15

MODEL-BUILDING STRATEGIES

OBJECTIVES

After reading this chapter, you should be able to:

1. Develop a 'full' (maximal) model which incorporates your biological understanding of the system being investigated.

2. Carry out procedures to reduce a large number of predictors to a more manageable subset.

3. Address key issues related to the predictors (*eg* functional form of the relationship between a continuous predictor and the outcome dealing with missing values).

4. Build regression-type models while considering statistical and non-statistical criteria.

5. Evaluate the reliability of a regression-type model.

6. Present the results from your analysis in a meaningful way.

15.1 INTRODUCTION

When building a regression model, we need to decide on the goals of the analysis, to recognise the need to incorporate both statistical considerations and our subject matter knowledge into that process, and to balance the desire to get the model which 'best fits' the data with the desire for parsimony (simplicity in the model). As will become apparent, the definition of 'best fit' depends on the goal of the analysis. Throughout this chapter (unless otherwise specified), the principles discussed relate to all types of regression model but will generally be presented in the context of a linear regression model.

Goals of the analysis
Regression models are generally built to meet one of 2 broad objectives. One goal might be to come up with the best model for predicting future observations. The details of this model (*eg* the effects of specific predictors) might be of little consequence, but we want to keep any variables out of the model whose relationship with the dependent variable is questionable. If these variables are included, and a future observation has a relatively extreme value for one of those variables, the prediction might be inaccurate.

More often in epidemiology, the goal is to understand the relationship(s) (potentially causal) between one or more predictors and the outcome of interest. In this case, you want to obtain the most precise estimates of coefficients possible for the variables of interest. Careful attention must be paid to any interaction and confounding effects.

Role of subject matter knowledge
Subject matter knowledge must guide model-building. If the goal is simply to build a predictive model, the role of subject matter knowledge is to prevent the inclusion of variables not likely to be generally related to the outcome of interest. As noted, inclusion of these could make future predictions inaccurate.

If the goal is understanding biological relationships, it is important that factors which are likely to be confounders should be retained in the model, regardless of their statistical significance. On the other hand, inclusion of factors which are almost certainly not confounders (see Chapter 13 for criteria for confounding) may result in biased results. This is most likely to happen if intermediate (intervening) variables are included in the analysis. Building a causal diagram is an essential first step in any model-building exercise in which the objective is to understand the relationships between predictors and the outcome (more on this in Section 15.3).

Subject matter knowledge may also help in the selection of variables. For example, choosing among collinear variables is facilitated if you are able to take into consideration the difficulty of measuring each of the predictors and their perceived reliability.

Parsimony vs fit
In general, parsimony (using as few predictors as required to obtain a good fit) should be your guide, but do not exclude variables that you have reason to believe (*ie* for biological reasons) should be in the model. Remember, the goal of most statistical analyses is to extract meaningful results from a complex dataset. If the final results are almost as complex as the original data, nothing has been gained. (If the number of regression coefficients equalled the number of observations in the dataset, we could have a perfect fitting model, but would have gained nothing.) Simple models are more robust, less likely to be influenced by specific idiosyncrasies of the existing data, and consequently will perform better if applied to new data.

15.2 Steps in building a model

The steps involved in building a regression model are:

1. Specify the maximum model to be considered (*ie* identify the outcome and the full set of predictors that you want to consider).
2. Specify the criterion/criteria to be used in selecting the variables to be included in the model.
3. Specify the strategy for applying the criterion/criteria.
4. Conduct the analyses.
5. Evaluate the reliability of the model chosen.
6. Present the results.

15.2.1 Specifying the maximum model

The first step in specifying the maximum model is to identify the outcome variable and determine whether it is likely to need transformation (*eg* natural log transformation) or other form of manipulation (*eg* recategorisation of an outcome variable). Discussion of issues related to the outcome variable is presented in the chapters dealing with specific modelling techniques (*eg* Chapter 14 for linear regression models, Chapter 16 for logistic models).

The maximum model includes all possible predictors of interest. There are pros and cons to making the maximum model very large. On one hand, it will prevent you from overlooking potentially important predictors. On the other hand, however, adding a lot of predictors increases the chances of:

- collinearity among predictor variables (if 2 or more independent variables are highly correlated, the estimates of their coefficients in a regression model will be unstable), and
- including variables that are not important 'in the real world', but happen to be significant in your dataset. (Interpretation of these results might be difficult and the risk of identifying spurious associations is high.)

When specifying the maximum model, you need to identify which variables should be included in the model-building process, how many should be included, and whether or not interaction terms need to be considered. Bear in mind that building the maximum model is as much a scientific/clinical task as it is a statistical one. The steps involved in specifying the maximum model include:

- drawing a causal diagram
- potentially reducing the number of predictors being considered
- considering the impact of missing values
- evaluating the effects of continuous predictors
- deciding what interactions are to be considered.

15.3 Building a causal model

It is imperative that you have a causal model in place before you begin the model-building process. This model is usually presented as a causal diagram. These were introduced in Chapter 13 and a much more complete discussion of causal diagrams is presented elsewhere (Rothman *et al*, 2008, Chapter 12). The diagram will identify potential causal relationships among the predictors and the outcome of interest. For example, if you were interested in evaluating the

effects of cigarette smoking on birth weight, and also had recorded data on mothers' race and education level, total birth order (parity), gestation length, and number of babies born as well as weight gain during pregnancy, then a putative causal diagram might look like Fig. 15.1. (**Note** There are other data recorded in the dataset bw5k. This subset has been chosen for pedagogical purposes.)

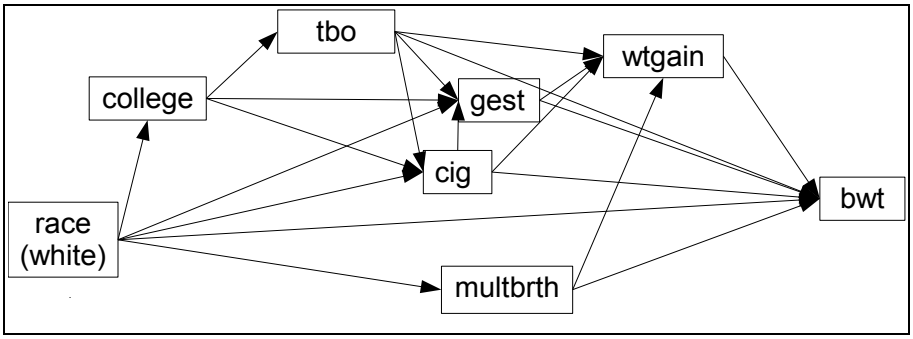

Fig. 15.1 Putative causal diagram for effects of cigarette smoking on birth weight

If the objective of the study was to quantify the effects of cigarette smoking on birth weight, you would NOT include either of the intervening variables (gestation length or weight gain) in the regression model. Inclusion of these intervening variables would remove any of the effect from cigarette smoking that was mediated through them. On the other hand, if race or college education are suspected to be important confounders, they might be designated to remain in the model regardless of whether or not they are statistically significant.

Even if a study has a large number of predictors, it is essential to start with a causal structure in mind and this can often be drawn by grouping variables into logical clusters (*eg* all demographic variables together, all behaviour measures together).

15.4 Reducing the number of predictors

It is sometimes necessary to reduce the number of predictors to be considered in a model-building process. However, before proceeding with an overview of the approaches for reducing a large number of predictors, we must point out that, in many cases, the most appropriate procedure would be to design a study which was much more focused and which collected high-quality data on far fewer predictors. This will greatly reduce the risk of identifying associations for which making a causal inference is precarious.

There are various ways to reduce the number of predictors that need to be considered for inclusion in a regression model. These include:

- screening predictors based on descriptive statistics
- correlation analysis of independent variables
- creation of indices
- screening variables based on unconditional associations
- principal components analysis/factor analysis
- correspondence analysis.

These will each be reviewed briefly and more detail can be found in Dohoo *et al* (1997). However, before any reduction in the number of independent variables is undertaken, it is essential to identify the primary variables of interest and any other variables for which there is already evidence that they might be confounders or interacting variables. These should always be retained for consideration in the model.

15.4.1 Screening predictors based on descriptive statistics

It is crucial to become thoroughly familiar with your data before starting any model-building (Chatfield, 2002). Descriptive statistics (means, variances, percentiles *etc* for continuous variables and frequency tabulations for categorical variables) can be helpful in identifying variables which might be of little value in your model. Keep in mind that, in general, you want to keep variables that you are confident have been measured accurately and precisely, and which are relatively complete. Some specific guidelines follow.

- Avoid variables with large numbers of missing observations (see Section 15.5 for dealing with missing data).
- Select only variables with substantial variability (*eg* if almost all of the babies in a study are males, adding sex as a predictor is not likely to be helpful).
- If a categorical variable has many categories with small numbers of observations in each, consider combining categories (if this makes biological sense), or eliminating the variable.

15.4.2 Correlation analysis

Examining all pairwise correlations among predictor variables will identify pairs of variables that contain essentially the same information. Inclusion of highly correlated variables will result in multicollinearity in the model, potentially producing unstable estimates of coefficients and incorrect standard errors. Collinearity will often be a problem with correlation coefficients greater than 0.9, but could occur at lower levels. If pairs of highly correlated variables are found, one of them should be selected for inclusion in the model based on criteria, such as biological plausibility, fewer missing observations, ease, and/or reliability of measurement.

