

---

## 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 technical and causal perspectives.
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 that have data which require a time-series approach to analysis.

## 14.1 INTRODUCTION

Up to this point, most of the 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 that is suitable for modelling the outcome when it is measured on a continuous or near-continuous scale. Examples of these would include birth weight, blood pressure, body mass index and, in some circumstances, disease frequency at a regional (*eg* county) 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 and Rao (2003), in which multiple regression was used to identify factors related to the severity of injury in falls from horses (based on a continuously distributed injury severity score). Two journal articles which provide a readable introduction to linear regression are Marill (2004a; 2004b).

In regression analysis, the relationship between the outcome and the predictors is asymmetric—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 continuous, categorical, or dichotomous scales.

## 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

$$Y = \beta_0 + \beta_1 X_1 + \varepsilon \quad \text{Eq 14.1}$$

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  $\beta_0$ , when  $X_1$  has the value 0, and for each unit increase in  $X_1$  the outcome  $Y$  changes by  $\beta_1$  units.  $\beta_0$  is usually referred to as the **constant**, or the **intercept term**, whereas  $\beta_1$  is usually referred to as the **regression coefficient**. The  $\varepsilon$  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  $\beta$ s represent population parameters which we estimate based on the observed data and our model. We will refer to predictor variables as  $X$ s. In general, we will denote the number of observations as  $n$ . Thus, our predicted values are:

$$\hat{Y}_i = \beta_0 + \beta_1 X_{1i} \quad , \quad i = 1, \dots, n \quad \text{Eq 14.2}$$

where  $\hat{Y}_i$  is the predicted value of the outcome for the  $i^{\text{th}}$  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  $i$ ; however, 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 to 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 1 predictor.) With 2 predictor variables, the regression model could be written as:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon \quad \text{Eq 14.3}$$

which suggests that we can predict the value of the outcome  $Y$  knowing the baseline (intercept or constant)  $\beta_0$  and the values of the 2 predictor variables (*ie* the  $X$ s). The parameters  $\beta_1$  and  $\beta_2$  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,  $\beta_1$  is an estimate of the effect of  $X_1$  on  $Y$  after controlling for the effects of  $X_2$ , and  $\beta_2$  is the estimated effect of  $X_2$  on  $Y$  after controlling for the effects of  $X_1$ . Expressed another way,  $\beta_1$  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 ( $\varepsilon$ ) takes this into account. Thus, our prediction equation is:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \quad \text{Eq 14.4}$$

where  $Y$  is the predicted value of the outcome for specific values of the 2 predictors  $X_1$  and  $X_2$ . In this equation,  $\beta_1$  describes the number of units change in  $Y$  as  $X_1$  changes by one unit,  $X_2$  being held constant, whereas  $\beta_2$  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  $\beta$ s 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 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  $\beta$ s, while not

including variables of little importance in the equation as this will increase the number of  $\beta$ s 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 birth weights of babies in the USA in 2007. The names of the variables used in this chapter and their descriptions are shown in Table 14.1; further details are in Chapter 31.

**Table 14.1 Selected variables from the dataset bwt5k**

Variable	Scale of measurement	Description
obs	N/A	observation number
bwt	continuous	birth weight (gm)
mrace_c4	nominal	mother's race (4 categories) also as 3 categories (mrace_c3)
white	dichotomous	mrace_c4 recoded to white vs other
meduc_c4	ordinal	mother's education
college	dichotomous	meduc_c4 recoded to college vs <college
tbo	ordinal	total birth order
tbo_c2	dichotomous	tbo recoded to primiparous vs multiparous
multbrth	dichotomous	single birth vs multiple babies (twins, etc)
wtgain	continuous	maternal weight gain (lb)
wtgain_c2		wtgain recoded to low (<30 lb) and high ( $\geq$ 30 lb)
cig_1, cig_2, cig_3	continuous	cigarettes 1 <sup>st</sup> , 2 <sup>nd</sup> , and 3 <sup>rd</sup> trimesters
gest	continuous	gestation length (weeks)

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 -bwt- were: median=3328 gm and mean (average)=3295 gm, the standard deviation was 566 gm and the range was 480 gm to 5,550 gm.

Without more 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. 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 an individual (baby) 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	Sums of squares	Degrees of freedom	Mean squares	F-test
Model (or regression)	$SSM = \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2$	$dfM = k$	$MSM = SSM/dfM$	$MSM/MSE$
Error (or residual)	$SSE = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$	$dfE = n - (k+1)$	$MSE = SSE/dfE$	
Total	$SST = \sum_{i=1}^n (Y_i - \bar{Y})^2$	$dfT = n - 1$	$MST = SST/dfT$	

In the formulae in the table,  $\bar{Y}$  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 sums of squares (SST) is shown in the second column of Table 14.2 (*ie*  $SST = SSM + SSE$ ; also,  $dfT = dfM + dfE$ ). For our example, gestation length will be the  $X$ -variable of interest. The MSE is our estimate of the error variance and therefore, also denoted as  $\sigma^2$ . Furthermore,  $\sigma$ , the square root of  $\sigma^2$ , 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  $\beta$ s 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  $\beta$ s can be determined using:

$$\beta_0 = \bar{Y} - \beta_1 \bar{X}_1 \quad \text{and} \quad \beta_1 = \sum (X_{1i} - \bar{X}_1)(Y_i - \bar{Y}) / SSX_1 \quad (\text{with } SSX_1 = \sum (X_{1i} - \bar{X}_1)^2) \quad \text{Eq 14.5}$$

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 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  $H_0: \beta_1 = \beta_2 = \dots = \beta_k = 0$  (*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  $\beta$ s 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 (1,790) is highly significant ( $p < 0.001$ ), indicating that the  $X$ -variable(s) in the model (-gest- in this instance) explains some of the variation in -bwt-. One feature of the ANOVA table that 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 as more predictors are added to the model.

Some care is necessary when interpreting the model  $F$ -statistic, as its meaning changes with the method of model-building (Livingston and Salt, 2005). The  $F$ -test probably has a straightforward meaning only 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

**Example 14.1 A simple linear regression model of birth weight on gestation length**

data = bw5k

A linear regression model with -bwt- as the outcome and -gest- as the sole predictor was fit using the birth-weight 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	422130383	1	422130383
Residual	1.1786e+09	4998	235814.6
Total	1.6007e+09	4999	320210.4

Number of obs = 5000  
F(1,4998) = 1790.09  
Prob > F = 0.0000  
R-squared = 0.2637  
Adj R-squared = 0.2636  
Root MSE = 485.61

Note that the variance (MS) of -bwt- is 320,210.4 and this is quite a bit larger than the MS residual suggesting that gestation length does explain some of the variation in -bwt-. The root MSE has the same scale as -bwt- (*ie* grams) and, because -gest- is associated with -bwt-, it is smaller than the standard deviation (566 gm) of -bwt-.

The regression coefficients from the model are shown below.

bwt	Coef	SE	t	P>t	95% CI
gest	124.487	2.942	42.31	0.000	118.719    130.255
constant	-1513.854	113.868	-13.29	0.000	-1737.085    -1290.623

The coefficient for -gest- suggests that, for each additional week in gestation length, -bwt- increases by 124.5 gm. Given the SE (2.94) of this statistic, the *t*-statistic (42.3) is significant at <0.001 so we can assume, at this point, that -gest- has an association with (or effect on) -bwt-. This is consistent with the 95% confidence interval values which do not include 0 (the no-effect level). The CI suggests that a reasonable range for the effect of a 1-week change in gestation length is between 119 gm and 130 gm.

