CHAPTER 4

Model-Building Strategies and Methods for Logistic Regression

4.1 INTRODUCTION

In previous chapters we focused on estimating, testing, and interpreting the coefficients and fitted values from a logistic regression model. The examples discussed were characterized by having few independent variables, and there was perceived to be only one possible model. While there may be situations where this is the case, it is more typical that there are many independent variables that could potentially be included in the model. Hence, we need to develop a strategy and associated methods for handling these more complex situations.

The goal of any method is to select those variables that result in a "best" model within the scientific context of the problem. In order to achieve this goal we must have: (i) a basic plan for selecting the variables for the model and (ii) a set of methods for assessing the adequacy of the model both in terms of its individual variables and its overall performance. In this chapter and the next we discuss methods that address both of these areas.

The methods to be discussed in this chapter are not to be used as a substitute, but rather as an addition to clear and careful thought. Successful modeling of a complex data set is part science, part statistical methods, and part experience and common sense. It is our goal to provide the reader with a paradigm that, when applied thoughtfully, yields the best possible model within the constraints of the available data.

4.2 PURPOSEFUL SELECTION OF COVARIATES

The criteria for including a variable in a model may vary from one problem to the next and from one scientific discipline to another. The traditional approach to

Applied Logistic Regression, Third Edition.
David W. Hosmer, Jr., Stanley Lemeshow, and Rodney X. Sturdivant.

statistical model building involves seeking the most parsimonious model that still accurately reflects the true outcome experience of the data. The rationale for minimizing the number of variables in the model is that the resultant model is more likely to be numerically stable, and is more easily adopted for use. The more variables included in a model, the greater the estimated standard errors become, and the more dependent the model becomes on the observed data. Epidemiologic methodologists suggest including all clinically and intuitively relevant variables in the model, regardless of their "statistical significance." The rationale for this approach is to provide as complete control of confounding as possible within the given data set. This is based on the fact that it is possible for individual variables not to exhibit strong confounding, but when taken collectively, considerable confounding can be present in the data, see Rothman et al. (2008), Maldonado and Greenland (1993), Greenland (1989), and Miettinen (1976). The major problem with this approach is that the model may be "overfit," producing numerically unstable estimates. Overfitting is typically characterized by unrealistically large estimated coefficients and/or estimated standard errors. This may be especially troublesome in problems where the number of variables in the model is large relative to the number of subjects and/or when the overall proportion responding ($y = 1$) is close to either 0 or 1. In an excellent tutorial paper, Harrell et al. (1996) discuss overfitting along with other model building issues.

The following seven steps describe a method of selecting variables that we call purposeful selection. The rationale behind the method is that it follows the steps that many applied investigators employ when examining a set of data and then building a multivariable regression model.

Step 1: Purposeful selection begins with a careful univariable analysis of each independent variable. For categorical variables we suggest doing this via a standard contingency table analysis of the outcome ($y = 0,\ 1$) versus the k levels of the independent variable. The usual likelihood ratio chi-square test with $k - 1$ degrees of freedom is exactly equal to the value of the likelihood ratio test for the significance of the coefficients for the $k - 1$ design variables in a univariable logistic regression model that contains that single independent variable. Since the Pearson chi-square test is asymptotically equivalent to the likelihood ratio chi-square test, it may also be used. In addition to the overall test, it is a good idea, for those variables exhibiting at least a moderate level of association, to estimate the individual odds ratios (along with confidence limits) using one of the levels as the reference group.

Particular attention should be paid to any contingency table with a zero (frequency) cell, since in that situation, most standard logistic regression software packages will fail to converge and produce a point estimate for one of the odds ratios of either zero or infinity. An intermediate strategy for dealing with this problem is to collapse categories of the independent variable in some sensible fashion to eliminate the zero cell. If the covariate with the zero cell turns out to be statistically significant, we can revisit the problem at a later stage using one of the special programs discussed in Section 10.3. Fortunately, the zero cell problem does not occur too frequently.