Note Examining correlations among variables in a pairwise manner will not necessarily prevent multicollinearity, because the problem can also arise from correlations among linear combinations of predictors. However, screening based on pairwise correlations will remove one potential source of the problem.

Note Correlations are really only valid for continuous predictors, but in practise, checking correlations among dichotomous predictors is a convenient way of identifying highly collinear predictors. These relationships can be further assessed using cross-tabulations.

15.4.3 Creation of indices

It might be possible to combine a number of related predictor variables into a single index that represents some overall level of a factor. This might be done subjectively based on the perceived importance of the contribution of a number of factors. For example, the Hamilton Rating Scale for Depression combines information about 22 characteristics (*eg* feelings of guilt, agitation) into an overall depression scale (Hamilton, 1960). The weight assigned to each factor might be subjectively assigned although, if possible, they should be based on evidence from

previous research. Alternatively, data on a number of factors can be combined in an objective manner if procedures to do so exist. For example, data on fan capacity, size and shape of air inlets, and room size might be used to compute the number of air changes per hour in a hospital room. This might then be expressed as the proportion of a recommended ventilation level. One drawback to the creation of indices is that it precludes the evaluation of the effects of individual factors which were used to create the index (see discussion of suppressor variables in Section 13.11.8).

In a situation in which data on a number of related predictors have been recorded, and it is reasonable to assume that the individual predictors are all reflective of some underlying but unmeasured characteristic (also called a latent variable), those items may be combined into an index (or **scale**). **Cronbach's alpha** may be used to evaluate the internal consistency of the scale (*ie* evaluate how well each predictor correlates with the scale). The scale is simply the sum, or average, of the values of the individual predictors (called items), so these must be standardised if they are not measured on the same scale. Cronbach's alpha (also called a reliability coefficient) is the square of the correlation between the scale and the underlying characteristic. Suggested guidelines for the interpretation of Cronbach's alpha are: <0.60 unacceptable, 0.60–0.65 undesirable, 0.66–0.70 minimally acceptable, 0.71–0.80 respectable, 0.81–0.90 very good, and >0.90 consider shortening the scale by reducing the number of items (Dukes, 2007).

In addition to looking at Cronbach's alpha, it is useful to evaluate the correlations between each item and the scale (or a scale generated without the item of interest) and with other items in the scale. This will identify items which do not fit well in the scale. Example 15.1 shows the use of Cronbach's alpha for the simple task of combining 3 measures of smoking into an overall measure of smoking during pregnancy. A more substantial example—the use of Cronbach's alpha to evaluate the reliability of scales measuring patient satisfaction with out-of-hours primary medical care—can be found in McKinley *et al* (1997).

15.4.4 Screening variables based on unconditional associations

One of the most commonly used approaches to reducing the number of predictor variables is to select only those that have unconditional associations with the outcome that are significant at some very liberal P-values (*eg* 0.15 or 0.2). The types of test used to evaluate these associations will depend on the form of the outcome and predictor variables. However, simple forms of a regression model (*eg* a linear or logistic regression model with a single predictor) will always be appropriate for this investigation. For example, univariable multinomial regression models were used to identify variables to be used subsequently in a multivariable multinomial (proportional odds) regression model (Polgreen *et al*, 2008).

One drawback to this approach is that an important predictor might be excluded if its effect is masked by another variable (*ie* the effect of a predictor only becomes evident once a confounder is controlled) (see distorter variables, Section 13.11.7). Using a liberal P-value helps prevent this problem. Another approach is to build a model with the statistically significant predictors and subsequently add all eliminated predictors, one at a time, back into the final model. If the confounder was included in the final model, the eliminated predictor might then turn out to have a statistically significant association and be added back into the model.

This process of screening predictors individually can be extended to include building multivariable models using mutually exclusive logical subsets of predictors to identify the key predictors in each subset, which are then retained for consideration in a final multivariable model.

Example 15.1 Cronbach's alpha
data = bw5k

Three highly correlated variables related to cigarette smoking during pregnancy were considered:
cig_1 - daily cigarette consumption during the 1st trimester
cig_2 - daily cigarette consumption during the 2nd trimester
cig_3 - daily cigarette consumption during the 3rd trimester

Because the variance of these 3 variables differed, Cronbach's alpha was constructed using standardised values for each predictor.

item	obs	sign	item-test correlation[a]	item-rest correlation[b]	average inter-item correlation[c]	Cronbach's alpha[d]
cig_1	5000	+	0.943	0.873	0.945	0.972
cig_2	5000	+	0.982	0.958	0.833	0.909
cig_3	5000	+	0.962	0.915	0.888	0.941
Test scale					0.889	0.960

[a] correlation between item and the scale (average of all items)
[b] correlation between item and a scale based on all other items
[c] average correlation among all other items
[d] Cronbach's alpha for a scale based on all other items

The overall reliability coefficient 0.96 is very high (estimated correlation between the scale and underlying characteristic is $\sqrt{0.96}$=0.98). While the correlations between individual items and the scale are also very high (0.94–0.98), with only 3 items in the scale, each item contributes substantially to the scale. A better evaluation of each item is found by looking at the correlations between items and a scale built without the item of interest included (item-rest correlation). This identifies -cig_1- as the item with the lowest correlation to other items. The average correlation among the other items is also highest if -cig_1- is omitted (0.945).

15.4.5 Principal components analysis, factor analysis, and correspondence analysis

Principal components analysis and factor analysis are 2 closely related techniques that can be used to consolidate the information contained in a set of predictor variables into a new set of uncorrelated (*ie* orthogonal) predictor variables. A detailed discussion of the techniques is beyond the scope of this book, but they will be summarised briefly. Both are designed primarily to work with quantitative (continuous) predictors, but techniques are available to allow categorical predictors to be included.

Principal components analysis is used to convert a set of k predictor variables into a set of k orthogonal, principal components with each successive component containing a decreasing proportion of the total variation among the original predictor variables. Because most of the variation is often contained in the first few principal components, a small subset is often selected for use as predictors in the regression model. The composition of the principal components does not vary depending on the number of components selected for retention. Once the regression model has been built with this subset of the principal components, the resulting coefficients can be back-transformed to obtain coefficients for the full set of original predictors. This resulting set of coefficients will be more stable than those from a model built directly from the original predictors because the problem of multicollinearity has been eliminated. However, there will be no evaluation of the statistical significance of each of the predictors and hence, no determination of which ones are most 'important'.

Factor analysis is a closely related technique, but is based on the assumption that a set of factors that have inherent meaning can be created from the original variables. For example, Khader *et al* (2011) used factor analysis to evaluate interrelationships among cardiovascular risk factors which are components of metabolic syndrome X. Unlike principal components, the composition of the factors does vary as the number of factors selected for creation varies. The strength of a factor analysis rests with the plausibility of the assumption that the factors are truly measuring an underlying latent structure (*eg* having a set of factors related to high blood pressure). If this assumption is valid, then knowing which of those underlying structures are associated with the outcome (*eg* metabolic syndrome X) might be as important as information about individual predictor variables. Establishing which of the original predictors are important determinants of the outcome is a subjective process based on the predictors that are highly correlated (or have high 'factor loadings') with factors found to be significant predictors of the outcome. As with principal components analysis, there is no statistical testing of individual predictors.

Correspondence analysis is a form of exploratory data analysis designed to analyse the relationships within a set of categorical variables. One of the main objectives of correspondence analysis is to produce a visual summary (usually 2-dimensional) of the complex relationships that exist within a set of categorical variables (both predictors and the outcome). The 2 axes are factorial and reflect the most 'inertia' (variability) in the original predictor variables. The result is a scatterplot which identifies clusters of predictors that are closely associated, with clusters farther from the intersection of the axes having stronger associations. After considering relationships among the predictors, the values of the outcome variable (also categorical) can also be projected on the same axes to determine which clusters of predictor variable values are associated with the outcome(s) of interest. A correspondence analysis of a subset of the risk factors for birth weight is presented in Example 15.2.

While principal components analysis, factor analysis, and correspondence analysis can be used to deal with the problem of large numbers of independent variables, they are perhaps better viewed as complementary techniques to model-building procedures. They provide insight into how predictor variables are related to each other, and ultimately into how groups of predictors are related to the outcome of interest.

15.5 The problem of missing values

Missing data are common in observational studies. Statistical programs for building regression models work on the basis of **complete case analysis**—that is, they only use observations for which there are no missing values for the outcome variable or any of the predictors. Consequently, even a relatively low overall percentage of missing values can result in a substantial reduction of the sample available for analysis if those missing data points are spread across observations. The complete case analysis can therefore be severely inefficient (*ie* have reduced power), but it can also induce bias if the complete cases are not representative of the full sample. To further discuss this, it is useful to distinguish between 3 possible mechanisms underlying missing values, and also between whether the missing data occur among the outcomes (Y) or the predictors (X). The missing-data mechanism concerns the reasons why some values are missing and, in particular, how these reasons might relate to values in the dataset.