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

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 (*ie* the *df*<sub>E</sub>) is used to evaluate the significance of any of the regression coefficients (*eg* the *j*<sup>th</sup> coefficient). The usual null hypothesis is *H*<sub>0</sub>: β<sub>*j*</sub>=0 but any other value of β\* can be used in *H*<sub>0</sub>: β<sub>*j*</sub>=β\* depending on the context. The *t*-test formula is:

$$t = \frac{\beta_j - \beta^*}{SE(\beta_j)}$$

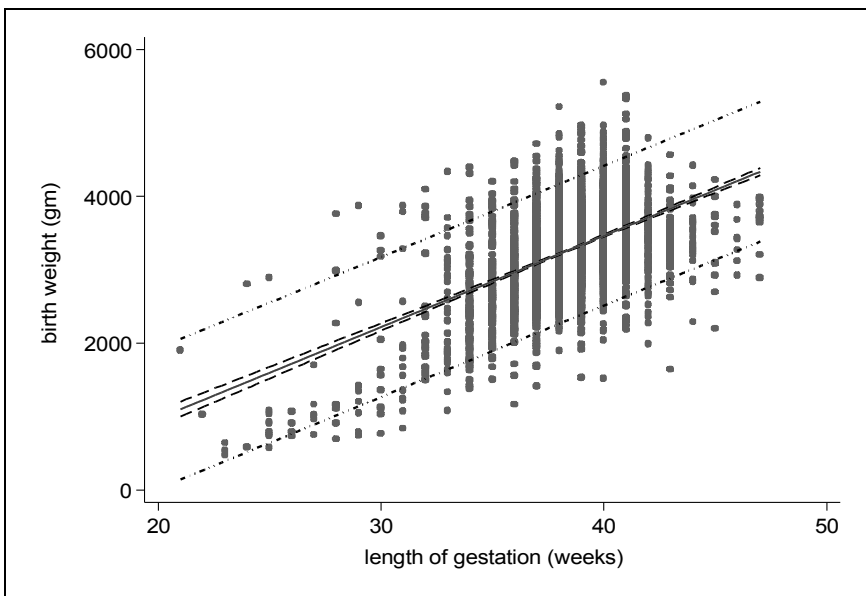
*Eq 14.6*

where  $SE(\beta_j)$  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:

$$SE(\beta_1) = \sqrt{MSE/SSX_1} \quad \text{Eq 14.7}$$

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 42.3 has a P-value of  $<0.001$  so we would reject the null hypothesis that the true regression coefficient has the value  $\beta_1=0$  which would indicate no association of -gest- with -bwt-. Fig. 14.1 shows the trend of increasing birth weight with increasing gestation length.

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 program 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 observed level of significance is higher than the nominal level of significance (usually termed  $\alpha$ ) 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.



**Fig. 14.1** Prediction (confidence) intervals for mean (dash line) and new observation (dash-dot line)

### 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  $SE_{\text{mean}}$ ; sometimes called the prediction error) of

$$SE_{\text{mean}}(Y|x^*) = \sigma \sqrt{\frac{1}{n} + \frac{(x^* - \bar{X}_1)^2}{SSX_1}} \quad \text{Eq 14.8}$$

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 gestation length of 40 weeks, the predicted outcome is 3,466 gm with a prediction SE of 7.96 gm.

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

$$SE_{\text{obs}}(Y|x^*) = \sigma \sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{X}_1)^2}{SSX_1}} \quad \text{Eq 14.9}$$

Using the data in Example 14.1, for a specific 40-week gestation, the predicted outcome is 3,466 gm with a forecast SE of 486 gm. Two points can be made here. First, the variation associated with predicting the mean outcome is much less and prediction intervals much narrower 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 intervals for the predictions are found using:

$$95\% \text{ CI} = Y \pm t_{.05}(\text{SE}) \quad \text{Eq 14.10}$$

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

The association between birth weight and gestation length, as determined by a linear regression of -bwt- on -gest-, 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 26.4%). Given that more than 73% of the variation in -bwt- is unexplained, this suggests that we cannot predict birth weight very precisely if we only know the gestation length. Perhaps additional variables can add to the explained proportion (a



rationale for a multivariable model). One formula for  $R^2$  is  $R^2 = \text{SSM} / \text{SST} = 1 - (\text{SSE} / \text{SST})$ . 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  $R^2 = 1 - (\text{MSE} / \text{MST})$ .

In multivariable models, the adjusted  $R^2$  is slightly lower than the  $R^2$ . The adjusted  $R^2$  is useful 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 has 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 appropriate 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 -multbrth-.

### 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 physical activity or nutrition) 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, one might use the model sum of squares, as indicated below.) That is, note  $\text{SSE}_{\text{full}}$  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  $\text{SSE}_{\text{red}}$  (for the ‘reduced model’). If variables  $X_j$  and  $X_k$  are important, then  $\text{SSE}_{\text{full}} \ll \text{SSE}_{\text{red}}$  (and  $\text{SSM}_{\text{full}} \gg \text{SSM}_{\text{red}}$ ).

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  $\text{dfe}_{\text{red}} - \text{dfe}_{\text{full}}$ ) to give us the MS from these variables. Dividing this MS by the  $\text{MSE}_{\text{full}}$  provides an  $F$ -test of the significance of  $X_j$  and  $X_k$  conditional on the other variables in the model. The formula to assess a set of variables is:

**Example 14.2 A simple regression with a dichotomous predictor**  
data = bw5k

A simple linear regression model of birth weight (-bwt-) whether or not the birth was multiple (twins or triplets) (-multbrth-) as the only predictor was fit.

				Number of obs = 5000
				F(1,4998) = 543.52
				Prob > F = 0.0000
				R-squared = 0.0981
				Adj R-squared = 0.0979
				Root MSE = 537.46

Source	SS	df	MS
Model	157000944	1	157000944
Residual	1.4437e+09	4998	288861.7
Total	1.6007e+09	4999	320210.4

Note that in this model, the  $X$ -variable(s) in the model is deemed to be significantly ( $P<0.001$ ) associated with -bwt-, and by itself, it explains a reasonable amount (9.8%) of the variation in -bwt-.

bwt	Coef	SE	t	P>t	95% CI
multbrth	-943.675	40.478	-23.31	0.000	-1023.029 -864.321
constant	3329.614	7.744	429.97	0.000	3314.433 3344.796

The regression coefficient for -multbrth- is -943.7, indicating that as -multbrth- increases by 1 unit the -bwt- decreases by 943.7 gm (recall that the coding for -multbrth- is 0 if it was a single birth and 1 if the birth was of twins or triplets; so an increase of 1 unit is the difference in outcome between babies that were part of a multiple birth vs those that were not). The P-value indicates that the apparent effect of 943.7 gm is significantly different from 0, so ‘chance’ is not a likely explanation for the association. The substantial  $R^2$  suggests that this is going to be an important predictor of birth weight.

$$F_{\text{group}} = \frac{\left( \frac{\text{SSE}_{\text{red}} - \text{SSE}_{\text{full}}}{\text{df } E_{\text{red}} - \text{df } E_{\text{full}}} \right)}{\text{MSE}_{\text{full}}} \sim F((\text{df } E_{\text{red}} - \text{df } E_{\text{full}}), \text{df } E_{\text{full}}) \text{ under } H_0$$

Eq 14.11

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  $\text{SSM}_{\text{full}} - \text{SSM}_{\text{red}} = \text{SSE}_{\text{red}} - \text{SSE}_{\text{full}}$ , 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 the 3 cigarette variables that were added to the simple linear model of Example 14.1.

