.9.4 Linearity of predictor-outcome association

In a regression model, we assume that the relationship between the continuous predictor and the outcome is linear. Most software regression packages will allow graphical assessment of linearity, some only in a univariable model, others in multivariable models. With multiple continuous variables in the model, one approach to detecting non-linearity is to plot the residuals against each of the continuous predictor variables (see Example 14.15). The sensitivity of this process can be increased by using a kernel smoothing function to help you visualise any pattern that might be present, but be careful of patterns in areas where the data are sparse. Methods for assessing linearity and dealing with non-linearity are discussed much more fully in Section 15.6. However, 3 possible approaches to resolving a nonlinearity problem will be mentioned here. The first is to add a power term of *X* (*eg* quadratic). The second approach is to try to transform the *Y*-variable (as discussed below). The third is to categorise the continuous predictor and include either regular or hierarchical indicator variables in the model in place of the continuous predictor variable. Example 14.15 shows a lowess smoothed curve to help evaluate the linearity of the relationship between herd size and -wpc-.

Suggestions for correcting a lack of linearity by transformation
In order to correct a lack of linearity, we can transform the outcome or the predictor(s) or both. As will become apparent, we often have to use transformations to correct for heteroscedasticity and lack of normality also. Sometimes correcting for one problem solves others, but sometimes correcting one problem makes a new problem on the other fronts. If we transform the outcome variable to improve linearity, this will definitely affect the variance and normality of residuals so these must be checked after transforming the outcome variable. Indeed, we might have to rebuild the model. If we transform the offending predictor(s), then the variance and normality of residuals are likely to remain relatively stable. Thus, often the route of choice for improving linearity is to test quadratic, or other power transformations of the predictor(s) within a power of ±2 to assess their significance. The following are guidelines:

- if the outcome increases, at a decreasing rate with *X*, then try a $\ln X$ or a $X^{1/2}$ transformation
- if the outcome increases, at an increasing rate with *X*, then try X^2 or e^X
- if the outcome decreases, at a decreasing rate with *X*, then try X^{-1} or e^{-X}

If the relationship is more complex, it may be necessary to use more complex polynomial models or hierarchical indicators instead of the continuous-scaled variable (see Section 15.6).

Example 14.15 Evaluating linearity between herd size and -wpc_sqrt-
data = daisy2

A lowess smoothed curve was fit to a scatterplot of the standardised residuals (derived from a model in which -hs100_ct- was entered only as a linear term) against herd size (-hs100_ct-).

The lowess curve shows the curvilinear nature of the association so adding a quadratic term for -hs100_ct- was deemed necessary.

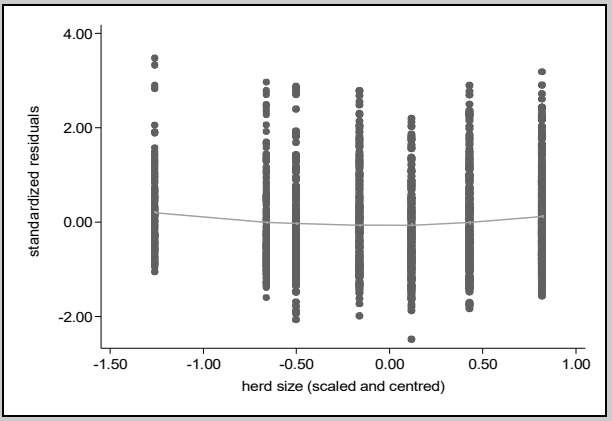

Fig. 14.7 Lowess smoothed curve through scatterplot of residuals vs herd size

We can choose the important cutpoints for the hierarchical indicators by identifying which ones are statistically significant (Section 14.4.2).

14.9.5 Correcting distribution problems using robust standard errors

A number of distributional problems can be dealt with using robust standard errors. These are discussed in more detail in Section 20.5.4 as they might also play a role in dealing with clustered data. Robust SEs are generally larger than regular SEs and hence, the resulting CIs for the coefficients are wider. If robust errors are used, be careful not to use the F-test to assess the model as it is no longer valid. Also, the MSE no longer estimates σ^2 as there is no single parametric value. After examining the residuals on a case-by-case basis, we refit the model of Example 14.16 using robust standard errors to help assess the importance of our disease variables and interactions. The variable -dyst- became non-significant but the coefficients and remaining P-values were similar to the non-robust error model (data not shown).