For continuous variables, the best univariable analysis involves fitting a univariable logistic regression model to obtain the estimated coefficient, the estimated standard error, the likelihood ratio test for the significance of the coefficient, and the univariable Wald statistic. An alternative analysis, which is nearly equivalent at the univariable level and that may be preferred in an applied setting, is based on the two-sample t-test. Descriptive statistics available from this analysis generally include group means, standard deviations, the t statistic, and its p-value. The similarity of this approach to the logistic regression analysis follows from the fact that the univariable linear discriminant function estimate of the logistic regression coefficient is

$$\frac{(\overline{x}_1 - \overline{x}_0)}{s_p^2} = \frac{t}{s_p}\sqrt{\frac{1}{n_1} + \frac{1}{n_0}}$$

and that the linear discriminant function and the maximum likelihood estimate of the logistic regression coefficient are usually quite close when the independent variable is approximately normally distributed within each of the outcome groups, $y = 0,\ 1$, [see Halpern et al. (1971)]. Thus, the univariable analysis based on the t-test can be used to determine whether the variable should be included in the model since the p-value should be of the same order of magnitude as that of the Wald statistic, Score test, or likelihood ratio test from logistic regression.

Through the use of these univariable analyses we identify, as candidates for a first multivariable model, any variable whose univariable test has a p-value less than 0.25 along with all variables of known clinical importance.

Our recommendation for using a significance level as high as 0.20 or 0.25 as a screening criterion for initial variable selection is based on the work by Bendel and Afifi (1977) on linear regression and on the work by Mickey and Greenland (1989) on logistic regression. These authors show that use of a more traditional level (such as 0.05) often fails to identify variables known to be important. Use of the higher level has the disadvantage of including variables that are of questionable importance at this initial stage of model development. For this reason, it is important to review all variables added to a model critically before a decision is reached regarding the final model.

Step 2: Fit the multivariable model containing all covariates identified for inclusion at Step 1. Following the fit of this model, we assess the importance of each covariate using the p-value of its Wald statistic. Variables that do not contribute, at traditional levels of statistical significance, should be eliminated and a new model fit. The new, smaller, model should be compared to the old, larger, model using the partial likelihood ratio test. This is especially important if more than one term has been removed from the model, which is always the case when a categorical variable with more than two levels has been included using two or more design variables that appear to be not significant. Also, one must pay attention to make sure that the samples used to fit the larger and smaller models are the same. This becomes an issue when there are missing data. We discuss strategies for handling missing data in Section 10.4.

Step 3: Following the fit of the smaller, reduced model we compare the values of the estimated coefficients in the smaller model to their respective values from the larger model. In particular, we should be concerned about any variable whose coefficient has changed markedly in magnitude [e.g., having a value of $\Delta\hat{\beta} > 20\%$, see equation (3.9)]. This indicates that one or more of the excluded variables are important in the sense of providing a needed adjustment of the effect of the variables that remained in the model. Such variable(s) should be added back into the model. This process of deleting, refitting, and verifying continues, cycling through Step 2 and Step 3, until it appears that all of the important variables are included in the model and those excluded are clinically and/or statistically unimportant. In this process we recommend that one should proceed slowly by deleting only a few covariates at a time.

Step 4: Add each variable not selected in Step 1 to the model obtained at the conclusion of cycling through Step 2 and Step 3, one at a time, and check its significance either by the Wald statistic p-value or the partial likelihood ratio test, if it is a categorical variable with more than two levels. This step is vital for identifying variables that, by themselves, are not significantly related to the outcome but make an important contribution in the presence of other variables. We refer to the model at the end of Step 4 as the *preliminary main effects model*.

Step 5: Once we have obtained a model that we feel contains the essential variables, we examine more closely the variables in the model. The question of the appropriate categories for categorical variables should have been addressed during the univariable analysis in Step 1. For each continuous variable in this model we must check the assumption that the logit increases/decreases linearly as a function of the covariate. There are a number of techniques and methods to do this and we discuss them in Section 4.2.1. We refer to the model at the end of Step 5 as the *main effects model*.

Step 6: Once we have the main effects model, we check for interactions among the variables in the model. In any model, as discussed and illustrated with examples in Section 3.5, an interaction between two variables implies that the effect of each variable is not constant over levels of the other variable. As noted in Section 3.5, the final decision as to whether an interaction term should be included in a model should be based on statistical as well as practical considerations. Any interaction term in the model must make sense from a clinical perspective.

We address the clinical plausibility issue by creating a list of possible pairs of variables in the model that have some realistic possibility of interacting with each other. The interaction variables are created as the arithmetic product of the pairs of main effect variables. This can result in more than one interaction term. For example, the interaction of two categorical variables, each with three levels (i.e., two dummy variables), generates four interaction variables. We add the interactions, one at a time, to the main effects model from Step 5. (This may involve adding more than one term at a time to the

model.) We then assess the statistical significance of the interaction using a likelihood ratio test. Unlike main effects where we consider adjustment as well as significance, we only consider the statistical significance of interactions and as such, they must contribute to the model at traditional levels, such as 5% or even 1%. Inclusion of an interaction term in the model that is not significant typically just increases the estimated standard errors without much change in the point estimates of effect.

