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. In this case, the details of the 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. In this strategy, careful attention must be paid to possible 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 guiding light, 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 a categorical 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 some potentially important predictors. However, on the other, 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.

Each of these will be discussed below.

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 retained placenta on reproductive performance (as measured by the calving-to-conception interval) in multiparous dairy cows and also had recorded data on: parity, twins, dystocia, vaginal discharge (an indicator of metritis), calving to first service interval and first service conception, then a putative causal diagram might look like Fig. 15.1. (These are the variables recorded in the dataset daisy2 used in Chapter 14.) **Note** In the causal diagram, it is assumed that twins and dystocia only affect reproductive performance through their effects on retained placenta and/or vaginal discharge. (You might propose other causal relationships).

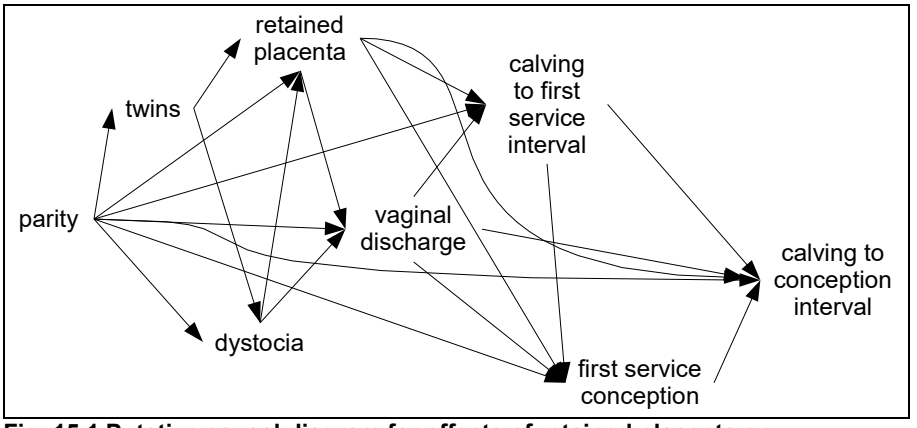

Fig. 15.1 Putative causal diagram for effects of retained placenta on reproductive performance

If the objective of the study was to quantify the effects of retained placenta on the calving to conception interval, you would NOT include any intervening variables (vaginal discharge, calving to first service interval, first service conception) in the regression model. Inclusion of these intervening variables would remove any of the effect from retained placenta that was mediated through them. On the other hand, if parity is suspected to be an important confounder, it might be designated to remain in the model regardless of whether or not it is statistically significant.

Even if a study has a very 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 farm management variables together, all measures of disease levels 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 procedure. 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 very precarious.

There are a variety of ways of reducing 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 essential 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 very 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 animals 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, an index representing the level of hygiene in stalls for dairy cows might be created as a linear combination of scores for factors such as quantity of bedding present, wetness of the bedding, amount of manure present and amount of fecal soiling of the udder and flanks of the cows. The weights 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 barn size might be used to compute the number of air changes per hour in a swine barn. 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 the 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 (unsuccessful) use of Cronbach's alpha for evaluating a scale to represent animal density in pig barns. Cronbach's alpha has also been used to demonstrate the internal consistency of a scale for dairy cow hygiene measures (Mounchili *et al*, 2004).

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-value (*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.

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

Example 15.1 Cronbach's alpha
data = pig_farm

Four variables related to the density of pigs in a pig barn were considered for inclusion in a scale reflecting animal density.

strdnst—floor space per starter pig (m^2)
grwdnst—floor space per grower pig (m^2)
fnrdnst—floor space per finisher pig (m^2)
shipm2—pigs shipped per m^2 in the barn

Because the items are measured on different scales, Cronbach's alpha was constructed as the mean of 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]
strdnst	69	+	0.7161	0.4962	0.5710	0.7997
grwdnst	69	+	0.8666	0.7403	0.4128	0.6783
fnrdnst	69	+	0.8389	0.6919	0.4420	0.7038
shipm2	69	-	0.7315	0.5193	0.5548	0.7889
Test scale					0.4951	0.7969