14.9.6 Interpreting transformed models

The assumptions of homoscedasticity and normality for the model presented in Example 14.12 were both violated. Subsequently, a Box-Cox analysis suggested that a log transformation might be appropriate (details not shown). The log transformation improved the normality of the residuals but resulted in totally unacceptable heteroscedasticity. Thus, we tried a square root transform of -wpc- (-wpc_sqrt-) and the Breusch-Pagan/Cook-Weisberg test for heteroscedasticity had a P-value of 0.52 indicating no significant departure from homoscedasticity. This solved the heteroscedasticity problem but normality remained a problem (visually the normality assumption seemed reasonable, but the formal test was still highly significant). After examining the residuals, and trying to correct distributional problems, with emphasis on homoscedasticity, we decided to use the square root transform of -wpc- as the outcome for our final model. This model is shown in Example 14.16 and is the model we will

Example 14.16 Our final model of disease impact on -wpc_sqrt-
data = daisy2

A model was built with square root transformed -wpc- as the outcome and with robust SEs.

Source	SS	df	MS	
Model	1133.93219	9	125.992465	Number of obs = 1574
Residual	12172.3347	1564	7.78282267	F(9, 1564) = 16.19
Total	13306.2668	1573	8.45916519	Prob > F = 0.0000
				R-squared = 0.0852
				Adj R-squared = 0.0800
				Root MSE = 2.7898

wpc_sqrt	Coef	SE	t	P>t	95% CI	
aut_calv	-0.514	0.142	-3.62	0.000	-0.792	-0.235
hs100_ct	1.230	0.121	10.17	0.000	0.993	1.467
hs100_ctsq	0.709	0.174	4.07	0.000	0.367	1.050
parity1	0.058	0.048	1.21	0.225	-0.036	0.152
twin	1.385	0.551	2.52	0.012	0.305	2.465
dyst	0.542	0.305	1.78	0.076	-0.057	1.141
rp	0.389	0.269	1.45	0.148	-0.138	0.917
vag_disch	-0.013	0.400	-0.03	0.974	-0.799	0.772
rp*vag_d_1	1.491	0.700	2.13	0.033	0.118	2.864
constant	7.517	0.147	51.03	0.000	7.228	7.806

The interpretation of the results of this model are presented in the text (Section 14.9.6).

use for assessment of individual observations. Since the residuals of our model still lack normality, we used robust standard errors when assessing the significance of each predictor, as suggested by Pires and Rodrigues (2007). (**Note** This means that we cannot interpret the *F*-statistic of the model.)

One problem with transformations is that they change the structure of the model and interpretation can become more difficult. Among transformations of the outcome, only the log transformation allows for back-transformation of regression coefficients (to give multiplicative effects on original scale). In general, rather than trying to explain the model in a mathematical sense, we suggest that you make extensive use of graphical techniques, compute the predicted values and plot the back-transformed outcomes. The key is to obtain the predicted outcome (and any confidence limits) in the transformed scale and then use the back-transform to determine the outcome in the original scale—on the assumption that explanations of effect are much easier in the original scale. Sometimes it is advantageous to leave the model in its transformed format. For example, it has now become standard practise to use log transformed somatic cell counts in models of risk factors that affect cell counts (or mastitis).

Because our main interest is on the effect (association) of the diseases on -wpc-, we obtained predicted values of -wpc_sqrt- only for combinations of the diseases in first parity cows in an average size herd, with no autumn-born calves or twins (this sets all these values to 0 so the arithmetic is easy).

Table 14.3 Predicted days from waiting period to conception by disease status

	dyst -		dyst +	
	vag_disch -	vag_disch +	vag_disch -	vag_disch +
rp -	59.9	56.7	60.9	63.7
rp +	66.1	95.6	67.2	101.8

Note Estimates derived from a model based on the -wpc_sqrt- (Example 14.16) are averaged over other predictors in the model.