Data may be **missing completely at random** (MCAR) if the missing values are truly randomly distributed throughout the dataset (*eg* due to a sample being split and the results of that test

Example 15.2 Correspondence analysis of risk factors associated with birth weight
data = bw5k

Multiple correspondence analysis was used to visually assess the relationships among several risk factors and birth weight. Risk factors were dichotomised and birth weight was converted to a 3-level categorical variable as follows.

Factor	Description
race	white, non-white
mother's education	college, less than college (no_coll)
cigarettes in 2nd trimester	smoker, non-smoker
birth weight	light (<3000 gm), med. (3000–3499 gm), heavy (≥3500 gm)

Correspondence analysis was used to visually evaluate the relationships among these variables with the results presented in Fig. 15.2.

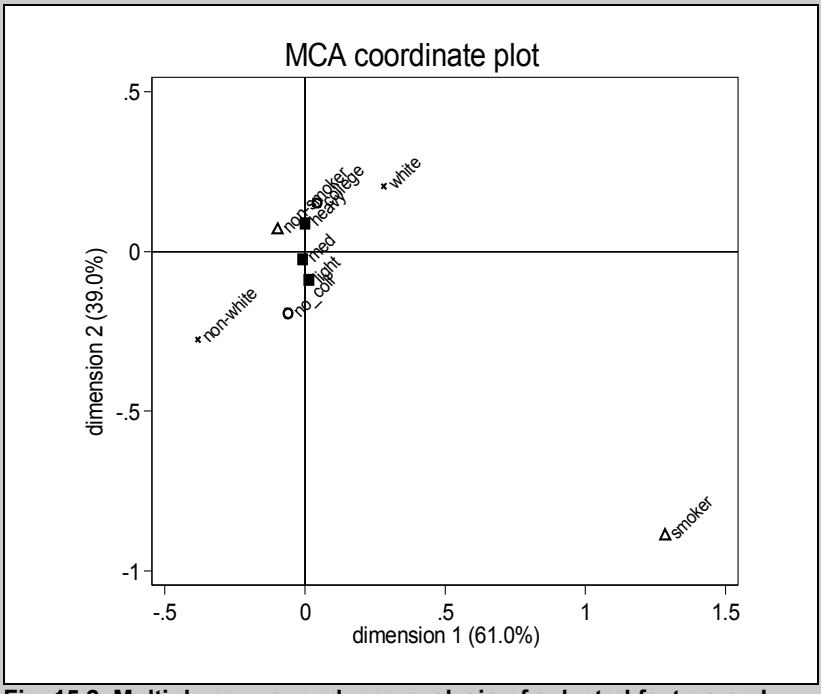

Fig. 15.2 Multiple correspondence analysis of selected factors and their relationship with birth weight

Among the risk factors, non-smoking was closely associated with college education, and having less than a college education was somewhat associated with being non-white. Variation in birth weight was nearly all along dimension 2 which was heavily influenced by smoking status (and, to a lesser extent, college education and race).

consequently being missing). It could be said that whether or not a value is missing could be likened to the tossing of a coin. However, MCAR does not require the probability of being missing to be equal to 0.5, nor even to be constant across the entire dataset. When considering missing values of outcomes, the probability of missingness is allowed to depend on the (observed) predictors, because the inference in regression models is conditional on the predictors. Therefore, for an MCAR assumption to hold, it is important to include as predictors any variables that may be associated with the missingness (*eg* time in a repeated measures study) (Fitzmaurice *et al*, 2004). Missing values of predictors may similarly be allowed to depend on either outcomes or other predictors without missing values. Under MCAR missingness, complete case analysis estimates will not be biased (Little and Rubin, 2002), but for missingness among the predictors only this also holds true under less restrictive assumptions (Donders *et al*, 2006; Harel and Zhou, 2007; Rubin, 2004).

If the observed data do not constitute a random sample of the full (unobserved) data, the missingness is no longer MCAR. If the probability of being missing can be completely explained by non-missing values in the data, either for the subject itself (if multiple outcomes are available at each subject) or for other subjects, then the missing data are called **missing at random** (MAR—*ie* they are missing at random, conditional on the observed values). It may be useful to contrast MAR with the alternative scenario (beyond MCAR): **missing not at random** (MNAR, or sometimes NMAR). Here the missingness depends on the unobserved data, *ie* the data one would have obtained if the missingness had not occurred. If the fact that an observation was not obtained was linked to its (potential) value, this information is part of the evidence obtained in the study and must be included in the analysis to avoid bias. Complete case analysis will generally produce biased estimates in MAR and MNAR scenarios for outcomes; the bias depends on the proportion of missing values and the strength of their association with the observed or unobserved outcomes.

The 2 main alternative methods to a complete case analysis are: (i) **imputation**, and (ii) analysis of the incomplete data by methods where the missing data are **ignorable**, *ie* the method is robust to missing data of the assumed form (Little, 2007). **Imputation** involves replacing the missing data points with values predicted from the available data for that observation. For missing values of a predictor variable, this prediction can be based solely on other predictors, or can include the observed outcome value for that variable (Moons *et al*, 2006). Single imputation involves deriving a single estimate for each missing value. However, an analysis based on single imputed data does not take into account the uncertainty associated with the estimated values. Multiple imputation involves generating multiple imputed datasets and combining results from the analyses of all of these datasets. It is generally accepted that multiple imputation is preferred to single imputation. Imputation may eliminate (MAR) or reduce (MNAR) the bias resulting from missing values. Methods for imputation is an active research area and a detailed discussion of the topic is beyond the scope of this text; 2 recent review publications which introduce the subject are Donders *et al* (2006) and Harel and Zhou (2007), and a relatively recent text on the subject is Rubin (2004).

Maximum likelihood (ML) estimation and Bayesian estimation (which in this context is closely linked to multiple imputation, see Chapter 24) are the main examples of procedures that make MAR missing values **ignorable**. In principle, ML estimation requires specification of the distribution of the missing values, but for outcome missing values, this is unnecessary under the MAR assumption (Fitzmaurice *et al*, 2004; Little, 2007). Implementation of ML procedures for missing covariates in logistic regression has been described (Vach, 1994; Vach and Blettner, 2007). In addition to imputation and use of robust procedures, a wealth of models and

procedures exist for dealing with missing values under MNAR assumptions in different contexts. This is also an active research area, and in particular "Statistics in Medicine" is a valuable source for current (and older) developments. For further discussion of missing data, we refer to Little and Rubin (2002), the standard statistical text on missing data.

15.6 Effects of continuous predictors

It is important to evaluate the structure of the relationship between a continuous predictor and an outcome (which could be a quantity as in a linear regression, the log-odds of disease in a logistic model *etc*). The underlying assumption of linearity can be evaluated when carrying out diagnostics for the model (*eg* evaluation of residuals) and this has the advantage that it evaluates the effects of a continuous predictor after adjustment for other predictors in the model. However, for practical purposes, it is useful to explore the nature of the relationship before starting model-building.

Some approaches to evaluating this relationship include:

- scatterplots and smoothed line plots
- converting the predictor to an ordinal variable (categorisation)
- exploring polynomial models
- using linear or cubic splines.

15.6.1 Scatterplots/smoothed line plots

Scatterplots are 2-way plots of the outcome (on the Y-axis) vs the continuous predictor (as shown in Fig. 15.3—a plot of the relationship between gestation length and birth weight. They are primarily useful for models with continuous outcomes (a scatterplot of a dichotomous outcome presents as 2 lines of dots at Y=0 and Y=1). By themselves, they rarely provide a clear indication of the nature of the functional relationship between the predictor and the outcome (you can imagine how difficult it would be to identify a curvilinear relationship just looking at the 'dots' in Fig. 15.3).

Smoothed lines
Scatterplots can be greatly improved by the addition of a **smoothed line** through the centre of the data, and there are multiple ways that this line can be constructed. All smoothed lines have a **local-influence property** in that the position of the line at any value of x (x_i) is influenced by points close to x_i, but not by points at a large distance from x_i. Smoothed-line plots are constructed as follows:

- for each value of the predictor (x_i), select a number of points on either side of that value (usually done symmetrically)—this set of points will be the 'neighbourhood'.
- compute an expected value of the outcome at x_i—this can be computed as:
  - a simple average of the y values of the observations in the neighbourhood (**running mean smoother**)
  - the predicted value from a simple linear regression through the observations in the neighbourhood (**running line smoother**)
  - the predicted value from a weighted linear regression through the observations in the neighbourhood (**lowess smoother**) so that points close to x_i get larger weight—the most usual form of weighting is Cleveland's tricube weighting (Cleveland, 1979)

- the predicted value from a weighted polynomial regression through the observations in the neighbourhood (**local polynomial smoother**)—weight can be based on a variety of distributions (*eg* normal, Epanechnikov *etc*) (beyond the scope of this book).
- repeat the process for all values of x in the range of the dataset.

The size of the neighbourhood can be controlled by setting the bandwidth. A bandwidth of 0.8 means 80% of all of the data goes into the neighbourhood used to estimate each point. The larger the neighbourhood used for each point, the smoother the line will be, but the greater the danger of missing important features of the relationship. Fig. 15.3 shows a lowess smoothed line (bandwidth=0.8) superimposed on the scatterplot of birth weight vs gestation length. Fig. 15.4 shows running mean, running line, and lowess smoothed lines for the same data.