[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 reliability coefficient is 0.797 indicating respectable reliability (estimated correlation between the scale and underlying characteristic is $\sqrt{0.797}=0.89$). While the correlations between individual items and the scale are reasonable (0.72–0.87), with only 4 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 -strdnst- as the item with the lowest correlation to other items. The average correlation among the other items is also highest if -strdnst- is omitted (0.571).

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, then 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. For example, Lofstedt *et al* (1999), when evaluating a wide range of possible predictors of septicemia in diarrheic calves, built separate models using demographic and physical examination data, clinical chemistry data and hematology data. The important predictors from each of these 3 models were then evaluated in an overall model.

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 identification 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, Berghaus *et al* (2005) used factor analysis to evaluate interrelationships among variables collected as part of a risk assessment for Johne's disease. 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 common environment for weaned calves and cows). If this assumption is valid, then knowing which of those underlying structures are associated with the outcome (*eg* Johne's disease) might be as important as information about individual predictor variables. Determining which of the original predictors are important determinants of the outcome is a subjective process based on determining which predictors 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 among 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 among a set of categorical variables (both predictors and the outcome). The 2 axes are factorial axes which 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 elevated bacterial counts in bulk tank milk is presented in Example 15.2.

Example 15.2 Correspondence analysis of risk factors for elevated total aerobic bacterial counts in bulk tank milk
data = tac_mca

In a study aimed at identifying risk factors for elevated total aerobic bacterial counts (TAC) in bulk tank milk, a large number of factors related to cow hygiene and milking system function were evaluated (Elmoslemany *et al*, 2009a; Elmoslemany *et al*, 2009b). All factors were dichotomised and coded so that the presence of the factor was a risk factor (*eg* X1(cow hygiene) was coded so that herds with clean cows were coded 0 and those with dirty cows coded 1). Factors identified as having unconditional significant associations with being a case herd (*ie* elevated TAC) were as follows.

Factor	Description
X1	predipping of teats (risk factor = not predipping)
X2	clipping udder hair (risk factor = not clipping)
X3	washing vs dry-wiping of udders (risk factor = washing)
X4	teat end cleanliness (risk factor = dirty)
X5	cow (udder flank and leg) hygiene (risk factor = dirty)
X6	alkalinity of pipeline wash water (risk factor = high alkalinity)

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

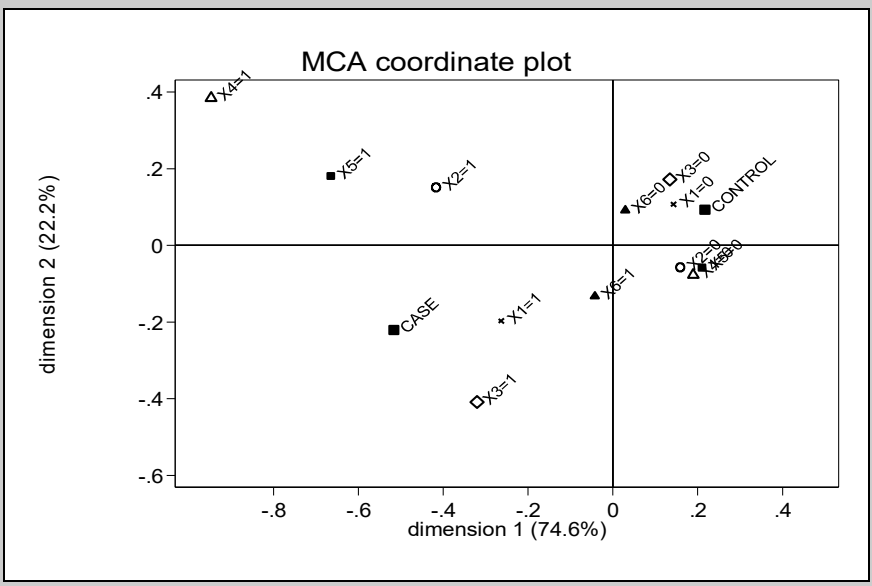

Fig. 15.2 Multiple correspondence analysis of risk factors for milk bacteria counts (case-control)