Note In our interpretation that follows, for simplicity, we ignore the SEs of these predicted times. If vaginal discharge is not present, -rp- delays conception by approx. 6 days. If -rp- is not present, -vag_disch- has very little effect ($\pm$3 days). If both -rp- and -vag_disch- are present, conception is delayed by approximately 35-41 days depending on whether -dyst- was absent or present. Although there was no interaction between -dyst- and the other 2 diseases, the effect of -dyst- does vary across -rp- and -vag_disch- categories (varying from 1 to 7 days) as a result of modelling on a transformed scale. For a more detailed discussion of back transformations see Afifi *et al* (2007).

When applying transformations to multivariable models we need to be careful when making predictions because additive and linear models in one scale become (possibly strongly) non-linear and non-additive (*ie* showing interaction) in another scale. Thus, the outcome depends on the values of all of the variables in the model even though there is no actual interaction. A recommended practise here is to use the mean values for variables not of direct interest and a range of values for those variables of primary interest when computing the predicted values. Again, all confidence limits *etc* are determined in the transformed scale and then back-transformed into the original scale as necessary.

14.9.7 Specification bias

If the model is correct, the residuals are uncorrelated with the predicted outcome ($\hat{Y}$). However, if an important variable is missing from the equation, the model suffers from specification bias. This might reflect itself in a linear pattern of the standardised residuals with the predicted values of Y. For example, small (negative) residuals might be associated with lower values of $\hat{Y}$ and large (positive) residuals with large values of $\hat{Y}$ suggesting that one or more important predictor variables are missing. Specifically, the sampling units with positive residuals have something in common that also gives them large observed values of Y, and this feature might help identify the missing variable. Unfortunately, with a 'weak' (low R^2 as in our model) model, it is difficult to discern some of these patterns because of the relatively large variability in r_i. There are formal tests for specification bias, but they are beyond the scope of this text.

14.10 ASSESSMENT OF INDIVIDUAL OBSERVATIONS

Our previous efforts were directed toward evaluating the major assumptions on which linear regression models are based. Here we assess the fit of the model on an observation by observation basis. Specifically, we look for:

- cases that are not well fit by the model and hence, have large residuals; some of these might be deemed **outliers**. In a technical sense, outliers have large values of -rstu- that are very unlikely to have arisen due to chance.

- cases with unusual *X*-values; these are called **leverage** observations.
- cases that have an unduly large impact on the model; these are **influential** observations.

Our rationale for pursuing this observation-by-observation analysis is that we want to be sure the model is correct for the majority of the study subjects, and if we can identify specific instances of observations that do not fit, or have a big influence on our model, it can help us identify the reason(s) for that impact. In addition, this pursuit can often provide insight into features of the data that can be useful in clarifying the model results or in planning studies.

There are 2 general approaches to assist in this task, one is to use graphical techniques to detect observations with an unusual value (*ie* atypical relative to the others) on the test statistic, and the other is based on identifying observations that exceed a specific cutpoint. Both have their advantages. The key is to try a variety of approaches and see which you prefer, but there is no need to use all possible approaches in a given dataset. Although we use graphical techniques regularly, here we present only tabular results. If a predictor variable is interval censored (*ie* treated as a continuous variable but only takes selected values (such has -hs100- in our examples)) special methods, beyond the level of this text, are available for the assessment of the residuals (Topp and Gomez, 2004).

14.10.1 Outliers

In general, an outlier is an observation in a dataset which is far removed in value from the others in the dataset. In multivariable datasets, we need to make precise the meaning of 'far removed in value', because it may be only in the combination of several variables that an observation becomes unusual (*eg* having twins and the other diseases of interest). In regression analysis, we distinguish between outliers in the outcome variable and outliers among the predictor variables (not involving the outcome).

An outlier in the outcome is detected by a (numerically) large residual, where 'large' is viewed relative to the other observations and to what would be expected for a dataset of the same size.

It is important to note that, although we are interested in identifying outliers, we do so largely to try and explain/understand why they fit poorly, not to remove them without reason. Outliers inflate the standard error of the estimate and hence, reduce the power of statistical tests. Unusual values of the outcome, or predictors, might reflect the state of nature, they might arise because of transcription or data entry errors, or they might signal that we are missing important covariates that could 'explain' the poor fitting points. In most instances, one should not be unduly concerned about these data points unless their standardised value is greater than 3, although values between 2 and 3 might be having an impact on the model. Recall that with normally distributed residuals, a small percentage (0.3%) of standardised residuals would be expected to lie outside of ± 3.