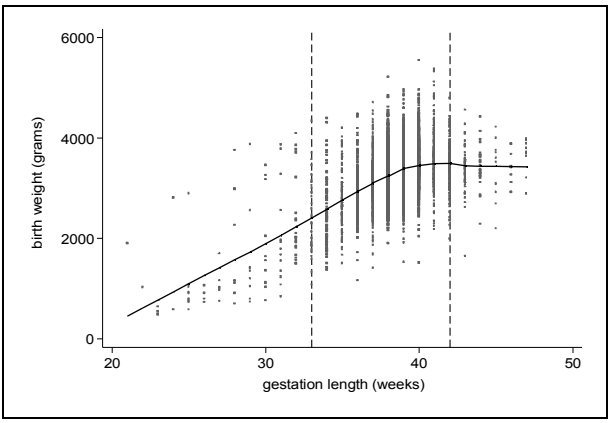

Fig. 15.3 Scatterplot of birth weight vs gestation length with lowess smoothed curve added

Note Vertical dashed lines mark the 2.5th and 97.5th percentiles of gestation length

Note All smoothed line functions can have problems reliably portraying the data at the extreme values of the distribution, because the neighbourhood is not symmetrical about x_i and may, in fact, contain relatively few data points. For this reason, it is important not to pay much attention to the position of the line at each end. This can be facilitated by adding an element to the graph that delineates where most of the data fall (in this case, dashed vertical lines). Adding 95% CI to the smoothed line (Fig. 15.5) also shows the problem of predicting the nature of the relationship at the extremes of the predictor.

Smoothed lines on a logit scale

Skip this section unless you are familiar with logits and logistic regression (Chapter 16). Although scatterplots of a dichotomous outcome are uninformative, smoothed lines can be computed on the logit scale. This is done by computing the smoothed value (probability) for all of the data points in the neighbourhood and then converting this value to the logit scale. Fig. 15.6 shows a lowess smoothed curve for the relationship between gestation length and the log odds of a baby having low birth weight (relationship appears non-linear).

15.6.2 Categorising continuous predictors

The assumption of linearity can be avoided by categorising the continuous predictor into 2 or more categories. While this might provide some insight into the nature of the relationship, it is not generally advisable for 3 reasons. First, categorisation involves the loss of information. Second, it is unlikely that biological processes have a step-function relationship (*ie* sudden changes in the outcome at specific values of the predictor). Finally, the choice of the cutpoints is arbitrary and, if points are chosen based on the observed data, this may lead to biased results (Royston *et al*, 2006). However, if a continuous variable is categorised, it has been suggested

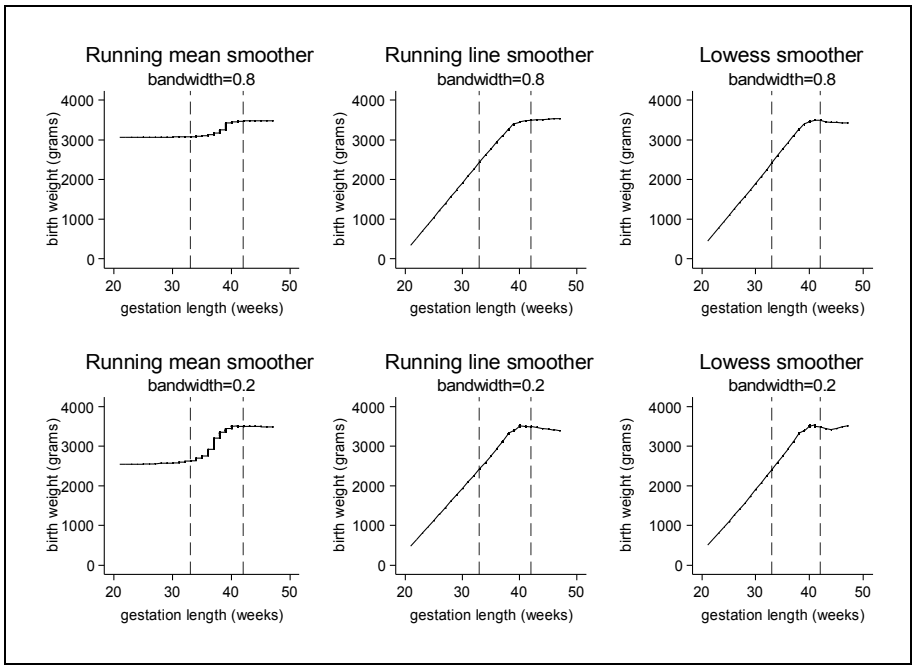

Fig. 15.4 Smoothed line estimates of the relationship between birth weight and gestation length

Note Vertical dashed lines mark the 2.5th and 97.5th percentiles of gestation length.

that 5 categories will usually suffice to control the confounding effects of that variable adequately (Cochran, 1968). A model containing a categorised variable can be compared with one with a continuous variable (linear effect) by using their AIC or BIC values (see Section 15.8.1).

15.6.3 Polynomial models

Polynomials arise when power terms (*eg* x^2 or x^3) are added to a linear model to allow the regression line to follow a curve rather than a straight line through the data. The complexity of the curve (*ie* number of bends) depends on the number of power terms included in the polynomial. Quadratic polynomials are the most commonly used, but fractional polynomials deserve careful consideration as well. Polynomial models have a **global-influence property** in that the shape of the line is influenced by

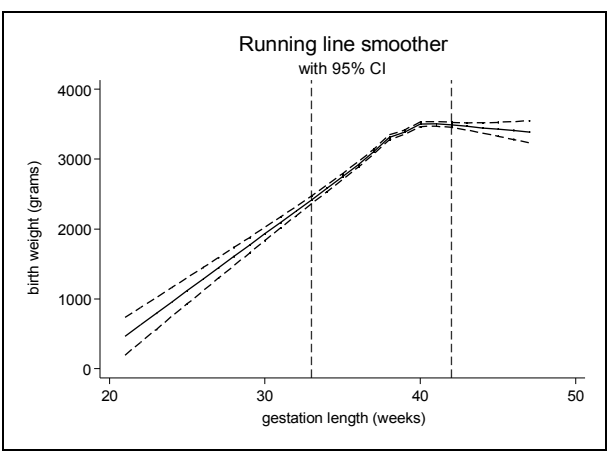

Fig. 15.5 Lowess smoothed line estimates of the relationship between birth weight and gestation length and its 95% CI

Note Vertical dashed lines mark the 2.5th and 97.5th percentiles of gestation length

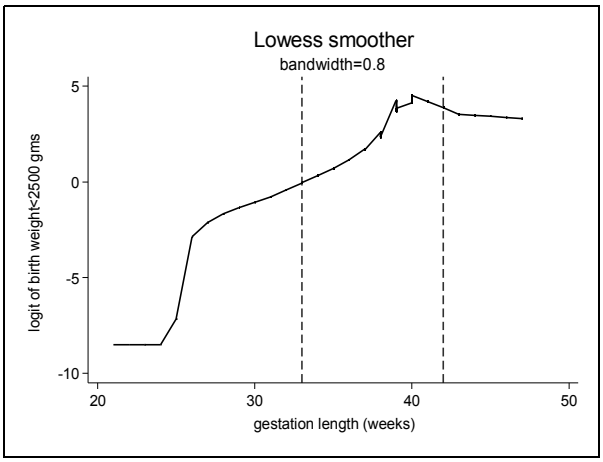

Fig. 15.6 Lowess smoothed curve estimates of the relationship between the logit of low birth weight (<2500 gm) and gestation length

Note Vertical dashed lines mark the 2.5th and 97.5th percentiles of gestation length

the full set of the data, not just the observations within the neighbourhood. One postulated advantage of global-influence models is that they may perform better on future data. Their disadvantage is that they are less sensitive to local disturbances in the data and hence localised effects may be overlooked. Caution must be used when interpreting results from polynomial models. They might be heavily influenced by points at the ends of the range of values for the predictor. It is also dangerous to make any predictions outside the range of observed values.

Quadratic models

The most common way to fit a curve (rather than a straight line) through the data is to add a **quadratic term** (the predictor squared, x^2). This fits a simple curve which bends in only one direction. The significance of the quadratic term can be used to check whether the assumption of linearity is acceptable (provided the data do not follow a more complex pattern than suggested by the simple curve of a quadratic model). One issue to keep in mind is that the original value is often highly correlated with its squared term and collinearity might be a problem in the model. The usual way to avoid this problem is to centre the original variable before squaring it. Example 15.3 shows that the quadratic term for gestation length is highly significant, indicating that a linear model is not appropriate. If a more complex curve is required to fit the data, a **cubic term** (x^3) can be added.

One way to ensure the new variables that are replacing the original variables are uncorrelated is to create **orthogonal polynomials**. These are variables that are constructed from the original data, but are on a new scale with each variable having a mean of 0 and possibly also a standard deviation (SD) of 1. The correlation between any pair of these variables is 0. These new variables can be used in place of the original ones in the regression model. Removal of the collinearity makes it possible to interpret the lower order terms, but the fact that they are not on the original scale makes this difficult (data not shown).