Absence of risk factors tended to cluster quite closely and were strongly associated with being a control herd (*ie* control herds did not have any of the risk factors evaluated). On the other hand, the presence of risk factors was more diffuse and not tightly clustered around the case herds, suggesting that case herds did not necessarily have all of the risk factors.

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 spilt and the results of that test consequently being missing). It could be said that the event that a particular value becomes missing is to be likened by 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 & Rubin, 2002), but for missingness among the predictors only this also holds true under less restrictive assumptions (Little, 1992; Vach & Blettner, 2007).

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**, that is, 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 & 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 & 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 the standard statistical text on missing data (Little & Rubin, 2002).

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 milk production in the first 120 days of

lactation (a measure of early lactation milk production) and the calving to first-service interval from the daisy2 data. They are only 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:

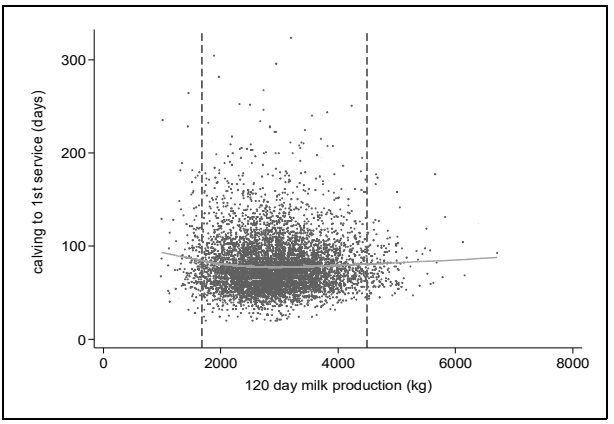

Fig. 15.3 Scatterplot of 120-day milk production and the calving to first service interval with lowess smoothed curve added

Note Vertical dashed lines mark the 2.5th and 97.5th percentiles of 120-day milk production

- 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**)—weights 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 go 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 calving to first service interval vs 120-day milk production. Fig. 15.4 shows running mean, running line and lowess smoothed lines for the same data.