If an observation is suspected to be an outlier, it can be assessed with a 2-tailed *t*-test based on the studentised (stu) residual. However, the probability associated with this test depends on whether the observation was suspected of being an outlier *a priori* or not. If an observation was suspected beforehand, then the P-value is found by comparing the studentised residual to the value of a *t*-distribution with dfE-1 degrees of freedom. However, if we are testing a specific data point subsequent to observing the residuals, we should multiply the above probability by the number of observations (n) which is equivalent to using the Bonferroni adjustment (Eq 14.18). If the studentised residual is larger than this number then it can be considered to be a

statistically confirmed outlier. In this dataset, a studentised residual greater than 4.17 would be considered to be an outlier.

$$P = 2 * n * t(dfE\text{-}1, r_{ti}) \qquad \textit{Eq 14.18}$$

Some general rules in managing outlier observations include:

- identify observations with large studentised residuals;
- try and find an explanation for them, such as a recording error or erroneous test result (ie equipment or operator problem);
- if there is no recording error, then think about what factors the outliers might have in common that, if measured, could explain their lack of fit;
- try refitting the model without the outliers to see the effect on the model; and
- if the observations are to be deleted (which they rarely are), be sure to explicitly record this for yourself and those who read your research report. (It is hard to justify the deletion of observations.)

Although deleting outliers will improve the fit of the model to the sample data, it might actually decrease the model's validity as a predictor for future observations. In Example 14.17, we have presented the 5 largest positive and negative residuals from our model along with the values of the key predictor variables; this presentation often helps you understand the reason for the departures from expectation.

14.10.2 Detecting 'unusual' observations—leverage

This activity focuses on identifying subjects with unusual values in the *X*s and is particularly applicable when many continuous variables are present in the model. For this purpose, we use the leverage from Eq 14.15 which indicates the potential for the i^{th} observation to have a major impact on the model.

In general, observations with at least one of the predictors that is far from the mean will tend to have a large leverage; note that leverage lies between $1/n < h_i < 1$. Observations with a very high leverage may have a large influence on the regression model; whether they do or not depends on the observed *Y*-values. A common rule is to examine observations that have leverage values $>2(k+1)/n$, where k is the number of predictors in the model (or the number of regression parameters, excluding the intercept). There is a fair bit of arbitrariness in this cutpoint (another commonly used value is $3(k+1)/n$), and hence, one should initially look for observations with relatively extreme leverage values regardless of the cutpoints. Using the latter guideline, for our example, any observation with a leverage above 0.019 can be considered as extreme in its predictor values. The 5 cases with the largest leverage values are shown in Example 14.18. Having identified potentially influential observations, we then proceed to evaluate their actual influence on the model.

14.10.3 Detecting influential observations—Cook's distance and DFITS

An intuitive test of an observation's overall influence is to omit it from the model, recalculate the model and note the amount of change in the predicted outcome. If an observation is influential, the change will be large; if not, the change will be small (see Example 14.19). This approach forms the basis of **Cook's Distance** D_i which is the sum of squared differences in fitted values with and without observation i (summed over all other observations and scaled

Example 14.17 Examination of standardised and studentised residuals from model with -wpc_sqrt- as outcome
data = daisy2

Standardised and studentised residuals were computed based on the model presented in Example 14.16 with ordinary SEs. Given the relatively large size of the dataset, the differences between the 2 sets of residuals were minimal. The 5 smallest standardised residuals were:

cow	herd size	parity	twin	dyst	rp	vaginal discharge	wpc sqrt	pred. value	std. resid.
2272	263	2	no	no	no	no	1.00	7.73	-2.42
1032	201	2	no	no	no	no	1.00	6.62	-2.02
403	235	4	no	no	no	no	1.73	7.00	-1.89
983	201	4	no	no	no	no	2.00	7.25	-1.89
1130	201	5	no	no	no	no	2.24	7.31	-1.82

These generally are cows that were predicted to have a longer than average -wpc- interval, but instead had very short intervals.