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: city of residence, categories representing different races *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  $\beta$ s would be meaningless (*eg*

**Example 14.3 Testing the significance of multiple variables**

data = bw5k

In this example, we have added variables representing the number of cigarettes the mother smoked in each of the three trimesters to the model which just contained gestation length. We want to check the overall significance of these smoking variables. The ANOVA table from the full model is shown below:

				Number of obs = 5000
				F(4,4995) = 462.95
				Prob > F = 0.0000
Source	SS	df	MS	R-squared = 0.2705
Model	432936925	4	108234231	Adj R-squared = 0.2699
Residual	1.1678e+09	4995	233792.7	Root MSE = 483.52
Total	1.6007e+09	4999	320210.4	

In the simple model with only -gest- as a predictor, we had  $SSE_{red} = 1178.6 \times 10^6$  with 4,998 df. Hence the  $F$ -test is:

$$F = \frac{(1178.6 \times 10^6 - 1167.8 \times 10^6) / (4998 - 4995)}{0.2338 \times 10^6} = 15.4$$

This  $F$ -statistic is highly significant with 3 and 4,995 df ( $P < 0.0001$ ). It is clear that smoking is associated with birth weight. (**Note** None of the individual regression coefficients for the number of cigarettes smoked was significant. This apparent contradiction with the overall significance of these variables is due to collinearity—see Section 14.5.)

because race 4 is not twice something in race 2, or race 5 is not 3 units more than race 2 *etc*), and would not achieve the desired effect (*eg* removing variation among races when examining the relationship between smoking and birth weights of babies).

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 -black- and -college-; 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) of the variable.

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 Section 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, first let's 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 gestation length was a predictor variable, it has no meaningful interpretation at the value 0 because a baby can't be born after a 0-week gestation). 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  $\beta_0$  will be the value of the outcome at the lowest possible value of the original  $X$ -variable(s), instead of at zero. Alternatively, an ‘average’ value may be subtracted (instead of the lowest possible one). As an example, we might rescale -gest- by subtracting the average gestation length (39 weeks) so  $\beta_0$  will reflect the birth weight of a baby with a 39-week gestation. 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). Scaling can be done by adding or subtracting any meaningful value—not just the two possibilities (lowest possible or average) identified above.

Another use of scaling is when the  $X$ -variable is measured with much greater accuracy than needed (eg regressing -bwt- in grams on number of cigarettes smoked per day during the first trimester which ranged from 0 to 90 in our example). In its original form, the coefficient for -cig\_1- represents the expected effect of one additional cigarette, which might be a very small value. This problem can be circumvented by dividing the value of  $X$  by a suitable constant (eg 20 to convert the measure from cigarettes per day to packs per day). Here, a unit change in the new variable reflects the expected change in birth weight for each pack of cigarettes smoked.

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 a variable called -mrace\_c3- that specifies the mother’s race (in 3 categories). Further, suppose categories are coded as 1, 2, or 3 (or A, B, C) and we wish to control for mother’s race when examining the potential effect of cigarette smoking on birth weight. To do this, we create 2 regular indicator (sometimes called **disjoint**) variables ( $X_1$  and  $X_2$ ) as logical answers to the following questions: was the race 1?; if yes, then  $X_1=1$ , else  $X_1=0$ . For the next indicator variable we ask: was the race 2?; if yes, then  $X_2=1$ , else  $X_2=0$ . With respect to these variables, the values in the table would be present in the dataset.

race	$X_1$	$X_2$
1	1	0
2	0	1
3	0	0

Thus, race 3 is identified as the race with both indicator variables equal to 0, and will be the **referent** (or comparison level or **reference category**) for assessing the effect of races 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  $j^{\text{th}}$  level takes the value 0 for all the indicators (see Example 14.5). As the third race has become the referent level (when all the indicator variables are in the equation),  $\beta_1$  (the coefficient of  $X_1$ ) estimates the difference in the outcome between races 1 and 3, whereas  $\beta_2$  estimates the difference in the outcome between races 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

**Example 14.4 Scaling predictor variables**

data = bw5k

Here we scale -gest- by subtracting 39 (the average observed gestation length) from the actual gestation length, so our new variable is  $\text{gest39} = \text{gest} - 39$ .

bwt	Coef	SE	t	P>t	95% CI	
gest21	124.487	2.942	42.31	0.000	118.719	130.255
constant	3341.136	6.953278	480.51	0.000	3327.505	3354.768

The effect of an increase of 1 week in the scaled variable is the same as in the unscaled variable (Example 14.1; 124.5 gm). In the original scale, -bwt- was predicted to be -1,514 gm for a 0-week gestation; here it is 3,341 gm for a 39-week gestation ( $\text{gest39}=0$ ). In general, the constant will be a sensible number that is easy to interpret and explain if the predictor(s) is appropriately scaled.

(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 race indicators and the race 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 mother’s education level.

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 mother’s education when predicting the effect of gestation length with birth weight.

**Example 14.5 Coding indicator (regular dummy) variables**

data = bw5k

We will demonstrate forming regular (*ie* disjoint) indicator variables from a nominal variable. For example, when considering the variable for the mother’s education (meduc\_c4), the categories are coded 1=less than high school diploma, 2=high school diploma, 3=some college or university, and 4=university/college degree. Let’s assume that it makes sense to use the lowest level as the baseline or referent, then the coding for the 3 indicator variables could be completed by writing code to answer the following questions.

```

If meduc_c4=2      then hs=1          else hs=0
If meduc_c4=3      then coll=1        else coll=0
If meduc_c4=4      then degree=1      else degree=0

```

The effect and significance of each new variable (-hs-, -coll-, and -degree-) would be in relation to having less than a high school diploma. Whether or not the information in the original variable -meduc\_c4- added significantly to the model should be assessed by an *F*-test as shown in Example 14.3.

**Example 14.6 Using and interpreting regular indicator variables in linear regression**

data = bw5k

A model for -bwt- was fit with -gest-, and 3 indicator variables for mother’s education level as the predictors.

bwt	Coef	SE	t	P>t	95% CI	
gest21	124.383	2.938	42.33	0.000	118.622	130.144
meduc_c4=2	20.046	20.980	0.96	0.339	-21.083	61.176
meduc_c4=3	53.270	21.887	2.43	0.015	10.361	96.179
meduc_c4=4	80.599	19.272	4.18	0.000	42.818	118.380
constant	1057.097	53.930	19.60	0.000	951.370	1162.824

In this instance, the referent level is ‘less than high school diploma’; the -bwt- of these babies for a 21-week gestation (gest21=0) is 1,057 gm (the intercept). The coefficient for each education level variable reflects the difference in -bwt- between babies born to those mothers and those born to mothers with less than a high school diploma. We can use a multiple partial *F*-test to test the overall significance of the education level predictors (*P*=0.001). Individually, the 2 highest levels of education are significantly different from the baseline, but having a high school diploma is not.

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’ 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 mother’s race (as above), if only indicator *X*<sub>1</sub> is in the model, the referent will be the weighted average of the outcome in races 2 and 3, and the coefficient associated with *X*<sub>1</sub> will represent the difference in response between race 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 (reflect relative changes in an underlying characteristic, *eg* severity of a disease), 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 mild, moderate, or severe pregnancy-associated hypertension), there might be concern when using these codes as a continuous predictor (*eg* is the biological effect of the difference between mild and moderate the same as between moderate and severe?). 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 immediately preceding it (assuming all hierarchical variables are in the model). As with regular indicator variables, it is possible to include just 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.htm>.

In Example 14.7, we compare mother's education levels using disjoint (*ie* dummy) and hierarchical indicators. The disjoint coefficients reflect the difference in -bwt- in each education level relative to the lowest (baseline) education level (less than high school diploma). With the hierarchical indicators, say for level 4, the regression coefficient reflects the difference in -bwt- between level 3 (some college) and level 4 (university/college degree). It can be seen that, in general, birth weights increased with education level.

#### 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.

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

#### Example 14.7 Indicator vs hierarchical coding of variables

data = bw5k

The effect of mother's education on -bwt- was estimated by using ordinary (disjoint) indicator and hierarchical dummy variables in a linear regression model (which also included -gest-).