Following the univariable analysis of the interaction terms we add each interaction that was significant to the model at the end of Step 5. We then follow Step 2 to simplify the model, considering only the removal of the interaction terms, not any main effects. At this point we view the main effect terms as being "locked" and they cannot be removed from the model. One implication of "locking the main effects" is that we do not consider statistical adjustment, $\Delta\hat{\beta}\%$, when winnowing insignificant interactions.

We refer to the model at the conclusion of Step 6 as the *preliminary final model*.

Step 7: Before any model becomes the final model we must assess its adequacy and check its fit. We discuss these methods in Chapter 5. Note that regardless of what method is used to obtain a multivariable statistical model, purposeful selection or any of the other methods discussed in this chapter, one must perform Step 7 before using the fitted model for inferential purposes.

Bursac et al. (2008) studied the properties of purposeful selection compared to stepwise selection via simulations. The results showed that purposeful selection retained significant covariates and also included covariates that were confounders of other model covariates in a manner superior to stepwise selection.

As noted above, the issue of variable selection is made more complicated by different analytic philosophies as well as by different statistical methods. One school of thought argues for the inclusion of all scientifically relevant variables into the multivariable model regardless of the results of univariable analyses. In general, the appropriateness of the decision to begin the multivariable model with all possible variables depends on the overall sample size and the number in each outcome group relative to the total number of candidate variables. When the data are adequate to support such an analysis it may be useful to begin the multivariable modeling from this point. However, when the data are inadequate, this approach can produce a numerically unstable multivariable model, discussed in greater detail in Section 4.5. In this case the Wald statistics should not be used to select variables because of the unstable nature of the results. Instead, we should select a subset of variables based on results of the univariable analyses and refine the definition of "scientifically relevant."

Another approach to variable selection is to use a stepwise method in which variables are selected either for inclusion or exclusion from the model in a sequential fashion based solely on statistical criteria. There are two main versions of the stepwise procedure: (i) forward selection with a test for backward elimination and (ii) backward elimination followed by a test for forward selection. The algorithms

used to define these procedures in logistic regression are discussed in Section 4.3. The stepwise approach is useful and intuitively appealing in that it builds models in a sequential fashion and it allows for the examination of a collection of models that might not otherwise have been examined.

"Best subsets selection" is a selection method that has not been used extensively in logistic regression. With this procedure a number of models containing one, two, three variables, and so on, are examined to determine which are considered the "best" according to some specified criteria. Best subsets linear regression software has been available for a number of years. A parallel theory has been worked out for nonnormal errors models [Lawless and Singhal (1978, 1987a, 1987b)]. We show in Section 4.4 how logistic regression may be performed using any best subsets linear regression program.

Stepwise, best subsets, and other mechanical selection procedures have been criticized because they can yield a biologically implausible model [Greenland (1989)] and can select irrelevant, or noise, variables [Flack and Chang (1987); Griffiths and Pope (1987)]. They may also fail to select variables that narrowly fail to achieve the pre-designated threshold for inclusion into a model. The problem is not the fact that the computer can select such models, but rather that the judgment of the analyst is taken out of the process and, as a result, has no opportunity to scrutinize the resulting model carefully before the final, best model is reported. The wide availability and ease with which stepwise methods can be used has undoubtedly reduced some analysts to the role of assisting the computer in model selection rather than the more appropriate alternative. It is only when the analyst understands the strengths, and especially the limitations of the methods that these methods can serve as useful tools in the model-building process. The analyst, not the computer, is ultimately responsible for the review and evaluation of the model.

4.2.1 Methods to Examine the Scale of a Continuous Covariate in the Logit

An important step in refining the main effects model is to determine whether the model is linear in the logit for each continuous variable. In this section we discuss four methods to address this assumption: (i) smoothed scatter plots, (ii) design variables, (iii) fractional polynomials and (iv) spline functions.

As a first step, it is useful to begin checking linearity in the logit with a smoothed scatterplot. This plot is helpful, not only as a graphical assessment of linearity but also as a tool for identifying extreme (large or small) observations that could unduly influence the assessment of linearity when using fractional polynomials or spline functions. One simple and easily computed form of a smoothed scatterplot was illustrated in Figure 1.2 using the data in Table 1.2. Other more complicated methods that have greater precision are preferred at this stage.