The 5 largest standardised residuals were:

cow	herd size	parity	twin	dyst	rp	vaginal discharge	wpc sqrt	pred. value	std. resid.
199	294	2	no	no	no	no	15.91	7.72	2.94
4939	185	4	no	yes	no	no	15.49	7.22	2.99
805	333	4	no	no	no	no	17.26	8.66	3.09
1226	125	2	no	no	no	no	15.30	6.64	3.12
1257	125	5	no	no	no	no	15.91	6.81	3.27

These are cows that had -wpc- intervals much longer than predicted. (None of the 3 cows with the largest standardised residuals had any of the 3 diseases of interest, but they had long -wpc-). Based on our data, with 1,574 cases and 1,564 degrees of freedom, a case with a studentised residual more extreme than $\pm$4.17 would be 'unusual' with a P-value of <0.05. Since we have no cows with such an extreme residual we conclude there are no serious outliers.

suitably). A more direct interpretation of Cook's distance derives from the formula:

$$D_i = \frac{r_{si}^2}{(k+1)} * \frac{h_i}{(1-h_i)} \qquad \textit{Eq 14.19}$$

emphasising that a large standardised residual, a large leverage, or both can lead to undue influence.

A commonly suggested cutpoint is to compare the Cook's value with the $F(k+1, n-k-1)$ distribution. If it exceeds the 50% percentile (not 5%), which is essentially 1, then the observation should be investigated. However, in our practical experience, the values of D_i rarely exceed this cutpoint, so it is recommended to look instead for values that are extreme relative to the others in the data. In our data set if we use *4/n* as the cutpoint, a large Cook's D_i would have a value exceeding 0.0025 and 91 cows have this value or greater.

Example 14.18 Examination of leverage cases from model with -wpc_sqrt- as outcome
data = daisy2

Leverage values were computed based on the model presented in Example 14.16 with ordinary SEs. The 5 largest values were:

cow	herd size	parity	twin	dyst	rp	vaginal discharge	std. resid.	leverage
2389	263	5	yes	no	no	yes	-0.57	0.059
2433	263	3	yes	no	yes	yes	-0.52	0.063
163	294	2	yes	no	yes	yes	0.06	0.064
1122	201	4	yes	no	yes	yes	-1.24	0.064
4916	185	4	yes	no	yes	yes	-0.66	0.064

Cows with twins (but without dystocia) and with one or more of -rp- or -vag_disch- have high leverage. Because delivering twins without dystocia is unusual and the diseases are infrequent, these cases are essentially 'unusual' and are highlighted by the leverage statistic.

A similar approach is used with a statistic known as **DFITS** (or DFFITS) (Example 14.19). It is an acronym that stands for 'difference in fit' between when the observation is in the model versus when it is out. DFITS indicates the number of standard errors change to the model when that observation is deleted. The following formula for DFITS shows its strong similarity to Cook's distance:

$$\text{DFITS}_i = r_{ti}\sqrt{\frac{h_i}{(1-h_i)}} \qquad \textit{Eq 14.20}$$

Thus, the DFITS statistic is based on the studentised residual and retains its sign. Again, if the DFITS numerically exceeds a value of, for example, 1 for n<120 or $2\sqrt{((k+1)/n)}$ in a larger dataset, it means that if that observation was deleted, the model would change by a relatively large amount (recall that k is the number of predictor variables in the model). As with outliers, we should be hesitant to remove influential observations without good reason. In general, we do not remove influential observations unless the data are known to be incorrect, or there is a clear explanation for their influence. If observations are removed, the reason(s) for their removal, must be drawn to the attention of those reading your research results.

In our model, a large value for DFITS is 0.16. There are a number (68) of cases with larger values than these, almost all have one or more of the 3 diseases of interest. While it is true that these diseases increase the time to conception, the cases shown in Example 14.19 fit the model reasonably well (*ie* reasonable -rsta- values) and there seems little reason to remove them.

14.10.4 Detecting influential values of specific predictors

Given an exposure variable of interest, one can assess the impact of deleting a specific observation on the value of the regression coefficient for that variable. The statistic used for this is known as a delta-beta (DB) and reflects the number of standard errors by which the specific regression coefficient changes when that observation is deleted. Thus it helps identify if a particular observation has a large influence on the β for that variable. Critical values for n<120

Example 14.19 Examination of influential cases from model with -wpc_sqrt- as outcome
data = daisy2