Variable	Indicator coding	Hierarchical coding
meduc_c4=2	20.046	20.046
meduc_c4=3	53.270	33.224
meduc_c4=4	80.599	27.329

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 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 error of regression coefficients become very large in a highly collinear model (Section 14.5.1); hence we become less certain of the likely magnitude of the association (*ie* of the true value of  $\beta$ ).

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 ( $\beta$ s) 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 variables collectively contribute significantly to the model. This situation happens when the 3 measures of cigarette consumption are included in a regression model of -bwt-. Collectively, the 3 measures are highly significant, but individually all of their  $P$ -values are  $> 0.3$  (data not shown).

Extreme values of odds ratios (*eg* 8-10 or more) can be used to detect collinearity between dichotomous variables, and extreme correlation coefficients ( $> 0.8-0.9$ ) 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 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 provides 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:

$$VIF = \frac{1}{1 - R_X^2} \quad \text{Eq 14.12}$$

where  $R_X^2$  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  $\beta_1$  for  $X_1$  is from Eq 14.5.

$$\text{var}^{(1)}(\beta_1) = \frac{\text{MSE}^{(1)}}{\text{SSX}_1} \quad \text{Eq 14.13}$$

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  $\beta_1$  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  $\beta_1$  might increase by an amount roughly equal to  $\sqrt{VIF}$ ; specifically,  $\text{var}(\beta_1)$  in the combined model (2) with both  $X_1$  and  $X_2$  is:

$$\text{var}^{(2)}(\beta_1) = \frac{\text{MSE}^{(2)}}{\text{SSX}_1} \times \frac{1}{(1 - R_2^2)} \quad \text{Eq 14.14}$$

where  $R_2^2$  is the coefficient of determination from a regressing  $X_2$  on  $X_1$ . Thus, the standard error of  $\beta_1$  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  $X_2$  can also cause the variance of  $\beta_1$  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) guide 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 *VIFs* 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 previous sections of this chapter, the models 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 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  $X_1 \cdot X_2$  which can be tested in the following model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \varepsilon \quad \text{Eq 14.15}$$

by assessing if  $\beta_3 = 0$  (see Example 14.9). If interaction is absent ( $\beta_3$  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  $\beta_3$  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  $\beta_1$  and  $\beta_2$  as linear effects when the interaction cancels (eg  $\beta_1$  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  $(3-1) \cdot (4-1) = 6$  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

**Example 14.8 The use of centring to avoid collinearity problems**

data = bw5k

There is some evidence that the relationship between gestation length and birth weight is not linear (*ie* not a straight line). One way to deal with this problem is to add a quadratic term to the model (-gest\_sq-). (See Chapter 15 for a much more complete discussion of the issue of non-linearity.) Consequently, a quadratic model with -gest- and -gest\_sq- (-gest- squared) was built.

bwt	Coef	SE	t	P>t	95% CI	
gest	590.914	32.990	17.91	0.000	526.239	655.589
gest_sq	-6.365	0.448	-14.19	0.000	-7.244	-5.485
constant	-9999.664	608.231	-16.44	0.000	-11192.060	-8807.264

Both terms are statistically significant but the effect of -gest- seems absurd (adding 0.59 kg per week). The correlation between -gest- and -gest\_sq- is 0.99 with a resulting *VIF* of 131. We can see the impact of this by noting that the SE of -gest- increased by over 11 times (from 2.94 (simple linear model in Ex 14.1) to 33) when the quadratic term was added.

In order to deal with this problem of collinearity, we can centre the gestation length variable by subtracting its mean (for -gest- this is 39) to create the centred variable -gest39-, and then, we create the squared centred gestation -gest39\_sq-. The summary of this model is shown below.

bwt	Coef	SE	t	P>t	95% CI	
gest39	94.483	3.577	26.42	0.000	87.471	101.495
gest39_sq	-6.365	0.448	-14.19	0.000	-7.244	-5.485
constant	3365.579	7.032	478.60	0.000	3351.793	3379.365

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  $R^2$  (0.292) and root MSE (476.2). Second, we note that the SE of the linear component -gest39- is approximately back to what it was when only the linear term (-gest39-) was in the model. Centring has reduced the correlation between -gest39- and -gest39\_sq- to -0.59 and the *VIF* is now reduced to 1.54. Because gestation length was scaled, the constant in this model represents the predicted -bwt- for a 39-week gestation.

predictor. Example 14.11 shows an interaction between 2 continuous variables.

## 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 and causal interpretation of a multivariable linear model designed to assess the effects of 4 factors (-white-, -college-, -cig\_2-, and -gest-) on birth weight. When making causal inferences, care is needed to ensure that only the appropriate variables are included in the analysis (see Section 13.1). In this regard, a causal diagram is very helpful (our hypothesised causal diagram is shown in Fig. 14.3).

Let us assume that our main objective is to evaluate the effects of cigarette smoking on birth weights. To simplify matters, we will restrict the analyses to single births. The diagram indicates that we are assuming that gestation length (-gest-) is an intervening variable between

**Example 14.9 Interaction between 2 dichotomous variables**  
data = bw5k

The dichotomous versions of maternal weight gain (below or above 30 lb) and total birth order (primiparous/multiparous) were evaluated in a regression model. The model with just these 2 predictors of -bwt- is shown.

bwt	Coef	SE	t	P>t	95% CI	
wtgain_c2	166.970	16.005	10.43	0.000	135.594	198.346
tbo_c2	58.451	16.945	3.45	0.001	25.230	91.671
constant	3165.005	17.050	185.63	0.000	3131.579	3198.432

Both factors are significant. In order to assess if the effect of one of these variables depends on the level of the other, we form an interaction term (a product of the 2 variables) and add it to the model.

bwt	Coef	SE	t	P>t	95% CI	
wtgain_c2	227.464	28.477	7.99	0.000	171.637	283.292
tbo_c2	110.495	26.413	4.18	0.000	58.713	162.277
wg_c2*tbo_c2	-88.379	34.420	-2.57	0.010	-155.857	-20.901
constant	3126.995	22.573	138.53	0.000	3082.742	3171.248

Note that since both -wtgain\_ct- and -tbo\_c2- are dichotomous and coded 0 for ‘no’ and 1 for ‘yes’, the interaction term has the value 1 only when both factors (weight gain >30 and multiparous) are present. In this sense, if it is significant, it says that we need to adjust (using  $\beta_3$ ) the predicted outcome when both factors are present to better reflect what was observed. Otherwise the combined effect of the 2 is just the sum of their individual effects.

There is evidence that the effect of high weight gain depends on whether the birth is primiparous or multiparous (and vice versa) because the interaction term is significant. Note also that the main effect of each variable is significant:

- When neither factor is present (*ie* primiparous births with weight gain <30 lb), the predicted outcome is 3,127 gm.
- In primiparous births with weight gain >30 gm, the predicted outcome is 3127+227=3354 gm.
- In multiparous births with weight gain <30 gm, the predicted outcome is 3127+110=3237 gm.
- When both factors are present, the predicted outcome is 3127+227+110-88=3376 gm.
- All *VIFs* for the model are moderate (<4.3), so collinearity is not an issue.

What this model implies is that the positive effect of multiparous birth on birth weight is present if weight gain is low, but is negligible if weight gain is high (110-88=22 gm). Similarly, high weight gain has a bigger effect in primiparous births (227 gm) than in multiparous births (227-88=139 gm).

-cig- and -bwt- (as is -wtgain-). Consequently, -gest- and -wtgain- will be excluded from the model. See Section 13.11.6 for a discussion of intervening variables.