Kay and Little (1986) illustrate the use of a method proposed by Copas (1983). This method requires computing a smoothed value for the response variable for each subject that is a weighted average of the values of the outcome variable over all subjects. The weight for each subject is a continuous decreasing function of the distance of the value of the covariate for the subject under consideration from the

value of the covariate for all other cases. For example, for covariate x for the ith subject we compute the smoothed value as

$$\overline{y}_{si} = \frac{\sum_{j=i_l}^{i_u} w(x_i, x_j) y_j}{\sum_{j=i_l}^{i_u} w(x_i, x_j)},$$

where $w(x_i, x_j)$ represents a particular weight function. For example, if we use STATA's scatterplot lowess smooth command, with the mean option and bandwidth k, then

$$w(x_i, x_j) = \left[1 - \left(\frac{|x_i - x_j|^3}{\Delta}\right)\right]^3,$$

where Δ is defined so that the maximum value for the weight is ≤ 1 and the two indices defining the summation, i_l and i_u, include the k percent of the n subjects with x values closest to x_i. Other weight functions are possible as well as additional smoothing using locally weighted least squares regression, which is actually the default in STATA.

In general, when using STATA, we use the default bandwidth of $k = 0.8$ and obtain the plot of the triplet $(x_i, y_i, \overline{y}_{si})$, that is, the observed and smoothed values of y on the same set of axes. The shape of the smoothed plot should provide some idea about the parametric relationship between the outcome and the covariate. Some packages, such as STATA's lowess command, provide the option of plotting the smoothed values, $(x_i, \overline{l}_{si})$ where $\overline{l}_{si} = \ln[\overline{y}_{si}/(1 - \overline{y}_{si})]$, that is, plotting on the logit scale, thus making it a little easier to make decisions about linearity in the logit. The advantage of the smoothed scatter plot is that, if it looks linear then the logit is likely linear in the covariate. One disadvantage of the smoothed scatter plot is that if it does not look linear, most of us lack the experience to guess, with any reliability, what function would satisfactorily reflect the displayed nonlinearity. The parametric approaches discussed below are useful here since they specify a best nonlinear transformation. Another disadvantage is that a smoothed scatterplot does not easily extend to multivariable models.

The second suggested method is one that is easily performed in all statistical packages and may be used with a multivariable model. The steps are as follows: (i) using the descriptive statistics capabilities of your statistical package, obtain the quartiles of the distribution of the continuous variable; (ii) create a categorical variable with four levels using three cutpoints based on the quartiles. We note that many other grouping strategies can be used but the one based on quartiles seems to work well in practice; (iii) fit the multivariable model replacing the continuous variable with the four-level categorical variable. To do this, one includes three design variables that use the lowest quartile as the reference group; (iv) following the fit of the model, plot the three estimated coefficients versus the midpoints

of the upper three quartiles. In addition, plot a coefficient equal to zero at the midpoint of the first quartile. To aid in the interpretation connect the four plotted points with straight lines. Visually inspect the plot. If it does not look linear then choose the most logical parametric shape(s) for the scale of the variable.

The next step is to refit the model using the possible parametric forms suggested by the plot and choose one that is significantly different from the linear model and makes clinical sense. It is possible that two or more different parameterizations of the covariate may yield similar results in the sense that they are significantly different from the linear model. However, it is our experience that one of the possible models will be more appealing clinically, thus yielding more easily interpreted parameter estimates.

The advantage of the first two methods is that they are graphical and easily performed. The disadvantage, as noted, is that it is sometimes difficult to postulate a parametric form from either a somewhat noisy plot (method 1) or from only four points (method 2).

The third method is an analytic approach based on the use of fractional polynomials as developed by Royston and Altman (1994). Since that key paper, Royston and colleagues have researched this method extensively and have written numerous papers providing guidance to applied investigators. For example, see Royston et al. (1999) and Sauerbrei and Royston (1999). The recent text on the method by Royston and Sauerbrei (2008) provides a detailed and highly readable account of the method along with its extensions and contains numerous numerical examples. Readers looking for more details are urged to consult this reference.