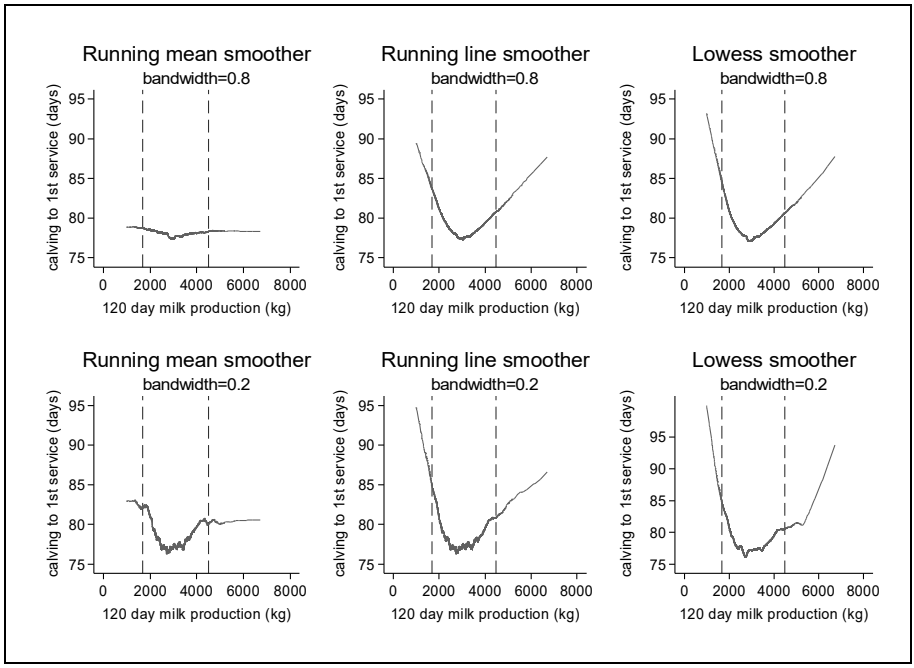

Fig. 15.4 Smoothed-line estimates of the relationship between 120-day milk production and the calving to first service interval

Note Vertical dashed lines mark the 2.5th and 97.5th percentiles of 120-day milk production.

Note All smoothed-line functions can have problems reliably portraying the data at the extreme values of the distribution because the neighbourhood is not symmetric 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. They do this 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 120-day milk production and the log odds of a cow conceiving at first service (relationship appears approximately 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 insights 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 categorisation of a continuous variable is done, it has been suggested that 5 categories will usually suffice to adequately control the confounding effects of that variable (Cochran, 1968). A model containing a categorised variable can be compared to one with a continuous variable (linear effect) by comparing their AIC or BIC values (see Section 15.8.1)

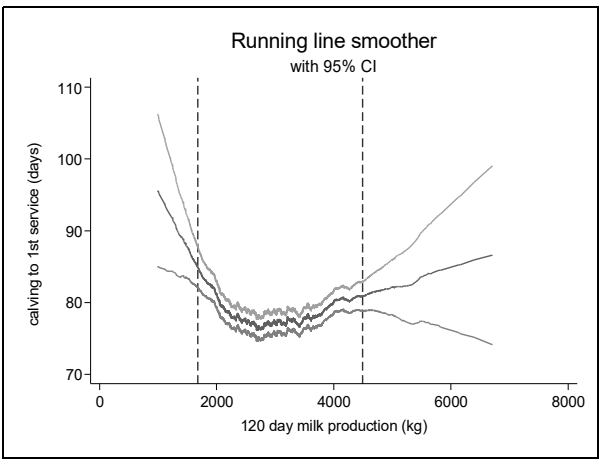

Fig. 15.5 Running-line smoothed estimates of relationship between 120-day milk production and calving to first-service interval and its 95% CI

Note Vertical dashed lines mark the 2.5th and 97.5th percentiles of 120-day milk production.

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 through the data rather than a straight line. 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 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 very dangerous to make any predictions outside the range of observed values.

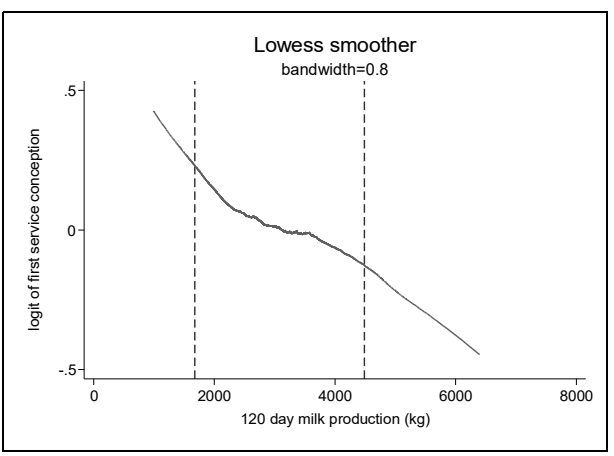

Fig. 15.6 Lowess smoothed-line estimates of the relationship between 120-day milk production and the logit of first-service conception

Note Vertical dashed lines mark the 2.5th and 97.5th percentiles of 120-day milk production.