Because we were primarily interested in the effects of smoking, we made an *a priori* decision to only consider interactions between -cig\_2- and other variables. If we had considered a larger number of potential interactions, we should probably have done something to address the problem of multiple comparisons (see Section 15.7). If we had selected a subset of variables from a much larger ‘pool’ of potential predictors, we should alter the *F*-statistic critical value for significance (Livingston and Salt, 2005).

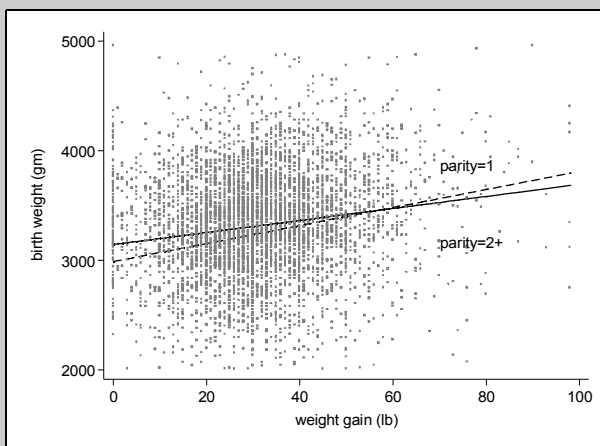
**Example 14.10 Interaction between a dichotomous and a continuous variable**

data = bw5k

We continue with the investigation of the interaction between birth order (primiparous (parity=1) vs multiparous (parity=2+)) and weight gain, but we leave the latter variable as a continuous (but centred) predictor.

bwt	Coef	SE	t	P>t	95% CI	
wtgain_ct	8.221	0.914	8.99	0.000	6.429	10.013
tbo_c2	70.733	17.040	4.15	0.000	37.326	104.140
wg_ct*tbo_c2	-2.728	1.133	-2.41	0.016	-4.949	-0.507
constant	3243.394	14.031	231.16	0.000	3215.887	3270.902

Once again, the interaction term is significant. As before, multiparous births seem to be heavier, but the coefficient (70.7) represents the effect of being multiparous when  $\text{wtgain\_ct}=0$  (ie average weight gain of 31 lb). In primiparous births, the effect of each additional lb of weight gain is to raise birth weight by 8.22 gm. In multiparous births, the positive effect of increasing weight gain appears to be reduced (compared with primiparous). Alternatively, the effect of multiparous appears to be reduced as weight gain increases. 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 -bwt- from the model and plotting it against the continuous predictor (-wtgain-) in primiparous and multiparous births (Fig. 14.2).



**Fig. 14.2 Interaction between birth order and weight gain**

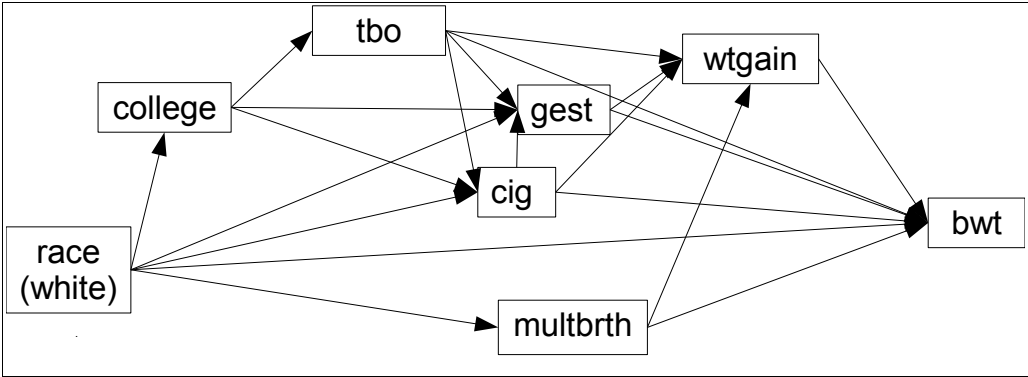
Here we can see the difference in effect of -wtgain\_ct- between multiparous (the solid sloped line) and primiparous (the dashed line) births. The graph indicates that weight gain always has a positive effect on birth weight, but the effect is more pronounced in primiparous births.

## 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 whether the model meets the key assumptions, we might have to alter it). We will use the model shown in Example 14.13 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 -bwt- value of 1 baby does not depend on the -bwt- value of other babies 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



**Fig. 14.3 Causal diagram of factors thought to influence birth weight**

**Note** Variables to the left are assumed to have potential effects on variables to the right with which they are connected by an arrow (eg -white- is assumed to affect -multbrth-, but -college- is not).

**Example 14.11 Interaction between 2 continuous variables**  
data = bw5k

Here we continue our investigation of the interaction between weight gain and total birth order (parity), but we now leave both variables as continuous predictors. The regression with the interaction term is as follows.

bwt	Coef	SE	t	P>t	95% CI	
wtgain	8.530	1.007	8.47	0.000	6.555	10.504
tbo	45.813	12.418	3.69	0.000	21.468	70.158
wg*tbo	-0.929	0.361	-2.58	0.010	-1.636	-0.223
constant	2986.017	36.126	82.65	0.000	2915.194	3056.841

Once again, the interaction term is statistically significant ( $P=0.01$ ) and negative in direction. It is easiest to understand the interaction effect by categorising one of the predictors (in this case -tbo-) and graphing separate lines showing the effect of -wtgain- on -bwt- for various birth orders. According to this model, the impact of weight gain on birth weight decreases as parity goes up (Fig. 14.4).

**Fig. 14.4 Interaction between -wtgain- and -tbo- (parity)**

**Example 14.12 An initial causal model of the impact of several factors on -bwt-  
data = bw5k**

A model was fit to evaluate the effects of cigarette smoking on birth weight. Based on the causal diagram shown in Fig. 14.2, -gest- and -wtgain- were excluded because they were intervening variables. The analysis was restricted to single births. Interactions between -cig\_2- and other variables were considered and the interaction with -white- (mother's race) was retained for didactic purposes even though it was not significant ( $P=0.012$ ).

				Number of obs = 4817
				F(5, 4811) = 21.58
				Prob > F = 0.0000
				R-squared = 0.0219
				Adj R-squared = 0.0209
				Root MSE = 530.58
Source	SS	df	MS	
Model	30372133.9	5	6074426.8	
Residual	1.3544e+09	4811	281516.8	
Total	1.3847e+09	4816	287531.0	

bwt	Coef	SE	t	P>t	95% CI	
white	102.567	16.588	6.18	0.000	70.047	135.088
cig_2	-30.679	10.679	-2.87	0.004	-51.615	-9.743
white*cig_2	17.178	10.945	1.57	0.117	-4.279	38.634
college	43.909	16.419	2.67	0.008	11.719	76.098
tbo	22.662	5.255	4.31	0.000	12.359	32.964
constant	3203.065	18.872	169.72	0.000	3166.067	3240.063

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 ( $F=21.58$   $P<0.001$ ) but it only explains 2.2% of the overall variation in -bwt-. The SE of prediction (531 gm) is only marginally smaller than the original crude SD of 536 gm.