15.6.4 Fractional polynomials

While any set of variables might be orthogonalised, orthogonal polynomials are usually limited to power terms that have positive integer values (*eg* x^2 and x^3). One way of exploring more flexible functional forms is to use **fractional polynomials** (FP). FPs are power terms that can take on both positive and negative integer values and fractional values. The most common set of values to consider is -3, -2, -1, -0.5, 0, 0.5, 1, 2, and 3 (where the power 0 refers to a natural log transformation). The combination of FP that best fits the data (*ie* the model with the smallest log likelihood) can be determined. A 2-degree FP (*ie* 2 power terms selected—x^{p1} and

Example 15.3 Quadratic model
data = bw5k

A quadratic model regressing birth weight on gestation length was fit after the gestation length was centred by subtracting the mean gestation length (39 weeks). The significance of the quadratic term suggests that the quadratic model fits significantly better than a simple linear model (which is consistent with the smoothed line plots).

Source	SS	df	MS	
Model	467800274	2	233900137	Number of obs = 5000
Residual	1.1329e+09	4997	226722.309	F(2, 4997) = 1031.66
Total	1.6007e+09	4999	320210.372	Prob > F = 0.0000
				R-squared = 0.2922
				Adj R-squared = 0.2920
				Root MSE = 476.15

bwt	Coef	SE	t	P>t	95% CI	
gest_ct	94.483	3.577	26.42	0.000	87.471	101.495
gest_sq	-6364.501	448.432	-14.19	0.000	-7243.624	-5485.378
constant	3365.579	7.032	478.60	0.000	3351.793	3379.365

x^{p2}) can fit a wide range of shapes, and it is usual to use 2 terms or less. (**Note** A 2-degree FP may choose the same value for p^1 and p^2 in which case the 2 power terms are: x^{p1} and $x^{p1}\ln(x)$.)

The main advantage of FP models is that a 2-degree FP can fit a wide range of non-linear functions and may well be the most parsimonious way to obtain a good fit with the data. However, some issues which need to be kept in mind when using fractional polynomials are:

- FP can only be used with positive values of x, so an initial transformation of x may be required (if the software implementation does not do this automatically, or a particular scale is preferred).
- FP models use more df than an ordinary polynomial model (*eg* quadratic). For example, when comparing a quadratic model with a linear model, the difference is one df (required to estimate the second β). However, a 2-degree FP model uses 2 extra dfs compared with a 1-degree FP model, because the process involves estimating both the β for the second term as well as the second power value.
- Scaling the x variable may be required to make the FP estimation procedure robust (to avoid numerical overflow or underflow in the estimation procedure). This may or may not be done automatically by the software implementation.
- Very small values of x may induce artifacts into an FP model.

The coefficients derived from an FP are impossible to interpret in a meaningful way. The only way to make sense of such a model is to display the function graphically (which is a good idea whenever there is a non-linear function of x in a model). However, if you want to control for the effect of a factor (*ie* a potential confounder) in a regression model, then fitting fractional polynomials can be a useful approach. A much more thorough discussion of the use of FP in regression modelling can be found in Royston and Sauerbrei (2008).

Example 15.4 shows the fitting of fractional polynomials to the birth-weight data used in the previous example. The best fitting model is based on power terms of -0.5 and ln(x). The shape of the FP model along with cubic, quadratic, and linear models is shown in Fig. 15.7.

Example 15.4 Fractional polynomials
data = bw5k

Fractional polynomials (up to 2-degree) were fit to explore the nature of the relationship between birth weight and gestation length.

Number of obs = 5000
F(2, 4997) = 1074.38

Source	SS	df	MS	Prob > F = 0.0000
Model	481348094	2	240674047	R-squared = 0.3007
Residual	1.1194e+09	4997	224011.118	Adj R-squared = 0.3004
Total	1.6007e+09	4999	320210.372	Root MSE = 473.3

bwt	Coef	SE	t	P>t	95% CI	
igest_1	374.0814	15.17498	24.65	0.000	344.3318	403.831
igest_2	-208.1388	9.116145	-22.83	0.000	-226.0105	-190.2672
constant	3340.234	7.239147	461.41	0.000	3326.042	3354.426

Deviance: 75783.64. Best powers of gest among 44 models fit: 3 3.

Fractional polynomial model comparisons:

gest	df	Deviance	Res SD	Dev dif	P (*)	Powers
not in model	0	77572.053	565.871	1788.413	0.000	
linear	1	76041.390	485.607	257.751	0.000	1
m = 1	2	75911.085	479.321	127.446	0.000	-1
m = 2	4	75783.639	473.298	–	–	3 3

(*) P-value from deviance difference comparing reported model with m = 2 model

The best fitting 2-degree model is based on the powers gest^3 and gest^3*ln(gest). This fits the data significantly better than a one-degree model (power is -1). As expected, it also fits significantly better than either a linear model or a null model. The shape of the resulting function is shown in Fig. 15.7 along with the fit from cubic, quadratic, and linear models.

15.6.5 Splines

An alternative to fitting a polynomial model is to fit a piecewise linear function. Points at which the slope of the relationship is observed (or expected) to change (known as knot points) are identified and the relationship is assumed to be linear between these points. In the absence of any evidence for the selection of points, they may be chosen based on percentiles of the predictor. Fig. 15.8 shows a piecewise linear function, with knot points at the 25th, 50th, and 75th percentiles of birth weight vs gestation length.

One drawback to a piecewise linear function is that it is not usually biologically reasonable to expect sudden shifts in the nature of the relationship at the chosen knot points. Piecewise linear functions are also called linear splines. Generally, spline functions are pieced together by polynomials between the knots. Cubic splines allow more flexible shapes and smoother transitions across the knots than linear splines. Details of cubic splines are beyond the scope of this text, but an example of cubic splines fit is also shown in Fig. 15.8 (using the same knot points as for the linear splines).

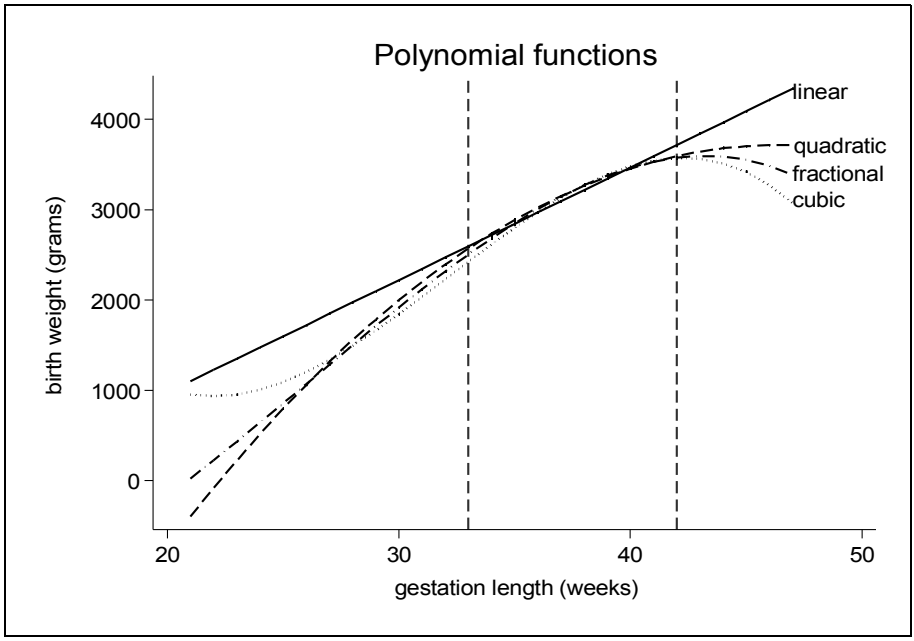

Fig. 15.7 Linear, quadratic, cubic, and fractional polynomial relationships between birth weight and gestation length

Note Vertical dashed lines mark the 2.5th and 97.5th percentiles of gestation length

One final comment about the selection of a functional form of a predictor is in order. In keeping with the idea that model-building should integrate subject matter knowledge with statistical considerations, it may not be appropriate to always use a 'best- fit' functional form that has been chosen based on the statistical significance of one form over another. In some situations (particularly with small datasets), there may not be sufficient evidence to conclude with certainty that a non-linear form is preferable to a linear form. However, if there are strong biological reasons to believe that a relationship is not likely to be linear, it may be appropriate to choose a polynomial function anyway. This is particularly important if the predictor is likely to be a strong confounder. In order to remove as much confounding effect as possible, it may be preferable to include a polynomial function of the predictor.

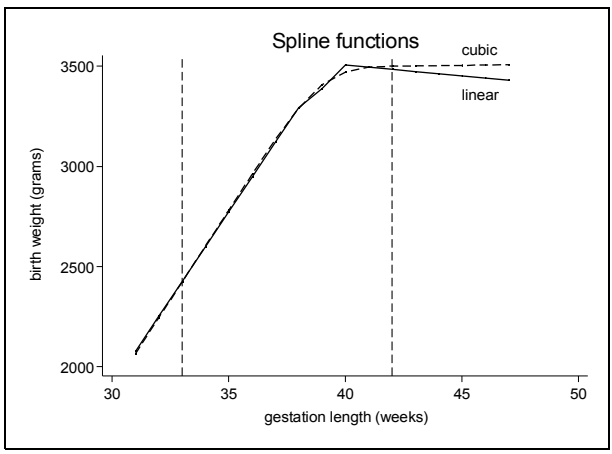

Fig. 15.8 Piecewise linear and cubic splines of relationship between birth weight and gestation length

Note Vertical dashed lines mark the 2.5th and 97.5th percentiles of gestation length. Plot limited to gestation lengths ≥30 weeks.