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 as a check of 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 120-day milk production 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 variable 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 variables 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. FP of degree-2 (*ie* 2 power terms selected—x^{p1} and 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

Example 15.3 Quadratic model
data = daisy2

A quadratic model regressing calving to first service interval(-cf-) on 120 day milk production was fit after the milk production variable was centred by subtracting the mean milk production for the period (~3000 kg). 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).

Number of obs = 7720
F(2, 7717) = 17.93
Prob > F = 0.0000
R-squared = 0.0046
Adj R-squared = 0.0044
Root MSE = 28.28

Source	SS	df	MS
Model	28681.5181	2	14340.7591
Residual	6171667.01	7717	14340.7591
Total	6200348.53	7719	803.258003

cf	Coef	SE	t	P>|t|	95% CI	
m120_ct	-.0009293	.000461	-2.02	0.044	-.0018329	-.0000257
m120_sq	.0253519	.0042909	5.91	0.000	.0169405	.0337633
_cons	76.96643	.3910863	196.80	0.000	76.1998	77.73307

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 as follows.

- 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 to a linear model, the difference is one *df* (required to estimate the second β). However, a 2-degree FP model uses 2 extra *df* 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 & Sauerbrei (2008).

Example 15.4 shows the fitting of fractional polynomials to the dairy 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.

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 25^{th}, 50^{th} and 75^{th} percentiles of 120-day milk production.

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).

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

Example 15.4 Fractional polynomials
data = daisy2

Fractional polynomials (up to degree-2) were fit to explore the nature of the relationship between 120-day milk production and the calving to first service interval.

Source	SS	df	MS	
Model	36587.3328	2	18293.6664	Number of obs = 7720
Residual	6163761.19	7717	798.725048	F(2, 7717) = 22.90
Total	6200348.53	7719	803.258003	Prob > F = 0.0000
				R-squared = 0.0059
				Adj R-squared = 0.0056
				Root MSE = 28.262

cf	Coef	SE	t	P>|t|	95% CI	
imilk_1	295.8967	46.01638	6.43	0.000	205.6922	386.1013
imilk_2	87.36896	14.07522	6.21	0.000	59.77771	114.9602
_cons	76.95351	.378534	203.29	0.000	76.21148	77.69554

Deviance: 73498.30. Best powers of milk120 among 44 models fit: -.5 0.

Fractional polynomial model comparisons:

milk120	df	Deviance	Res SD	Dev dif	P (*)	Powers
not in model	0	73543.989	28.3418	45.690	0.000	
linear	1	73543.037	28.3419	44.738	0.000	1
m = 1	2	73525.418	28.3096	27.118	0.000	-2
m = 2	4	73498.300	28.2617	–	–	-.5 and 0

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

The best fitting 2-degree model is based on the powers -0.5 and ln(x). This fits the data significantly better than a one-degree model (power is -2). As expected, it also fits significantly better than either a linear model and 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.

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.

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).