Although not significant, interpretation of the interaction between a dichotomous (-white-) and continuous predictor (-cig\_2-) suggests the following. Among non-white mothers, each additional cigarette smoked during the 2<sup>nd</sup> trimester reduced the -bwt- by 30.7 gm (with a 95% confidence interval of 9.7 gm to 51.6 gm). For white mothers, the effect was  $-30.7+17.2=-13.5$  gm per cigarette per day. Babies from non-smoking, white mothers were, on average, 102.6 gm heavier (than other races) and the discrepancy between white mothers and other races increased as the number of cigarettes smoked increased (by 17.2 gm/day). As the mother's parity increased, so too did the expected birth weight (by 22.7 gm for each additional previous baby) and college educated mothers had heavier babies (43.9 gm). However, all of these estimates are based on the assumption that other factors were held constant. For example, the effect of being white is based on comparing mothers of the 2 racial groups with comparable smoking habits, education levels, and parity.

might signal a lack of independence when there are multiple observations on a single individual, or on multiple individuals within a group (eg multiple babies from the same mother). Methods for dealing with clustered data are presented in Chapters 20–23. A specific type of clustering (serial correlation) is likely to occur when assessing regular measurements from an individual (eg daily weights taken during the first 3 months of life). 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

**Example 14.13 Evaluation of homoscedasticity (equal variances)**  
data = bw5k

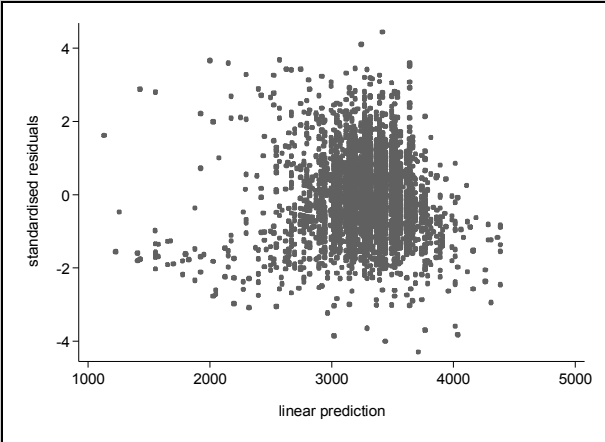
Source	SS	df	MS
Model	442004512	4	110501128
Residual	1.1587e+09	4995	231977.405
Total	1.6007e+09	4999	320210.372

Number of obs = 5000  
F(4, 4995) = 476.34  
Prob > F = 0.0000  
R-squared = 0.2761  
Adj R-squared = 0.2755  
Root MSE = 481.64

bwt	Coef	SE	t	P>t	95% CI	
white	76.842	14.683	5.23	0.000	48.058	105.626
college	25.223	14.570	1.73	0.083	-3.341	53.786
cig_2	-15.430	2.136	-7.22	0.000	-19.618	-11.243
gest	124.263	2.921	42.55	0.000	118.537	129.989
constant	-1552.323	113.118	-13.72	0.000	-1774.083	-1330.563

A scatterplot of standardised residuals vs predicted values based on the model above was generated.

Visually, it appears as if there is substantially more variability in the mid-range of predicted values. However, this can be difficult to assess given that 50% of predicted values fall between 3,170 and 3,443. The Cook-Weisberg test for heteroscedasticity yields a  $\chi^2$ -statistic of 173.7 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 2,500; 3,000; 3,500; and 4,000, suggests that the variance is actually highest in the lowest group (SD=1.76 if  $p_v<2,500$ ) but there are only 86 observations in this group. Although statistically significantly different, the range of SDs for the categories over 2,500 is only from 0.91 to 1.28 suggesting that the problem of heteroscedasticity may not be as serious as it first appears.



**Fig. 14.5 Scatterplot of standardised residuals vs fitted values**

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 -bwt- when the gestation length was 36 weeks should be the same as when it was 42 weeks *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 babies born to college-educated mothers should be normally distributed as they should be for babies born to mothers without a college education). 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-week increase in -gest- from 38 to 39 affects -bwt- by the same amount as a 1-week increase from 41 to 42). 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 is discussed in more detail below (Sections 14.9.1, 14.9.2, and 14.9.4), 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  $i^{\text{th}}$  observation and has the same units as the outcome variable,

$$r_i = Y_i - \hat{Y}_i \quad \text{Eq 14.16}$$

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:

$$\text{var}(r_i) = \sigma^2(1 - h_i) \quad \text{Eq 14.17}$$

where  $h_i$  is the weight of the  $i^{\text{th}}$  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:

$$h_i = \frac{1}{n} + \frac{(X_{1i} - \bar{X}_1)^2}{SSX_1} \quad \text{Eq 14.18}$$

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  $\sigma^2$ , this produces what are called **standardised (std) residuals** (these are also called internally studentised residuals):

$$r_{si} = \frac{r_i}{\sigma \sqrt{1 - h_i}} \quad \text{Eq 14.19}$$

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  $i^{\text{th}}$  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  $i^{\text{th}}$  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):

$$r_{ii} = \frac{r_{-i}}{\sigma_{-i} \sqrt{1 - h_i}} \quad \text{Eq 14.20}$$

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 practise 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 the Breusch-Pagan test (also known as the Cook-Weisberg test) (1983). The null hypothesis is homoscedasticity, so a significant ( $P < 0.05$ ) test result indicates the presence of heteroscedasticity. An evaluation of heteroscedasticity is presented in Example 14.13. (**Note** All subsequent model diagnostics are based on the model shown in Example 14.13.)

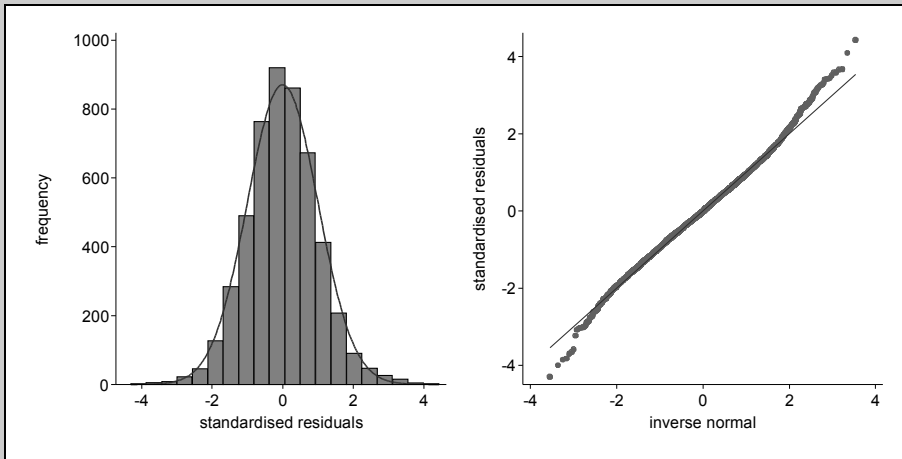
### 14.9.2 Normality of residuals

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  $45^\circ$  to the horizontal (see right side Fig. 14.6, Example 14.14). If the residuals are skewed to the right, the normal plot will curve below the  $45^\circ$  line (the curve is convex), whereas, if the residuals are left skewed, the normal plot will curve above the  $45^\circ$  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 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).

**Example 14.14 Evaluating normality of residuals**

data = bw5k

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

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) both suggest reasonable normality, although the Q-Q plot does identify the fact that there are more extreme (both negative and positive) residuals than would be expected.

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  $W=0.995$  (small values of  $W$  are critical for a normal distribution) with  $p<0.0001$ , indicating non-normality.

**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 is also 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 variance decreases with the outcome and the relationship of  $X$  and  $Y$  is nearly linear, a reciprocal transformation of  $Y$  could prove helpful,
- if  $Y$  is a proportion ( $p$ ) (or more generally, an outcome in a bounded interval but without a binomial denominator) the  $\{\arcsin(p^{1/2})\}$  transformation may stabilise the variance.

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  $Y^{\lambda}$  (except

for  $\lambda=0$ , 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  $\lambda$  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  $\lambda$  to the nearest 1/4 unit (*ie*  $\lambda=0.45$  would be  $\lambda=1/2$ ), or to pick a ‘nice’ value within the 95% confidence interval for  $\lambda$ . In the model from Example 14.13,  $\lambda=1.38$  and in this case, the model was refit with  $-\text{bwt}^{-1.38}$  as the outcome. While this improved the normality of the residuals, the Shapiro-Wilk test remained highly significant (data not shown). (See Section 14.9.6 for a discussion on interpretation of results from this transformed model.) **Note** If the Box-Cox procedure suggests a log transform and there are observations with the  $Y=0$ , a small number (usually the lowest observed value of  $Y$  in your data) should be added 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  $\lambda$  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 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 non-linearity 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 gestation length and  $-\text{bwt}$ .