15.7 Identifying interaction terms of interest

It is important to consider including interaction terms when specifying the maximum model. There are 5 general strategies for creating and evaluating 2-way interactions.

1. Create and evaluate all possible 2-way interaction terms. This will only be feasible if the total number of predictors is small (*eg* ≤ 8).
2. Create 2-way interactions among all predictors that are significant in the final main effects model (once you have completed the initial model-building (Section 15.8)).
3. Create 2-way interactions among all predictors found to have a significant unconditional association with the outcome.
4. Create 2-way interactions only among pairs of variables which you suspect (based on evidence from the literature *etc*) might interact. This will probably focus on interactions involving the primary predictor(s) of interest and important confounders.
5. Only create 2-way interactions that involve the exposure variable (predictor) of interest.

Regardless of how the set of interaction terms is created, you could subject them to the same sort of screening processes described above to reduce the number included in the model-building process. If an interaction term is to be included in the model, then the main effects that make up that interaction term must also be included. Evaluation of a large number of 2-way interactions could identify spurious associations due to the fact that a large number of associations are being evaluated. In this case, some form of adjustment for the fact that multiple factors are being considered (*eg* Bonferroni adjustment) should be undertaken. Two-way interactions between continuous predictors are difficult to interpret, and, whenever significant, should be evaluated by fitting a range of possible values for both predictors with a graphical display of the results (see Example 14.11).

Three-way interactions might be considered, but they are usually difficult to interpret. They should be included only if there is good reason (*a priori*) to suspect the existence of such an effect, or if they are made up of variables with significant 2-way interactions. Three-way interactions might also unnecessarily complicate the model, because all of the main effects and 2-way interactions among the predictors making up the 3-way interaction need to be included.

15.8 Building the model

15.8.1 Specify the selection criteria

Once a maximum model has been specified, you need to decide how you will determine which predictors need to be retained in the model. Criteria for retention can be based on non-statistical considerations or the statistical significance of the predictor. It is essential that both be considered.

Non-statistical considerations

Variables should be retained in the model if they:

- are a primary predictor of interest
- are thought, *a priori*, to be confounders for the primary predictor of interest
- show evidence of being a confounder in this dataset because their removal results in a substantial change in the coefficient for one of the primary predictors of interest. **Note**

Building an appropriate causal model before starting the model-building process will help ensure that the variable is not an intervening variable (see Section 13.11.6)

- are a component of an interaction term which is included in the model.

Statistical criteria—nested models

Nested models are based on the same set of observations; the predictors in one model are a subset of the predictors in the other model. By far the most common approach to evaluating the statistical significance of individual predictors is to use tests based on nested models. For a linear regression model, this would involve carrying out a partial *F*-test for the predictor, while in other types of regression model (*eg* logistic, Poisson) a Wald test (see Section 16.7), a score test (not covered in this text) or likelihood-ratio test (*LRT*—see Section 16.6) can be used. Of these, the *LRT* has the best statistical properties (Royston and Sauerbrei, 2008) although the tests usually produce similar results. Consequently, the Wald test, which is often the most convenient, can be relied on unless the statistical significance of the predictor is questionable (*eg* P-value close to 0.05) or the estimated SE appears suspect (as may happen when estimation is difficult). When evaluating the significance of a categorical variable (included in the model as a set of indicator variables), the overall significance of all the indicator variables in the model should be used, not the statistical significance of individual indicator variables.

Statistical considerations—non-nested models

A number of **information criteria** (IC) have been developed for comparing models that are not nested. The general formula for these criteria is:

$$\text{IC} = -2\ \text{lnL} + a * s \qquad \textit{Eq 15.1}$$

where a is a penalty constant, s is the number of parameters in the model ($s=(k+1)$ for a linear regression model (where k is the number of predictors) and lnL is the log-likelihood (see Section 16.4).

The most commonly used information criteria are the **Akaike's Information Criteria** (AIC) which has $a=2$ and the **Bayesian Information Criteria** (BIC)—also known as the Schwartz Bayesian Criteria—which has $a=\log n$. They are based on an overall assessment of the model and can be used to compare different models, regardless of whether they are nested. They can be used to compare linear regression models and discrete data models (*eg* logistic, Poisson). However, some words of caution are in order. First, these statistics should not be used to compare nested models—test-based comparisons (*eg* partial *F*-tests or likelihood-ratio tests) are superior. Second, these statistics cannot be used to compare models which are based on different sets of observations. Finally, these criteria should not be used to compare models in which the likelihoods are computed in different ways (*eg* comparing a Cox semi-parametric survival model and a Weibull parametric model would not be appropriate—see Chapter 19).

The smaller the value of the IC, the better the model. If 2 models have comparable log likelihoods, the more parsimonious model (*ie* fewer parameters) will have the smaller IC. The BIC has an advantage that guidelines for assessing the evidence of superiority of one model over another are available (Table 15.1) (Raftery, 1996) (guidelines based on a Bayesian approach to statistics—see Chapter 24). However, the BIC tends to strongly favour more parsimonious models. It also suffers from the disadvantage that it depends on the value of n (number of observations), but it is not always clear what value of n should be used if the data are clustered (*ie* you do not have n independent units). (**Note** Several variations in the formula for the BIC exist in statistical programs. However, regardless of the formula used, the difference in the BIC between 2 models will be the same for each of the formulae.)

Table 15.1 Guidelines for interpreting BIC values from non-nested models

Absolute difference in BIC	Evidence for superiority of the better model
0–<2	Weak
2–<6	Positive
6–<10	Strong
≥10	Very strong

Two additional approaches, applicable to linear regression models, are based on the adjusted R^2 or a statistic called Mallows' C_p. The model which maximises the adjusted R^2 (see Section 14.3.5) is, in effect, maximising the amount of variance explained by the model, while precluding the incorporation of predictors which explain only a very small amount of the variance. This approach is equivalent to finding the model which minimises the mean square error (MSE). (**Note** Adding unimportant terms to the model will actually increase the MSE because the *df* on which it is based becomes smaller.)

Mallows' C_p is computed as follows (Mallows, 1973). If k predictors are selected from a complete set of p predictors, then Mallows' C_p for that model is:

$$Cp = \sum \frac{(Y - \hat{Y})^2}{\sigma^2} - n + 2k \qquad \textit{Eq 15.2}$$

where Y and $\hat{Y}$ are the observed and predicted values of Y for a model based on the k predictors, σ^2 is the MSE from a model based on all (p) predictors, and n is the sample size. Mallows' C_p is a special case of the AIC. Models with the lowest C_p are generally considered the best (Example 15.5).

15.8.2 Specifying the selection strategy

Once the criteria (both statistical and non-statistical) to be used in the selection process have been specified, there are a number of ways to carry out the selection.

All possible/best subset regressions
If the number of predictors in the maximum model is small, then it is possible to examine all possible combinations of predictors. Once all of the models have been fit, it is relatively easy to apply both the non-statistical and statistical criteria described above in order to identify some of the 'better' models. This approach is best applied in a context that a researcher is searching for a number of good models, such as early on in an investigation.

This process is modified slightly with best subset regression. In this procedure, the software identifies the 'best' model (according to one of the criteria outlined above), with a given number of predictors. For example, it will identify the single-term model with the largest R^2, the 2-term model with the largest R^2, the 3-term model with the largest R^2 *etc*. The investigator can then identify the point at which increasing the number of predictors in the model is of little value in terms of improving the predictive ability of the model. Both nested and non-nested models can be compared using 'all possible' or 'best subset' selection procedures.

Forward selection/backward elimination/stepwise
When a **forward selection** process is used, the computer first fits a model with only the intercept, and then selectively adds terms that meet a specified criterion. The usual criterion is

Example 15.5 Automated model selection for factors affecting birth weight
data = bw5k

Forward selection and backward elimination procedures were applied to the birth weight data using a selection threshold of P=0.05. The predictors (and coefficients) selected by each approach were as shown. (Forward or backward stepwise selection produced the same model as backward elimination.)