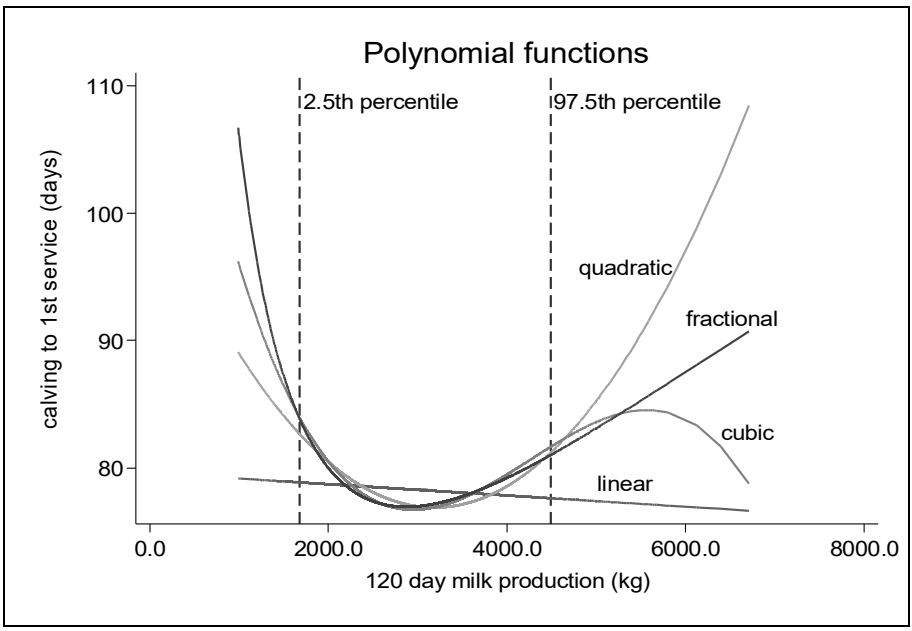

Fig. 15.7 Linear, quadratic, cubic and fractional polynomial relationships between 120-day milk production and calving to first- service interval

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

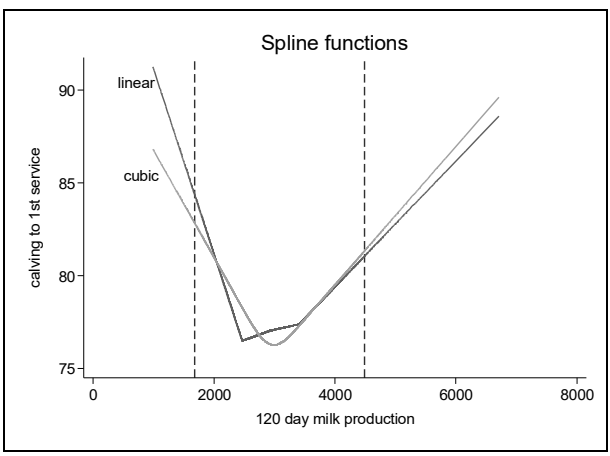

Fig. 15.8 Piecewise linear function and cubic splines of relationship between 120-day milk production and calving to first-service interval and its 95% CI

Note Vertical dashed lines mark the 2.5th and 97.5th percentiles of 120-day milk production.)

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 in the model.

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 and the non-statistical considerations will be discussed first.

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 13.12.6)
- are a component of an interaction term which is included in the model.

Statistical criteria—nested models
Nested models are models which are based on the same set of observations and in which 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, a score test (not covered in this text) or likelihood-ratio test (LRT—see Section 16.4) can be used. Of these, the LRT has the best statistical properties (Royston & 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** have been developed for comparing models that are not

nested. The general formula for these criteria is:

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

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 various 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 Mallow's Cp. 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.

Mallow's Cp is computed as follows (Mallows, 1973). If k predictors are selected from a complete set of p predictors, then Mallow's Cp 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. Mallow's Cp is a special case of the AIC. Models with the lowest Cp 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 in an investigation on a topic.

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 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 Mallow's *Cp* or the AIC as a selection criterion (Sauerbrei & Royston, 1999).

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.

Cautions 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 15.6)
- 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 only be based 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.

Example 15.5 Automated model selection for risk factors for pneumonia in swine
data = pig_farm

These data were obtained from a study evaluating the effects of various management factors on respiratory disease in swine (Hurnik *et al*, 1994a; Hurnik *et al*, 1994b). Starting with a full set of 43 predictors in this dataset and using the natural log of the prevalence of pneumonia (proportion of hogs with typical lung lesions at slaughter) as the outcome (n=66 observations), both forward and backward selection procedures were applied using a selection threshold of P=0.05. The predictors selected by each approach (and their coefficients) were:

Description of Predictor	Variable name	Forward selection	Backward elimination
Air inlet size (prop. of recommended)	inlet	-0.04	
Slow-growing pigs held back	hldbck	0.50	0.67
Herd size ('000)	size	0.43	0.67
Exhaust fan capacity (prop. of recommended)	exhaust	-0.37	-0.46
Producer's years of experience	exprnce	0.03	0.02
Slatted floor (vs solid)	floor		-0.51
Only home raised pigs in barn	hmrsd	-0.49	
	constant	-2.09	-2.62
Model parameters			
	SS_{tot}	62.9	62.9
	SSE	28.2	31.6
	√ MSE	0.69	0.73
	-2lnL	131.1	138.7
	adjusted R^2	0.51	0.46
	AIC	145.1	150.7
	BIC	160.4	163.8
	Cp	-11.1	-8.1

These data were introduced in Examples 15.3 and 15.4 and a full description of the dataset can be found in Chapter 31. The 2 procedures arrived at different final models, which was not surprising given the large number of variables relative to the number of observations in this dataset. The forward selection procedure has produced a superior model which explains more of the variation in the log-prevalence of pneumonia, has lower AIC and BIC scores and a lower Mallow's *Cp*. A best subset approach might be useful to identify a number of good models in situations such as this. However, variables that were selected in both procedures were consistent in their direction, although there were substantial differences in the coefficients. The model which gave the lowest Mallow's *Cp* (-11.7) was the same as the forward selection model except for one additional term (floor feeding). The model which maximised the adjusted R^2 (at a value of 0.59) contained 19 predictors and would have been totally unsuitable (results not shown).

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

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 group that are significant, 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 & 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 allows 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 a **split-sample** and **leave-one-out** analysis. 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. Split-sample validation of a model based on the daisy2 data is presented in Example 15.6.

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

Example 15.6 Cross-validation correlation
data = daisy2

The final model evaluating the effects of reproductive diseases on time to conception from Chapter 14 was used as a basis for this evaluation. (The outcome was square root transformed prior to analysis). 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 regression model was:

Number of obs = 775
F(9, 765) = 7.10
Prob > F = 0.0000
R-squared = 0.0771
Adj R-squared = 0.0662
Root MSE = 2.7338

Source	SS	df	MS
Model	477.42999	9	53.0477766
Residual	5717.36006	765	7.47367329
Total	6194.79005	774	8.00360472

wpc_sqrt	Coef	SE	t	P>|t|	95% CI	
hs_ct	.9891427	.1704134	5.80	0.000	.6546093	1.323676
hs_sq	.560026	.2461032	2.28	0.023	.0769083	1.043144
parity1	.1413507	.0678251	2.08	0.037	.0082053	.274496
calv_spr	-.7031682	.1992144	-3.53	0.000	-1.09424	-.3120964
twin	1.974691	.7838149	2.52	0.012	.4360079	3.513375
_ldyst_1	1.512377	.441814	3.42	0.001	.6450658	2.379689
vag_disch	.722121	.463742	1.56	0.120	-.188237	1.632479
ldysXvag~1	-2.444457	1.151216	-2.12	0.034	-4.704374	-.1845398
rp	-.030485	.3486269	-0.09	0.930	-.7148639	.6538939
_cons	7.503017	.2055855	36.50	0.000	7.099439	7.906596

The coefficient of determination (R^2) was reduced from 0.077 to 0.071 in the second half of the data. This represents minimal shrinkage, suggesting that the model is relatively reliable when applied to new datasets.

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 interval. 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 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 5 predictors used in Example 15.5 on the log-transformed prevalence of pneumonia in swine herds.

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 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.

Example 15.7 Effects of predictors
data = pig_farm

Based on the model selected using backward elimination (Example 15.5), the effects of the various predictors was evaluated by computing the expected change in the log-prevalence of pneumonia for defined changes in each of the predictors.