The essential idea is that we wish to determine what value of x^p yields the best model for the covariate. In theory, we could incorporate the power, p, as an additional parameter in the estimation procedure. However, this greatly increases the numerical complexity of the estimation problem. Royston and Altman (1994) propose replacing full maximum likelihood estimation of the power by a search through a small but reasonable set of possible values. The method is described in the second edition of this text, Hosmer and Lemeshow (2000) and Hosmer et al. (2008) provide a brief, but updated introduction to fractional polynomials when fitting a proportional hazards regression model. This material provides the basis for the discussion.

The method of fractional polynomials may be used with a multivariable logistic regression model, but for the sake of simplicity, we describe the procedure using a model with a single continuous covariate. The equation for a logit, that is linear in the covariate, is

$$g(x, \boldsymbol{\beta}) = \beta_0 + \beta_1 x,$$

where $\boldsymbol{\beta}$, in general, denotes the vector of model coefficients. One way to generalize this function is to specify it as

$$g(x, \boldsymbol{\beta}) = \beta_0 + \sum_{j=1}^{J} \beta_j \times F_j(x),$$

where the functions $F_j(x)$ are a particular type of power function. The value of the first function is $F_1(x) = x^{p_1}$. In theory, the power, p_1, could be any number, but in most applied settings it makes sense to try to use something simple. Royston and Altman (1994) propose restricting the power to be among those in the set $\wp = \{-2, -1, -0.5, 0, 0.5, 1, 2, 3\}$, where $p_1 = 0$ denotes the log of the variable. The remaining functions are defined as

$$F_j(x) = \begin{cases} x^{p_j}, & p_j \neq p_{j-1} \\ F_{j-1}(x)\ln(x), & p_j = p_{j-1} \end{cases}$$

for $j = 2, \ldots, J$ and restricting powers to those in $\wp$. For example, if we chose $J = 2$ with $p_1 = 0$ and $p_2 = -0.5$, then the logit is

$$g(x, \boldsymbol{\beta}) = \beta_0 + \beta_1 \ln(x) + \beta_2 \frac{1}{\sqrt{x}}.$$

As another example, if we chose $J = 2$ with $p_1 = 2$ and $p_2 = 2$, then the logit is

$$g(x, \boldsymbol{\beta}) = \beta_0 + \beta_1 x^2 + \beta_2 x^2 \ln(x).$$

The model is quadratic in x when $J = 2$ with $p_1 = 1$ and $p_2 = 2$. Again, we could allow the covariate to enter the model with any number of functions, J, but in most applied settings an adequate transformation is found if we use $J = 1$ or 2.

Implementation of the method requires, for $J = 1$, fitting 8 models, that is $p_1 \in \wp$. The best model is the one with the largest log-likelihood (or smallest deviance). The process is repeated with $J = 2$ by fitting the 36 models obtained from the distinct pairs of powers (i.e., $(p_1, p_2) \in \wp \times \wp$) and the best model is again the one with the largest log-likelihood (or smallest deviance).

The relevant question is whether either of the two best models is significantly better than the linear model. Let $L(1)$ denote the log-likelihood for the linear model (i.e., $J = 1$ and $p_1 = 1$) and let $L(p_1)$ denote the log-likelihood for the best $J = 1$ model and $L(p_1, p_2)$ denote the log-likelihood for the best $J = 2$ model. Royston and Altman (1994) and Ambler and Royston (2001) suggest, and verify with simulations, that each term in the fractional polynomial model contributes approximately 2 degrees of freedom to the model, effectively one for the power and one for the coefficient. Thus, the partial likelihood ratio test comparing the linear model to the best $J = 1$ model,

$$G(1, p_1) = -2\{L(1) - L(p_1)\},$$

is approximately distributed as chi-square with one degree of freedom under the null hypothesis that the logit is linear in x. The partial likelihood ratio test comparing the best $J = 1$ model to the best $J = 2$ model,

$$G[p_1, (p_1, p_2)] = -2\{L(p_1) - L(p_1, p_2)\},$$

is approximately distributed as chi-square with 2 degrees of freedom under the hypothesis that the $J = 2$ model is not significantly different from the $J = 1$ model.

Similarly, the partial likelihood ratio test comparing the linear model to the best $J = 2$ model is distributed approximately as chi-square with 3 degrees of freedom. (Note: to keep the notation simple, we use p_1 to denote the best power both when $J = 1$ and as the first of the two powers for $J = 2$. These are not likely to be the same numeric value in practice.)