Description of predictor	Variable name	Forward selection	Backward elimination
gestation length	gest	111.184	111.685
multiple birth	multbrth	-605.514	-606.543
mother's age	mage	2.983	
male child	male	109.991	109.754
race = white	_Imrace_c4_2	49.959	50.841
race = black	_Imrace_c4_3	-120.192	-121.513
race = other	_Imrace_c4_4	-101.136	-98.563
cigarettes in 2nd trimester	cig_2	-15.777	-16.157
total birth order (parity)	tbo	22.291	24.755
father's age = 24–29	_Ifage_c4_2	47.705	60.286
father's age = 30–34	_Ifage_c4_3	79.056	103.460
father's age = 35+	_Ifage_c4_4	70.113	104.679
# of prenatal visits	previs	3.660	3.845
pregnancy hypertension	phyper	69.057	
constant	_cons	-1333.053	-1229.904
Model parameters			
	SS_{tot}	5.37E+008	5.36E+008
	√ MSE	461.940	462.160
	-2lnL	75528.640	75535.520
	R^2	0.336	0.335
	adjusted R^2	0.334	0.333
	AIC	75558.650	75561.530
	BIC	75656.400	75646.250
	Cp	15.000	17.865

The two selection procedures produced slightly different models with forward selection identifying 2 predictors (pregnancy hypertension and mother's age) which were not included in the backward elimination. Mother's age is not significant if father's age is in the model, but is added prior to father's age in the forward selection procedure. Pregnancy hypertension is borderline significant (P≈0.05) and whether the P-value falls just below or above the cutpoint of 0.05 depends on whether mother's age is in the model. Because the models are quite similar, the root MSE and R^2 are all quite close. The AIC (and Mallows' C_p) would suggest that the larger (forward selection) model is superior while the BIC favours the smaller model (highlighting the BIC's predilection for parsimony).

Note This example is provided for pedagogical purposes only, not as a recommended approach to model-building.

having the largest Wald test statistic, provided it exceeds the value required to produce a P-value below a specified value (such as 0.05). The term with the largest Wald test statistic is added first and then the process is repeated. This continues until no term meets the entry criterion.

With **backward elimination**, the process is reversed. The maximum model is fit and then terms are removed sequentially until none of the terms remaining in the model has a Wald test statistic meeting the specified criterion. An advantage of backward elimination is that the statistical significance of terms is assessed after adjustment for the potential confounding effect of other variables in the model. With forward selection, this happens to a much more limited extent (only after confounders have been selected and incorporated into the model).

Stepwise regression is simply a combination of forward selection and backward elimination. **Forward stepwise** starts with forward selection but after the addition of each variable, the criterion for backward elimination is applied to each variable in the model to see if it should remain. **Backward stepwise** starts with a full model and sequentially removes predictors but after the removal of each variable, all removed variables are checked to see if any of them would meet the forward selection criterion for inclusion.

In general, backward stepwise regression is favoured over forward stepwise (Mantel, 1970). However, forward stepwise may have to be used when there are a very large number of predictors or a large number of interaction terms are being considered. Backward stepwise with a P-value for variable removal of 0.157 has been suggested as a reasonable substitute for an all-subset procedure using Mallows' C_p or the AIC as a selection criterion (Sauerbrei *et al*, 2007).

In general, different selection procedures will often result in the same final model. However, in small datasets and those with large numbers of predictors, this may not be the case as can be seen in Example 15.5.

Caution in using any automated selection procedures

While the automated selection procedures described above are convenient, easy to apply and quickly reduce a large complex dataset to a succinct regression model, they must be applied judiciously, and should be considered methods of data exploration rather than definitive approaches to building a model. Some scientific journals will no longer accept regression models which have been built solely using automated selection criteria.

Some of the problems with automated model-building procedures are that they:

- yield R^2 values which are too high (see more on validation in Section 15.9)
- are based on methods (*eg* partial *F*-tests) which were designed to test specific hypotheses in the data (as opposed to evaluating all possible relationships) so they produce P-values which are too small and confidence intervals for parameters which are too narrow (more on this below)
- can have severe problems in the face of collinearity
- cannot incorporate any of the non-statistical considerations identified above
- make the predictive ability of the model look better than it really is
- do not differentiate between exposures, confounders and intervening variables, and
- waste a lot of paper.

However, the most serious drawback in their use is that they allow the investigator to avoid thinking about their data and the questions to be asked. By turning the model-building procedure over to an automated process, the investigator abdicates all responsibility for the

results of their analysis. Most seriously, the ability to evaluate the confounding effect of predictors which may not be statistically significant is lost. Avoiding this problem involves combining an assessment of the statistical significance of predictors with some form of change-in-estimate criterion (*ie* do estimates of other predictors change by a specified amount when the confounder is removed or added) (Rothman *et al*, 2008) (see also Chapter 13).

However, when faced with a large number of predictor variables, using a variety of automated selection procedures might be helpful in identifying all of the predictors which potentially have statistically significant associations with the outcome.

Three additional points must be kept in mind when using any automated procedure. First, groups of indicator variables formed by breaking down a categorical variable must all be added or removed together. Second, if any interaction term is included, the main effects of both variables that make up the interaction term must be kept in the model. Third, the analysis will be based only on those observations for which all variables are not missing. If there are many missing observations in the dataset, the data used to estimate the model might be a very small subset of the full dataset.

P-values and automated selection procedures
It is important to note that if you allow an automated selection procedure to sift through all of your predictors and select a significant group, the actual level of significance of the selected predictors is less than the level that you set (*eg* 0.05). For example, if you select 'significant' predictors from a list of 10 unrelated variables (with α=0.05), then the probability of finding at least one predictor significant due to chance alone is:

$$\alpha^* = 1-(1-0.05)^{10} = 0.40 \qquad \textit{Eq 15.3}$$

There is a 40% chance that at least one predictor will be significant, even if none of them has any association with the outcome. This value (40%) is called the **experiment-wise error rate**.

Comparing predictions from competing models
If 2 models with different predictors have comparable predictive ability, it may be useful to compare actual predicted values from the 2 models. One approach to this is to use the Bland and Altman limits of agreement methods described in Section 5.2.5 (treating the predicted values from the 2 models as the diagnostic test results) (Royston and Sauerbrei, 2008).

15.8.3 Conduct the analysis

Once the issues described in the preceding sections have been addressed, the analysis should be relatively straightforward. However, it is inevitably an iterative process. As models are built and evaluated, the investigator gains insight into the complex relationships that exist among the variables in the dataset which allow for more refined, and biologically reasonable models to be built. In the process, investigators must incorporate their biological knowledge of the system being studied along with the results of the statistical analyses.

15.9 Evaluate the reliability of the model

Evaluating any regression model is a 2-step process. The first step is to thoroughly evaluate the model using regression 'diagnostics' (*eg* evaluating the normality of residuals from a linear regression model). This assesses the **validity** of the model, and procedures for doing this are

described in each chapter dealing with specific model types. The second step is to evaluate the **reliability** of the model. That is, to address the question of 'how well will the model predict observations in subsequent samples?' **Note** The term reliability is used differently by various authors, but we will use it to describe how well the conclusions from a regression model are likely to perform in terms of future predictions (Kleinbaum *et al*, 2007). Simply reporting the R^2 of a linear model or computing the '% correctly classified' by a logistic model does not evaluate reliability as these estimates will always overstate the true reliability of the model.

The 2 most common approaches to assessing reliability are **split-sample** and **leave-one-out** analyses. A **split-sample analysis** involves dividing the data randomly into 2 groups. A regression model is built using the data from one of the 2 groups and the model is then applied to the second group to obtain predicted values for the remaining observations. For linear regression models, the correlation between the predicted and observed values in the second group is called the **cross-validation correlation**. The difference between the R^2 obtained from the analysis of the first group's data and the square of the cross-correlation validation correlation is called the **shrinkage on cross-validation**. If it is small (a subjective decision, although 0.1 is generally considered small), then the model is considered reliable. For non-linear regression models (*eg* logistic models), the same general approach can be used but some other measure of predictive ability (*eg* replace R^2 with % correctly classified) needs to be used to compare the 2 sets of results.

If only a small dataset is available, it might be desirable to put more than 50% of the observations in the first group (the one used to build the prediction model). Alternatively, a 10-fold cross-validation can be carried out in which the data are divided into 10 subsets, with 9 being used to estimate the model, and that model used to generate predicted values for the 10th subset. This process is repeated with each subset being left out of the model estimation procedure. Example 15.6 shows split-sample validation of a model based on the birth weight data.

A leave-one-out approach to validation is based on fitting the model many times, with one observation left out each time (until all have been omitted). The residuals for the omitted observations are summed to provide an estimate of the prediction error which can then be compared with the prediction error from the model based on all observations. If the 2 values are close, it suggests that the model will predict future observations well.

An alternative approach to split-sample validation involves building separate regression models for each of the 2 halves of the dataset, and subjectively comparing the regression coefficient. **Note** This can be done for any type of regression model. If the coefficients are substantially different in the 2 models, then the model is not reliable.

15.10 Presenting the results

The standard method of presenting results from a regression model is to present the coefficients (don't forget to include the intercept), their standard errors, and/or their confidence intervals. Assuming the observed effects are causal, the coefficients represent the change that would be expected in the outcome for a unit change in the predictor. For dichotomous predictors (or categorical variables that have been converted to a set of dichotomous predictors), the coefficient represents the effect of the factor being present compared with when it is absent. However, for continuous variables, assessing their impact is more difficult because they are all measured on different scales (and hence, a 'unit change' might represent either a small or large change in the predictor). Consequently, it is difficult to determine the magnitude of the impact

Example 15.6 Cross-validation correlation
data = bw5k

The final model evaluating the effects of several factors on birth weight from Chapter 14 was used as a basis for this evaluation. The model was built using half of the data and the reliability evaluated by determining the models predictive ability in the second half of the data. The procedure for randomly selecting observations for the split dataset resulted in 2,466 observations being used to build the model and this dataset resulted in an R^2 of 0.274. When this model was applied to the 2nd half of the data, the R^2 rose slightly to 0.277, suggesting that the model is reliable.