#### Suggestions for correcting a lack of linearity by transformation

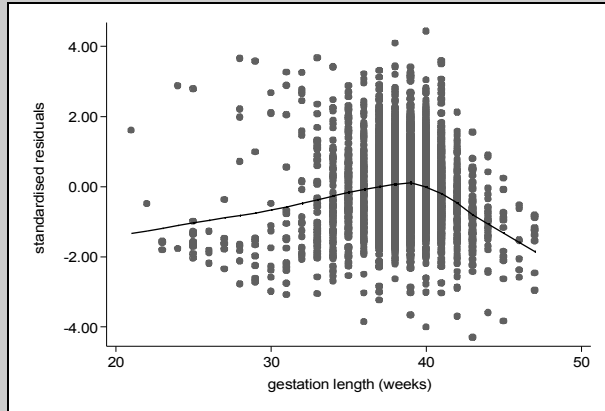
In order to correct a lack of linearity, we can transform the outcome or the predictor(s) or both. As will become apparent, we often have to use transformations to correct for both

**Example 14.15 Evaluating linearity between gestation length and birth weight**  
data = bw5k

A lowess smoothed curve was fit to a scatterplot of the standardised residuals (derived from a model in which -gest- was entered only as a linear term) against birth weight (-bwt-). It is clear that the effect of -gest- is not linear. **Note** This could have been predicted by using a lowess smoothed curve to evaluate the unconditional association between -gest- and -bwt- (plot not shown). See Chapter 15 for a more detailed discussion of evaluating linearity.

Adding a quadratic term for -gest- (-gest39- and -gest39\_sq-) improved the linearity assumption but evidence of non-linearity remained. The revised model had better predictive ability  $R^2$  of 30% compared with 28%, but the problems of

heteroscedasticity and non-normality became more severe (data not shown).



**Fig. 14.7 Lowess smoothed curve through scatterplot of residuals vs gestation length**

heteroscedasticity and lack of normality. 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  $\pm 2$  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). 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  $\sigma^2$  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 cigarette smoking and gestation length. Both remained highly statistically significant, but the SE of -gest- was approximately 50% larger (Example 14.16).

### 14.9.6 Interpreting transformed models

Although visual assessments of homoscedasticity and normality appeared reasonable for the model presented in Example 14.13, there was statistical evidence of violation of both assumptions. Subsequently, a Box-Cox analysis suggested that a  $\text{bwt}^{1.38}$  transformation might be appropriate (details not shown). This transformation reduced the amount of heteroscedasticity (although it remained statistically significant). The transformation also improved the normality of the residuals, but once again statistical evidence of departure from normality remained. The main problem with the distribution of the residuals appears to be a number of quite large positive residuals. Coefficients (or SEs) from the original model and Box-Cox transformed model cannot be compared because they are on different scales (discussed below), but the P-values for the significance can be. In this example, all predictors remain highly statistically significant, providing a degree of comfort that the original model may be satisfactory.

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 is standard practise to use log transformed bacteria counts from microbiological studies of factors that affect bacterial counts.

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-

#### Example 14.16 Final model with robust SEs

data = bw5k

A model was built with birth weight on the original scale with robust SEs.

bwt	Coef	Robust SE	t	P>t	95% CI	
white	76.842	14.584	5.27	0.000	48.252	105.433
college	25.223	14.576	1.73	0.084	-3.353	53.798
cig_2	-15.430	2.559	-6.03	0.000	-20.448	-10.413
gest	124.263	4.398	28.25	0.000	115.640	132.886
constant	-1552.323	171.298	-9.06	0.000	-1888.142	-1216.503

The interpretation of the results of this model are presented in the text (Section 14.9.5).

linear and non-additive (*ie* showing interaction) in another. Thus, the outcome depends on the values of all of the variables in the model even though there is no actual interaction. A recommended practice 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 intervals *etc* are determined in the transformed scale and then back-transformed into the original scale as necessary.

For comparison purposes, we present predicted values from the original model and the Box-Cox transformed model for the 2 levels of college education at 2 levels of gestation length (with all other predictors set to 0).

**Table 14.3 Predicted birth weights from the original and Box-Cox transformed models**

	gest=38		gest=42	
	college=0	college=1	college=0	college=1
original	3170	3195	3667	3692
Box-Cox transformed	3273	3299	3733	3758

**Note** Estimates computed with all other predictors set to 0.

**Note** On both scales, the effect of a college education is 25 gm. The effect of change in gestation length (from 38 to 42 weeks) increased birth weight by 497 gm on the original scale, but by approximately 460 gm in the transformed model. It appears that the effect of gestation length is slightly smaller in the transformed model. For a more detailed discussion of back transformations see Afifi *et al* (2007).

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. However, if a model has a low  $R^2$ , 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 studentised residuals 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), special methods, beyond the level of this text, are available for the assessment of the residuals (Topp and Gómez, 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* being college educated, a heavy smoker and having a long gestation). 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 to 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  $\pm 3$ .

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 with 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 14.20). 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 4.42 would be considered to be an outlier.

$$P = 2 * n * t(dfE, r_{ii}) \quad \text{Eq 14.21}$$

Some general rules in managing outlier observations include:

- identify observations with large studentised residuals
- try to 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
- 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 14.18 which indicates the potential for the  $i^{\text{th}}$  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 this last approach for our example, observation with a leverage above 0.017 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

**Example 14.17 Examination of standardised and studentised residuals**

data = bw5k

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. Based on our data, with 5,000 observations and 4,995 degrees of freedom, an observation with a studentised residual more extreme than  $\pm 4.42$  would be ‘unusual’ with a P-value of  $< 0.05$ . Since we have no babies with such an extreme residual, we conclude there are no serious outliers.

In a dataset of this size, we would expect to see 250 and 50 residuals with absolute values  $> 2$  and  $> 3$ , respectively. We observe that there are 243 and 41, respectively, suggesting a reasonable distribution of extreme residuals.

The 5 smallest (*ie* most extreme negative) standardised residuals (-rsta-) were:

obs	bwt	white	college	cig_2	gest	pv	rsta
2853524	1644	1	0	10	43	3714	-4.30
2946097	1515	0	1	0	40	3443	-4.01
1394806	1165	1	1	0	36	3023	-3.86
747439	2195	0	0	0	45	4040	-3.83
4005523	1985	1	1	0	42	3769	-3.71

Most of these were babies with long gestation periods ( $\geq 39$  weeks), but very low birth weights. If possible, it would be useful to attempt to validate the data about the length of the gestation period.

The 5 largest standardised residuals were:

obs	bwt	white	college	cig_2	gest	pv	rsta
591651	5370	1	1	0	41	3645	3.58
3144236	3759	1	0	0	28	2004	3.65
3666116	4338	0	1	0	33	2574	3.67
283005	5216	1	0	0	38	3247	4.09
368699	5550	0	0	0	40	3418	4.43

These were normal weight babies delivered from very short gestation periods, or exceptionally heavy babies from normal (or slightly prolonged) gestation periods. As above, it would be useful to attempt to verify the data about the length of the gestation.

fitted values with and without observation  $i$  (summed over all other observations and scaled suitably). A more direct interpretation of Cook’s distance derives from Eq 14.22.