In an applied setting we can use the partial likelihood ratio test in two ways to determine whether a transformation is significantly better than the linear model: a closed test and a sequential test [see Sauerbrei et al. (2006) and cited references]. We note that Sauerbrei, Meier-Hirmer, Benner, and Royston consider a model that does not contain x as the base model. We use the linear model as the base model since, at the end of step 3, we have eliminated all statistically nonsignificant or clinically unimportant covariates.

The closed test procedure begins by comparing the best two-term fractional polynomial model to the linear model using $G[1, (p_1, p_2)]$. If this test is not significant, at a typical level such as 0.05, then we stop and use the linear model. If the test is significant then the best two-term fractional polynomial model is compared to the best one-term fractional polynomial model using $G[p_1, (p_1, p_2)]$. If this test is significant then we select the two-term model; otherwise select the one-term model.

The sequential test procedure begins by comparing the best two-term fractional polynomial model to the best one-term fractional polynomial model using $G[p_1, (p_1, p_2)]$. If this test is significant we select the two-term model. If it is not significant then we compare the best one-term fractional polynomial model to the linear model using $G[1, (p_1, p_2)]$. If the test is significant then we select the best one-term model; otherwise we use the linear model.

Ambler and Royston (2001) examined the type I error rates of the two testing methods via simulations and concluded that the closed test is better than the sequential test at maintaining the overall error rate. Thus, we use the closed test method in this text.

Whenever a one or two-term model is selected we highly recommend that the resulting functional form be critically examined for subject matter plausibility. The best way to do this is by plotting the fitted model versus the covariate. We explain how to do this and illustrate it with the examples later in this chapter. One should always ask the obvious question: Does the functional form of the fractional polynomial transformation make sense within the context of the study? If it really does not make sense then we suggest using the linear model or possibly another fractional polynomial model. In almost every example we have encountered, where one of the two best fractional polynomial models is better than the linear model there is another fractional polynomial model that is also better whose deviance is trivially larger than the selected best model. This other model may provide a more clinically acceptable transformation. For example, assume that the closed test procedure selects the two-term model with powers $(2, 3)$. This transformation may have a deviance that is not much smaller than that of the two-term quadratic model $(1, 2)$. From a subject matter perspective the quadratic model may make more sense and be more easily explained than the best model. In this case we would not hesitate to use the quadratic model.

The only software package that has fully implemented the method of fractional polynomials within the distributed package is STATA. In addition to the method described above, STATA's fractional polynomial routine offers the user considerable flexibility in expanding the set of powers, $\wp$, searched; however, in most settings the default set of values should be more than adequate. STATA's implementation also includes valuable graphical displays of the transformed model. Sauerbrei et al. (2006) provide links to obtain macros for SAS and R code that can be used to perform all the fractional polynomial analyses done with STATA in this text.

So far the discussion of fractional polynomials has been in the setting of a simple univariable logistic regression model. In practice, most models are multivariable and can contain numerous continuous covariates, each of which must be checked for linearity. The approach we described above, where we checked for linearity one variable at a time, is the one we use in Step 5 of purposeful selection.

Royston and Ambler (1998, 1999) extended the original fractional polynomial software to incorporate an iterative examination for scale with multivariable models. The default method incorporates recommendations discussed in detail in Sauerbrei and Royston (1999). Multivariable modeling using fractional polynomials is available in distributed STATA and can be performed in SAS and R using the macros and code that can be obtained from links in Sauerbrei et al. (2006). We describe model building using multivariable fractional polynomials in Section 4.3.3.

We have found, in our practice, a level of reluctance by applied investigators to use fractional polynomial transformations, regardless of how much clinical sense they might make, because they think the model is too complicated to estimate odds ratios. We showed in Section 3.5 that by carefully following the four-step procedure for estimating odds ratios, one is able to obtain the correct expression involving the model coefficients to estimate any odds ratio, no matter how complicated the model might be.

The fourth method of checking for linearity in the logit is via spline functions. Spline functions have been used in statistical applications to model nonlinear functions for a long time; well before the advent of computers and modern statistical software brought computer intensive methods to the desk top [see, for example, Poirier (1973), who cites pioneering work on these functions by Schoenberg (1946)]. Harrell (2001, pp. 18–24) presents a concise mathematical treatment of the spline function methods we discuss in this section. Royston and Sauerbrei (2008, Chapter 9) compare spline functions to fractional polynomials.