	All		Half	
Variable	**β**	**P**	**β**	**P**
white	76.84	0.000	88.46	0.000
college	25.22	0.083	-2.91	0.888
cig_2	-15.43	0.000	-15.97	0.000
gest	124.26	0.000	122.07	0.000
constant	-1552.32	0.000	-1450.81	0.000

While the overall reliability of the model was high, the coefficient (and P-value) for -college- is substantially changed when only half of the data are used, suggesting that this parameter may not be estimated reliably.

of each predictor on the outcome. In order to obtain a better understanding of the effect of a predictor, it would be helpful to have an idea of what constitutes a reasonable change in any predictor measured on a continuous scale. Two approaches to presenting results in order that the relative impact of different predictors can be compared are to:

- use standardised coefficients, or
- compute predicted effects as a continuous predictor changes over its interquartile range.

Each of these will be discussed briefly. However, before proceeding it should be noted that there is evidence that non-numerical presentation of study results may be preferable, depending on the target audience (Akl *et al*, 2007), but this type of presentation will not be considered further in this text.

15.10.1 Standardised coefficients

In linear regression models, standardised coefficients represent the effect on the (standardised) outcome that results from a change of 1 SD in the predictor. They can be computed by rescaling the coefficient by multiplying it by the ratio of the SD of the predictor to the SD of the outcome [$\beta^*=\beta(\sigma_x/\sigma_y)$]. In the past, they have not only been used to evaluate the relative magnitude of effects for various predictors in a model, but to compare results across studies. However, there are 2 problems with this approach. First, the SD might not be a good measure of the variability of a continuous predictor variable. If the distribution is skewed to the right, a few large values might unduly inflate the estimate of the SD. More importantly, the SD of the predictor or the outcome might vary from population to population. If standardised coefficients are used to compare results across studies, identical results from 2 studies can appear different due to differences in the scaling factor. Consequently, standardised coefficients are no longer recommended for general use.

15.10.2 Interquartile ranges

The effect of a predictor can be represented by computing the change in the outcome that would be expected to accompany a change in the predictor across its interquartile range (IQR) (*ie* from its 25th to 75th percentile). This avoids the problem of outlying observations having a big impact on the standard deviation. Although the IQR might also vary across populations (as the SD does), the problem of comparability across studies can be avoided by supplementing the ordinary coefficients with the estimates of effect based on the IQR, rather than replacing the ordinary coefficients with standardised ones. Example 15.7 shows the effects of the 4 predictors used in Example 14.13.

15.10.3 Predictors eliminated from a model

When presenting results from a multivariable model, you might also want to discuss the potential effects of predictors not included in the model. Unless the P-value is very large, it is unwise to assume that the effect is zero. Some investigators will discuss unconditional associations between those predictors and the outcome. An alternative, if a backward elimination procedure has been used in the model-building process, is to use the coefficient of the predictor at the last step before it was removed from the model. A third approach is to force the predictor back into the final model and use its coefficient from that model as an estimate of its effect (adjusted for other predictors in the model).

15.10.4 Scale of results

In linear regression models, transformation of the outcome is often necessary to ensure that the assumptions underlying the model are satisfied. However, this makes the interpretation of the

Example 15.7 Effects of predictors
data = bw5k

Based on the final model evaluating the effects of several factors on birth weight from Chapter 14 , the effects of the various predictors was evaluated by computing the expected change in birth weight for each of the predictors.

Variable	Coef	Basis	Estimated effect change	Effect
white	76.84	dichotomous	0–1	76.84
college	25.22	dichotomous	0–1	25.22
cig_2	-15.43	arbitrary range	0–10	-154.30
gest	124.26	IQR	38–40	248.53

Of the 2 dichotomous predictors (-white- and -college-), being white had roughly 3 times the effect on birth weight as a college education did. As expected, a change in gestation length over the IQR had a large effect on birth weight. Because both the 25th and 75th percentiles of -cig_2- were 0, a moderate level of cigarette consumption (median consumption among smokers = 10 cigarettes per day) was used to evaluate the effect. The large negative effect of smoking is evident. **Note** It must be remembered that -gest- is an intervening variable so the effect of smoking is the direct effect (*ie* comparing two gestations of the same length). If smoking also effects gestation length, its total effect will be different than the effect shown above.

results more difficult and it is usually desirable to present results on a different scale than was used in the analysis. Back-transformations following linear regressions are discussed in Section 14.9.6. Converting results from the logit scale to the probability scale after logistic regression is discussed in Section 16.8.5. This issue was not a problem in Example 15.7 because the model was fit using an outcome (birth weight) on its original scale (grams).

REFERENCES

Akl EA, Maroun N, Guyatt G, Oxman AD, Alonso-Coello P, Vist GE, et al. Symbols were superior to numbers for presenting strength of recommendations to health care consumers: a randomized trial. J Clin Epidemiol. 2007;60(12):1298-305.

Chatfield C. Confessions of a pragmatic statistician. The Statistician. 2002;51:1-20.

Cleveland WS. Robust locally weighted regression and smoothing scatterplots. J Am Stat Assoc. 1979;74:829-36.

Cochran WG. The effectiveness of adjustment by subclassification in removing bias in observational studies. Biometrics. 1968;24(2):295-313.

Dohoo IR, Ducrot C, Fourichon C, Donald A, Hurnik D. An overview of techniques for dealing with large numbers of independent variables in epidemiologic studies. Prev Vet Med. 1997;29(3):221-39.

Donders ART, van der Heijden GJMG, Stijnen T, Moons KGM. Review: a gentle introduction to imputation of missing values. J Clin Epidemiol. 2006;59(10):1087-91.

Dukes K. Cronbach's alpha. Encyclopedia of Biostatistics, 2nd Ed. New York: J Wiley & Sons; 2007.

Fitzmaurice GM, Laird NM, Ware, JH. Applied Longitudinal Analysis. New York: Wiley; 2004.

Hamilton M. A rating scale for depression. J Neurol Neurosurg Psychiatry. 1960 Feb;23:56-62.

Harel O, Zhou X-H. Multiple imputation: review of theory, implementation and software. Stat Med. 2007;26(16):3057-77.

Khader YS, Batieha A, Jaddou H, Batieha Z, El-Khateeb M, Ajlouni K. Factor analysis of cardiometabolic risk factors clustering in children and adolescents. Metab Syndr Relat Disord. 2011 Apr;9(2):151-6.

Kleinbaum DG, Kupper LL, Mullen KE. Applied regression analysis and other multivariable models. 4th Ed. Pacific Grove, CA: Duxbury Press; 2007.

Little RJ. Missing data. Encyclopedia of Biostatistics, 2nd Ed. New York: J Wiley & Sons; 2007.

Little RJA, Rubin DB. Statistical Analysis with Missing Data. New York: Wiley; 2002.

Mallows C_p. Some comments on C_p. Technometrics. 1973;15:661-75.

Mantel N. Why stepdown procedures in variable selection? Technometrics. 1970;12:621-5.

McKinley RK, Manku-Scott T, Hastings AM, French DP, Baker R. Reliability and validity of a new measure of patient satisfaction with out of hours primary medical care in the United Kingdom: development of a patient questionnaire. Br Med J. 1997 Jan 18;314(7075):193-8.

Moons KGM, Donders RART, Stijnen T, Harrell FE, Jr. Using the outcome for imputation of missing predictor values was preferred. J Clin Epidemiol. 2006;59(10):1092-101.

Polgreen PM, Chen Y, Beekmann S, Srinivasan A, Neill MA, Gay T, et al. Elements of influenza vaccination programs that predict higher vaccination rates: results of an emerging infections network survey. Clin Infect Dis. 2008 Jan 1;46(1):14-9.

Raftery AE. Bayesian model selection in social research. Sociological Methodology. Oxford: Basil Blackwell; 1996. p. 111-63.

Rothman KJ, Greenland S, Lash TL. Modern Epidemiology, 3rd Ed. Philadelphia: Lippincott Williams & Wilkins; 2008.

Royston P, Altman DG, Sauerbrei W. Dichotomizing continuous predictors in multiple regression: a bad idea. Stat Med. 2006;25(1):127-41.

Royston P, Sauerbrei W. Multivariable model-building. A pragmatic approach to regression analysis based on fractional polynomials for modelling continuous variables. Chichester, England: John Wiley & Sons, Ltd; 2008.

Rubin DB. Multiple Imputation for Nonresponse in Surveys. New york: Wiley; 2004.

Sauerbrei W, Royston P, Binder H. Selection of important variables and determination of functional form for continuous predictors in multivariable model building. Stat Med. 2007;26(30):5512-28.

Vach W. Multiple Imputation for Nonresponse in Surveys. New York: Springer; 1994.

Vach W, Blettner M. Missing data in epidemiological studies. Encyclopedia of Biostatistics, 2nd Ed. New York: J Wiley & Sons; 2007.