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

Eq 14.22

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<sup>th</sup> percentile (not 5%), which is essentially 1, then the

**Example 14.18 Examination of high leverage observations**  
data = bw5k

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

obs	bwt	white	college	cig_2	gest	pv	rsta	lev
3503942	2920	1	0	30	38	2784	0.29	0.016
2797566	3856	1	0	40	40	2878	2.06	0.029
2718923	2997	1	0	40	38	2629	0.77	0.029
3402360	4250	1	0	40	38	2629	3.42	0.029
783858	3340	1	0	45	39	2676	1.40	0.037

All of the highest leverage cases were babies born to mothers with high cigarette consumption.

**Example 14.19 Examination of influential observations**  
data = bw5k

A total of 288 observations had Cook's *D* values above the suggested threshold ( $4/5000=0.0008$ ). The observations with the 10 largest Cook's *D* values are listed below. Most had positive DFIT values.

obs	bwt	white	college	cig_2	gest	pv	rsta	Cook's D	dfit
3406296	1503	1	1	20	35	2590	-2.27	0.009	-0.21
2853524	1644	1	0	10	43	3714	-4.30	0.010	-0.22
1358363	3874	1	1	0	29	2153	3.58	0.010	0.23
3321093	2892	0	0	0	25	1554	2.79	0.011	0.24
3144236	3759	1	0	0	28	2004	3.65	0.013	0.26
373122	2807	0	0	0	24	1430	2.87	0.014	0.26
3118691	4311	1	0	20	37	2814	3.12	0.014	0.27
783858	3340	1	0	45	39	2676	1.40	0.015	0.28
2797566	3856	1	0	40	40	2878	2.06	0.025	0.36
3402360	4250	1	0	40	38	2629	3.42	0.070	0.59

- The most influential observations fell into one of three groups:
- babies whose mothers smoked, but who had birth weights substantially above predicted values
  - babies with very short gestational lengths but with birth weights substantially above predicted values
  - two babies from mothers who smoked but had birth weights much lower than expected.

It will be important to make sure that these observations don't have an undue influence on the model. Removing these values had relatively little effect on the model with the exception that the detrimental effect of cigarette smoking becomes larger (coefficient for -cig\_2- was changed from -15 to -18).

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 dataset, if we use  $4/n$  as the cutpoint, a large Cook's  $D_i$  would have a value exceeding 0.0008 and 288 babies have this value or greater.

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:

$$\text{DFITS}_i = r_{ii} \sqrt{\frac{h_i}{(1-h_i)}} \quad \text{Eq 14.23}$$

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 0.063, and there are a number (292) of observations with larger values than these. The 10 observations with the largest Cook's  $D$  (and also large DFITS) are shown in Example 14.19. Characteristics of those babies are also discussed.

#### 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  $\beta$  for that variable. Critical values for  $n < 120$  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 value was  $\pm 0.028$  and 130 babies exceeded this threshold for the DB for `-cig_2-`. Not surprisingly, these influential observations are all from babies whose mothers smoked heavily ( $\geq 15$  cigarettes per day). However, there is no justification for removing any of these observations from 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 identify some problems with normality and heteroscedasticity, and with the linearity of the gestation length effect. This last issue will be investigated more thoroughly in Section 15.6. We were unable to identify a transformation that completely resolved the first two issues. However,

we felt justified in remaining with the original model because the statistical significance of all of our predictors was never in question and the effects of the predictors did not change dramatically when we used a transformed model. There were no statistically significant outliers.

### 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 daily weights of babies over the first 3 months of life. In this setting, the outcome in one time period (eg one 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, 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 babies with multiple weight measurements, when most data came from babies with only one measurement. 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 (for example, taking daily weight measurements). The set of predictors could be daily measurements of breast-feeding frequency or time spent sleeping. If we analyse such data, to estimate the impact of frequency of feeding on weight, 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 and 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 (Fleury *et al*, 2006), and temporal patterns of fox rabies in Ontario, Canada (Tinline and MacInnes, 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 the analysis of time-series data 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  $t = \text{day}$ ), we can use a variety of smoothing functions of length  $2m+1$  ( $m \geq t$ ). 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 (however, this is 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 ( $m=7$ ) evaluates the correlations between observations ranging from 1–7 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 daily weight would be expected to be most strongly correlated between adjacent days and the correlation tends to decrease as the number of days between measurements increases. 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 correlation 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 one 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  $X_s$ ). The  $X_s$  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 ( $X_{ts}$ ) 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.

$$Y_t = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + \alpha_1 Y_{t-1} + \alpha_2 Y_{t-2} + \varepsilon_t \quad \text{Eq 14.24}$$

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 weight on day  $t$  and  $X_{1t}$  could be nursing frequency on day  $t$ , and  $X_{2t}$  could be a measure of the time slept on day  $t$ . As shown in Fleury *et al* (2006),  $Y_t$  could be the daily number of enteric disease cases and the  $X_s$  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 auto-regressive and moving average processes. An ARMA-11 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 SJ. Factors affecting the severity of horse-related injuries. *Injury*. 2003;34(12):897-900.
- Afifi AA, Kotlerman JB, Ettner SL, Cowan M. Methods for improving regression analysis for skewed continuous or counted responses. *Ann Rev Public Health*. 2007;28:95-111.
- Austin PC, Hoch JSJ. Estimating linear regression models in the presence of a censored independent variable. *Stat Med*. 2004;23(3):411-29.
- Chao YC, Zhao Y, Kupper LL, Nylander-French LAJ. Quantifying the relative importance of predictors in multiple linear regression analyses for public health studies. *J Occup Environ Hyg*. 2008;5(8):519-29.
- Cheung YB. A modified least-squares regression approach to the estimation of risk difference. *Am J Epidemiol*. 2007;166(11):1337-44.
- Cook RD, Weisberg S. Diagnostic for heteroscedasticity in regression. *Biometrika*. 1983;70:1-10.
- Diggle PJ. Time series, a biostatistical introduction. Oxford, UK: Clarendon Press; 1990.
- 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 Jul;50(6):385-91.
- Fuller W. Measurement Error Models. New York: Wiley; 2006.
- Hartnack S, Doherr MG, Alter T, Toutounian-Mashad K, Greiner MJ. Campylobacter monitoring in German broiler flocks: an explorative time series analysis. *Zoonoses Public Health*. 2009;56(3):117-28.
- Livingston DJ, Salt DW. Judging the significance of multiple linear regression models. *J Med Chem*. 2005;48:661-3.
- Ljung GM, Box GEP. On a measure of a lack of fit in time series models. *Biometrika*. 1978;65:297-303.
- Marill KAJ. Advanced statistics: linear regression, part II: multiple linear regression. *Acad Emerg Med*. 2004a;11(1):94-102.



- Marill KAJ. Advanced statistics: linear regression, part I: simple linear regression. *Acad Emerg Med.* 2004b;11(1):87-93.
- Murad H, Freedman LSJ. Estimating and testing interactions in linear regression models when explanatory variables are subject to classical measurement error. *Stat Med.* 2007;26(23):4293-310.
- Pires AM, Rodrigues IM. Multiple linear regression with some correlated errors: classical and robust methods. *Stat Med.* 2007;26(15):2901-18.
- Pitard A, Viel JF. Some methods to address collinearity among pollutants in epidemiological time series. *Stat Med.* 1997 Mar 15;16(5):527-44.
- Poirier E, Watier L, Espie E, Weill FX, De Valk H, Desenclos JCJ. 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(9):1217-24.
- Tinline RR, MacInnes CD. Ecogeographic patterns of rabies in southern Ontario based on time series analysis. *J Wildl Dis.* 2004 Apr;40(2):212-21.
- Topp R, Gómez 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(2):319-23.
- Zeger SL, Irizarry R, Peng RD. On time series analysis of public health and biomedical data. *Ann Rev Public Health.* 2006;27:57-79.