The 5 largest negative DFIT cases are:

cow	herd size	twin	dyst	rp	vaginal discharge	Std. resid.	Cook's D	dfit
713	333	yes	no	no	no	-1.63	0.012	-0.346
444	235	yes	no	no	no	-1.68	0.011	-0.339
1122	201	yes	no	yes	yes	-1.24	0.010	-0.323
2480	263	no	no	yes	yes	-1.68	0.010	-0.320
5029	185	yes	yes	no	no	-1.34	0.009	-0.305

Cows with twins or one of the diseases and hence predicted to have long -wpc- intervals, but with short -wpc- intervals had an influence on the model. The 5 largest positive DFIT cases are:

cow	herd size	twin	dyst	rp	vaginal discharge	std. resid.	Cook's D	dfit
4939	185	no	yes	no	no	2.99	0.012	0.349
1124	201	no	yes	no	no	2.85	0.012	0.351
238	294	no	no	no	yes	2.69	0.016	0.398
4999	185	no	no	yes	yes	2.09	0.016	0.405
1238	125	no	no	no	yes	2.63	0.020	0.447

Cows without twins but with one or more of the diseases were the most influential cases. In all instances, had the 'influential' cows possessed covariate values that were of little interest to us, we might remove them and assess their impact as part of considering their complete removal from the model. Since we are interested in the effect of the diseases, we will not remove any cases. Not surprisingly, among the 10 most influential cases, those with high leverage had smaller residuals than those with lower leverage. Other factors that appeared to lead to influential cases included 1 herd that accounted for 40% of the 10 most influential cases, and cows with twins (1.7% of calvings) accounted for 4 of the 5 largest negative DFIT cases.

Due to the influence of cows with twins, we reran the model excluding cows with twins. None of the coefficients changed by more than 20%; however, the coefficient for dystocia became non-significant with P=0.076.

are 1 and for larger datasets $2/\sqrt{n}$. Again, this value might be too sensitive and initially one should just focus on observations with very extreme DB values.

In our model, the critical DB values for the variable -rp- are ±0.05. All cows with -rp- were influential; however, this is not surprising given that cows with -rp- are definitely in the minority. The same is true for DBs on -vag_disch- and -dyst-. There were 30 cows with both -rp- and -vag_disch- and they had large DBs also. There is no need to remove any of these cases in our dataset. In general, the DB statistics are much more useful if the variables of interest are continuous rather than dichotomous.

14.10.5 Comments on the model deficiencies

In our examples, we have taken you through the basic steps of assessing a linear regression model. We did have a few problem cows (subjects) in the dataset that were minimally influential, or poor fitting, but we had a serious problem relating to the lack of normality with more positive residuals than negative residuals. The reality is that if we correct the normality assumption, we create unequal variances. On balance our square root transformation corrected heteroscedasticity so we stayed with that model. We might be bothered by the lack of normality and hence we ran a model using robust variances just to check the significance of coefficients (we can't use the *F*-statistic in this instance). As indicated in Example 14.16, the variables selected for the model, with -wpc_sqrt- as the outcome, do not change drastically when robust standard errors are used. This provides some support for the likely validity of the former model despite the non-normality of residuals and the low R^2.

14.11 TIME-SERIES DATA

Time-series data are characterised by an outcome measured at equal time intervals over a reasonably long time period, such as hospital admissions per day for 1-5 years, or herd-level milk production per day for at least a year. In this setting, the outcome in one time period (*eg* day) is likely to be correlated with the outcome in adjacent time periods. This correlation of outcomes often leads to correlation of residuals and breaks one of the major assumptions of (ordinary least squares; OLS) linear regression. Often, we can predict that data will be correlated given the structure of our sampling of subjects (for example measuring the same outcome on selected animals within a herd, or taking repeated measurements of an outcome on the same individuals over time). Pires and Rodrigues (2007) describe methods for use when only some of the errors are correlated, such as would occur if a dataset had data from some cows with multiple lactations, when most data came from cows with only one lactation. Analyses to control for the correlations between repeated measurements on a group of study subjects are described in Chapter 23.