Variable	Coef	Basis	Estimated effect change	Effect
hldbck	0.666	dichotomous	0 - 1	0.666
size('000)	0.669	IQR	0.550 - 1.600	0.702
exhaust	-0.458	IQR	0.120 - 1.407	-0.589
exprnce	0.023	IQR	8.5 – 26.0	0.401
floor	-0.509	dichotomous	0 - 1	-0.509

It appears that herd size is one of the largest determinants of the prevalence of respiratory disease in this study population although all factors have comparable effects.

In Example 15.7, the effect of each predictor is assumed to be linear on the log scale, which is equivalent to having a multiplicative effect on the original scale. For example, holding back slow-growing pigs (β=0.666) increases the prevalence of pneumonia by a factor of 1.95 times ($e^{0.666}$=1.95). Consequently, the effect of holding back pigs will depend on the values of other factors in the model, because they will determine the prevalence of pneumonia that is multiplied by 1.95. It is often useful to compute the expected effects of key predictors on the original scale at various levels of other factors in the model.

REFERENCES

Akl EA, Maroun N, Guyatt G, Oxman AD, Alonso-Coello P, Vist GE, Devereaux PJ, Montori VM, Schünemann HJ. Symbols were superior to numbers for presenting strength of recommendations to health care consumers: a randomized trial J Clin Epidemiol. 2007; 60: 1298-305.

Berghaus RD, Lombard JE, Gardner IA, Farver TB. Factor analysis of a Johne's disease risk assessment questionnaire with evaluation of factor scores and a subset of original questions as predictors of observed clinical paratuberculosis Prev Vet Med. 2005; 72: 291-309.

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

Cleveland W. Robust locally weighted regression and smoothing scatterplots Journal of the 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: 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: 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: 1087-91.

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

Elmoslemany AM, Keefe GP, Dohoo IR, Jayarao BM. Risk factors for bacteriological quality of bulk tank milk in Prince Edward Island dairy herds. Part 2: bacteria count-specific risk factors J Dairy Sci. 2009a; 92: 2644-52.

Elmoslemany AM, Keefe GP, Dohoo IR, Jayarao BM. Risk factors for bacteriological quality of bulk tank milk in Prince Edward Island dairy herds. Part 1: overall risk factors J Dairy Sci. 2009b; 92: 2634-43.Fitzmaurice G, Laird N, J. Applied Longitudinal Analysis. Wiley: New York; 2004.

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

Hurnik D, Dohoo I, Bate L. Types of farm management as risk factors for swine respiratory disease Prev Vet Med. 1994a; 20: 147-57.

Hurnik D, Dohoo I, Donald A, Robinson N. Factor analysis of swine farm management practices on Prince Edward Island . Prev Vet Med. 1994b; 20: 135-46.

Kleinbaum D, Kupper L, Mullen K. Applied regression analysis and other multivariable models. 4th Ed. Duxbury Press: Pacific Grove; 2007.

Little R. Regression with missing Xs: A review J Am Stat Assoc. 1992; 87: 1227-37.

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

Little R, Rubin D. Statistical Analysis with Missing Data. Wiley: New York; 2002.

Lofstedt J, Dohoo IR, Duizer G. Model to predict septicemia in diarrheic calves J Vet Intern Med. 1999; 13: 81-8.

Mallows C. Some comments on *Cp*. Technometrics. 1973; 15: 661-75.

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

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

Mounchili A, Wichtel JJ, Dohoo IR, Keefe GP, Halliday LJ. Risk factors for milk off-flavours in dairy herds from Prince Edward Island, Canada. Prev Vet Med. 2004; 64: 133-45.

Raftery A. Bayesian model selection in social research. In: Sociological Methodology. Basil Blackwell: Oxford; 1996.

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

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

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

Rubin D. Multiple Imputation for Nonresponse in Surveys. Wiley: New York; 2004.

Sauerbrei W, Royston P. Building multivariable prognostic and diagnostic models: transformation of the predictors by using fractional polynomials J Royal Stat Soc. Series A. 1999; 162: 71-94.

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

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