In time-series data, a correlation between residuals on adjacent time periods (*eg* days) arises because we make repeated observations, at equally spaced intervals, on our study subjects regularly (for example taking daily measurements of milk production on a cow, or on a herd. The set of predictors could be disease occurrence and/or daily somatic cell counts (SCC)). If we analyse such data, to estimate the impact of SCC on milk production, the coefficients reflecting the 'effect' of the predictors are unbiased but the standard errors are likely to be incorrect. The correlation of residuals can lead to either inflated or deflated standard errors. If we suspect serial correlation, we can use the Durbin-Watson test to assess this. In general, a Durbin-Watson test value of approximately 2 indicates no correlation, and as the test statistic gets smaller this indicates increasing correlation between adjacent residuals. There are more advanced tests of serial correlation such as the Ljung-Box *Q*-test (Ljung & Box, 1978), that provide a specific P-value that is easier to interpret than the Durbin-Watson test.

Examples of time series analysis include the analysis of temporal patterns of *Campylobacter* spp in humans and poultry (Hartnack *et al*, 2009); relationships between ambient temperature and enteric infections in humans (Fleury *et al*, 2006), and temporal patterns of fox rabies in Ontario (Tinline *et al*, 2004). Poirier *et al* (2008) give an example of modelling the effect of an intervention in poultry production on the future monthly number of isolations of poultry-related *Salmonella* spp in humans. A useful text on time-series is Diggle (1990).

One of the early steps in analysing time-series data is to plot the outcome data and, in this regard, a smoothed curve is a good way to enhance visualisation of trends and other patterns such as seasonal changes in the outcome. If the time counter is '*t*' (*eg* for daily measurements *t*=day) we can use a variety of smoothing functions of length $2m+1$ ($m>=t$). The larger *m* is, the greater the smoothing (of all fluctuations of duration less than *m*). For example, if we have a daily time series and *m*=1, then a 3-day moving average will remove variation in the outcome measure of periodicity of 3 days or less. Before proceeding to detailed analyses it is important that the data be 'stationary'; that is, any trend, or seasonal variation be removed (beyond the level of this text).

Once this is accomplished, we should examine the correlation between residuals over a specified number of lag periods (for example, correlations between residuals over a 7-day lag (*m*=7) indicates the correlations between observations ranging from 1 to 7 days apart). Typically, the correlation is greatest for time points that are closest together (*ie* observations on the same subjects made close together in time tend to be strongly correlated with one another). For example, the residuals from predicting milk production of dairy cows are most strongly correlated between adjacent days and the correlation tends to decrease as the number of days between milk production measurements increases up to about 4 days when the correlation becomes essentially non-existent). The autocorrelation function can be used to ascertain the correlation structure for outcomes in periods up to *m* time units apart. The partial autocorrelation function between 2 outcomes *m* units apart takes into account the correlations between time units between 1 and *m* and is useful for identifying where sudden changes in the correlation structure occur. Most software packages have convenient commands to allow you to examine these correlations over a variety of lag periods. Knowledge of these correlations provides guidance about the desired model structure.

14.11.1 Adjusting for serial correlation

One way to correct for the correlation between residuals is to use a weighted least squares estimator, and 2 such estimators are the Cochrane-Orcutt and the Prais-Winsten estimators. These do not take the dynamic nature of the time series (*eg* trends, weekly or seasonal patterns) into account but they do make corrections to the standard errors, assuming a lag of 1 time unit suffices. Again, many software packages will allow you to run these regressions, and it is usual practise to rerun the tests for correlated residuals after running these models to ensure that the correlations have been removed.

A more advanced approach involves the use of what are termed autoregressive models (Zeger *et al*, 2006). The details of these are beyond the level of this text and will not be pursued here. However, we will (barely) introduce the subject at this point. Essentially, we model the outcome (Y_t) on a given day as a function of a number of predictors (*ie* the *Xs*). The *Xs* can be variables that we have reason to believe will account for some of the patterns evident in the time series such as seasonal and/or annual trends, or they could be time-dependent exposure variables ($X_t s$) whose 'effect' we are attempting to estimate. The choice of these would depend on our beliefs about what processes are 'driving' the patterns seen in the time series. Once these fixed effects are included it is common to find that the nature and strength of the lagged correlations have changed from the original naive values. The equation below is an autoregressive (AR) model because we have included the outcome on the previous 2 days as predictors; this would be an AR-2 model.

$$Y_t = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + \alpha_1 Y_{t-1} + \alpha_2 Y_{t-2} + \varepsilon_t \qquad \textit{Eq 14.21}$$

This model also implies that the predictor variables are time dependent (and measured on the same time scale as Y) variables. For example Y_t could be milk produced on day t and X_{1t} could be somatic cell count on day t, and X_{2t} could be a measure of daily temperature on day t. As shown in Fleury *et al* (2006), Y_t could be the daily number of enteric disease cases and the X_ts could be weather variables.

With an AR1 structure, correlations have an exponential decay over time (the structure of the decay is more complex for AR2, AR3 *etc*). It is useful to verify this, for example, visually through correlograms, in order to ensure that the expected decay is consistent with the data. If there are sudden changes in the correlation structure, or if the correlation of Y_t with Y_{t-1} drops very quickly, then a moving average model could be helpful to account for these. The moving average (MA) component uses the residuals of time periods at the specified lags to account for the correlation structure. As the name suggests, ARMA models use both autoregressive and moving average processes. An ARMA(1,1) model of AR-1 and MA-1 is useful if the AR-1 model includes measurement error. As noted above, for their validity, ARMA models must be stationary (this indicates that the mean, variance, and autocorrelation structures are the same over time) and this needs to be verified. Stationarity does not mean that we cannot model events that change over time, but we may need to adjust for them by removing trend, seasonality *etc*.

REFERENCES

Abu-Zidan FM, Rao S. Factors affecting the severity of horse-related injuries Injury. 2003; 34: 897-900.

Afifi AA, Kotlerman JB, Ettner SL, Cowan M. Methods for improving regression analysis for skewed continuous or counted responses. Annu Rev Public Health. 2007;28:95-111.

Austin PC, Hoch JS. Estimating linear regression models in the presence of a censored independent variable Stat Med. 2004; 23: 411-29.

Chao YC, Zhao Y, Kupper LL, Nylander-French LA. Quantifying the relative importance of predictors in multiple linear regression analyses for public health studies J Occup Environ Hyg. 2008; 5: 519-29.

Cheung YB. A modified least-squares regression approach to the estimation of risk difference Am J Epidemiol. 2007; 166: 1337-44.

Cook R, Weisberg S. Diagnostic for heteroscedasticity in regression Biometrika. 1983; 70: 1-10.

Fleury M, Charron DF, Holt JD, Allen OB, Maarouf AR. A time series analysis of the relationship of ambient temperature and common bacterial enteric infections in two Canadian provinces. Int J Biometeorol. 2006;50(6):385-91.

Fuller W. Measurement Error Models. New York: Wiley; 2006.

Hartnack S, Doherr MG, Alter T, Toutounian-Mashad K, Greiner M. Campylobacter monitoring in German broiler flocks: an explorative time series analysis Zoonoses Public

Health. 2009; 56: 117-28.

Livingstone DJ, Salt DW. Judging the significance of multiple linear regression models J Med Chem. 2005; 48: 661-3.

Ljung G, Box G. On a measure of lack of fit in time series models Biometrika. 1978; 65: 297-303.

Marill KA. Advanced statistics: linear regression, part I: simple linear regression Acad Emerg Med. 2004a; 11: 87-93.

Marill KA. Advanced statistics: linear regression, part II: multiple linear regression Acad Emerg Med. 2004b; 11: 94-102.

Murad H, Freedman LS. Estimating and testing interactions in linear regression models when explanatory variables are subject to classical measurement error Stat Med. 2007; 26: 4293-310.

Pires AM, Rodrigues IM. Multiple linear regression with some correlated errors: classical and robust methods Stat Med. 2007; 26: 2901-18.

Pitard A, Viel JF. Some methods to address collinearity among pollutants in epidemiological time series. Stat Med. 1997;16(5):527-44.

Poirier E, Watier L, Espie E, Weill FX, De Valk H, Desenclos JC. Evaluation of the impact on human salmonellosis of control measures targeted to *Salmonella Enteritidis* and *Typhimurium* in poultry breeding using time-series analysis and intervention models in France Epidemiol Infect. 2008; 136: 1217-24.

Topp R, Gomez G. Residual analysis in linear regression models with an interval-censored covariate. Stat Med. 2004;23(21):3377-91.

Walter SD, Feinstein AR, Wells CK. Coding ordinal independent variables in multiple regression analyses Am J Epidemiol. 1987; 125: 319-23.

Zeger SL, Irizarry R, Peng RD. On time series analysis of public health and biomedical data. Annu Rev Public Health. 2006;27:57-79.