The basic idea behind spline functions is to mathematically mimic the use of the draftsman's spline to fit a series of smooth curves that are joined at specified points, called "knots". In this section we consider linear and restricted cubic spines as these are the ones commonly available in statistical packages (e.g., STATA and SAS).

We begin our discussion by considering linear splines based on three knots. We discuss how to choose the number of knots and where these knots should be placed shortly. The linear spline variables used in the fit can be parameterized with coefficients representing the slope in each interval, or alternatively, by the slope in the first interval and the change in the slope from the previous interval. We use the

former parameterization, in which case the definitions of the four spline variables formed from three knots are as follows:

$$x_1 = \min(X, k_1)$$

and

$$x_j = \max[\min(X, k_j), k_{j-1}] - k_{j-1},\ j = 2, \ldots, 4$$

where k_1, k_2 and k_3 are the three knots. The four linear spline variables used in the fit are as follows:

$$x_{l1} = \begin{cases} X, & \text{if } X < k_1, \\ k_1, & \text{if } k_1 \le X, \end{cases}$$

$$x_{l2} = \begin{cases} 0, & \text{if } X < k_1, \\ X - k_1, & \text{if } k_1 \le X < k_2, \\ k_2 - k_1, & \text{if } k_2 \le X, \end{cases}$$

$$x_{l3} = \begin{cases} 0, & \text{if } X < k_2, \\ X - k_2, & \text{if } k_2 \le X < k_3, \\ k_3 - k_2, & \text{if } k_3 \le X, \end{cases}$$

$$x_{l4} = \begin{cases} 0, & \text{if } X < k_3, \\ X - k_3, & \text{if } k_3 \le X, \end{cases}$$

where the subscript "l" stands for linear spline.

The equation of the logit is

$$g(\mathbf{x}_l, \boldsymbol{\beta}_l) = \beta_{l0} + \beta_{l1}x_{l1} + \beta_{l2}x_{l2} + \beta_{l3}x_{l3} + \beta_{l4}x_{l4}. \tag{4.1}$$

Under the model in equation (4.1) the equation of the logit in the four intervals defined by the three knots is as follows:

$$g(\mathbf{x}_l, \boldsymbol{\beta}_l)$$

$$= \begin{cases} \beta_{l0} + \beta_{l1}X & \text{if } X < k_1, \\ \beta_{l0} + \beta_{l1}k_1 + \beta_{l2}\left(X - k_1\right) & \\ \qquad = [\beta_{l0} + \beta_{l1}k_1 - \beta_{l2}k_2] + \beta_{l2}X & \text{if } k_1 \le X < k_2, \\ \beta_{l0} + \beta_{l1}k_1 + \beta_{l2}(k_2 - k_1) + \beta_{l3}(X - k_3) & \\ \qquad = [\beta_{l0} + \beta_{l1}k_1 + \beta_{l2}(k_2 - k_1) - \beta_{l3}k_3] + \beta_{l3}X & \text{if } k_2 \le X < k_3, \\ \beta_{l0} + \beta_{l1}k_1 + \beta_{l2}(k_2 - k_1) + \beta_{l3}(k_3 - k_2) + \beta_{l4}(X - k_3) & \\ = [\beta_{l0} + \beta_{l1}k_1 + \beta_{l2}(k_2 - k_1) + \beta_{l3}(k_3 - k_2) - \beta_{l4}k_3] + \beta_{l4}X & \text{if } k_3 \le X. \end{cases}$$

Thus, the slopes of the lines in the four intervals are given by β_{lj}, $j = 1, 2, 3, 4$ and the four intercepts are functions of β_{lj}, $j = 0, 1, 2, 3, 4$ and the three knots.

While linear spline functions, like those in equation (4.1), are relatively easy and simple to describe they may not be sufficiently flexible to model a complex nonlinear relationship between an outcome and a covariate. In these settings restricted cubic splines are a good choice. In this approach the spline functions are linear

in the first and last intervals and are cubic functions in between, but join at the knots. Restricting the functions to be linear in the tails serves to eliminate wild fluctuations than can be a result of a few extreme data points. The definitions of the restricted cubic spline variables, used by STATA, formed from three knots are as follows:

$$x_{c1} = X,$$

and

$$\begin{aligned} x_{c2} &= \frac{1}{(k_3 - k_1)^2} \times \left\{ \left(X - k_1\right)_+^3 - (k_3 - k_2)^{-1} \left[\left(X - k_2\right)_+^3 (k_3 - k_1) \right.\right. \\ &\quad \left.\left. - \left(X - k_3\right)_+^3 (k_2 - k_1) \right] \right\} \\ &= \frac{1}{(k_3 - k_1)^2} \times \left\{ \left(X - k_1\right)_+^3 - \frac{(X - k_2)_+^3 (k_3 - k_1)}{(k_3 - k_2)} + \frac{(X - k_3)_+^3 (k_2 - k_1)}{(k_3 - k_2)} \right\}, \end{aligned}$$

where the function $(u)_+$ is defined as

$$(u)_+ = \begin{cases} 0, & u \le 0 \\ u, & u > 0 \end{cases}$$

and the logit is

$$g(\mathbf{x}_c, \boldsymbol{\beta}_c) = \beta_{c0} + \beta_{c1} x_{c1} + \beta_{c2} x_{c2}. \tag{4.2}$$

The restricted cubic spline covariate, x_{c2}, is obviously much more complex and more difficult to understand from its formula than the linear spline covariates. The value of this covariate in each of the four intervals is as follows:

$$x_{c2} = \begin{cases} 0 & \text{if } X < k_1, \\ \frac{(X-k_1)^3}{(k_3-k_1)^2} = \frac{X_*^3}{c^2} & \text{if } k_1 \le X < k_2, \\ \frac{1}{(k_3-k_1)^2} \left\{ \left(X - k_1\right)^3 - \frac{(X-k_2)^3 (k_3-k_1)}{(k_3-k_2)} \right\} & \text{if } k_2 \le X < k_3, \\ \quad = -\frac{a}{bc^2} \{ X_*^3 - 3cX_*^2 + 3acX_* - a^2 c \} & \\ \frac{1}{(k_3-k_1)^2} \left\{ \left(X - k_1\right)^3 - \frac{(X-k_2)^3 (k_3-k_1)}{(k_3-k_2)} + \frac{(X-k_3)^3 (k_2-k_1)}{(k_3-k_2)} \right\} & \text{if } k_3 \le X, \\ \quad = \frac{a}{c} [3X_* - (a + c)] & \end{cases}$$

where

$$X_* = X - k_1, a = k_2 - k_1, b = k_3 - k_2, \text{and} \quad c = a + b. \tag{4.3}$$

Obviously, one could use as many or as few knots as one wished. The more knots one chooses the more flexible the resulting fit, but at a price of more parameters to estimate. In most applications three to five knots are sufficient. One could choose the knots to be equally spaced over the range of the covariate. For example, if the range of the covariate was from 0 to 50 and one wanted four knots then one could choose values 10, 20, 30, and 40. One might choose equally spaced percentiles,

Table 4.1 Distribution Percentiles Defining Placement of Knots for Splines

# of Knots	Percentiles					
3	10	50	90			
4	5	35	65	95		
5	5	27.5	50	73.5	95	
6	5	23	41	59	77	95
7	2.5	18.33	34.17	65.83	81.67	97.5

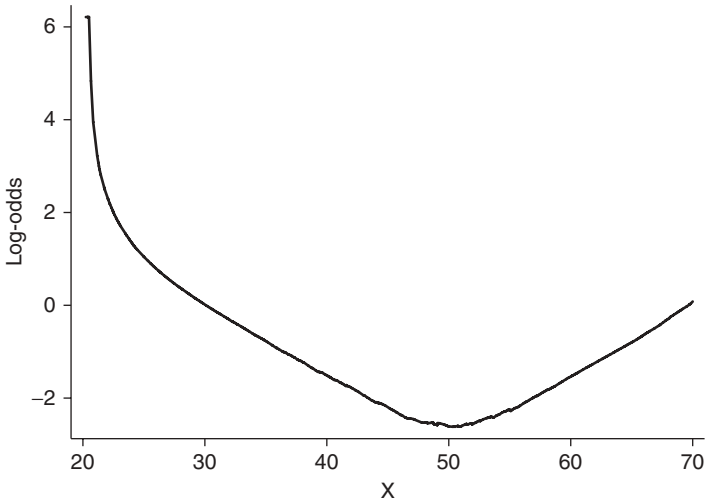

Figure 4.1 Lowess smooth on the log-odds scale of outcome Y versus the covariate X, $n = 500$.

for example, the 25th, 50th and 75th for three knots. Alternatively, Harrell (2001) provides percentiles, for three to seven knots, that have been shown in simulations to provide a good fit to wide range of shapes. These are given in Table 4.1.

Before we use purposeful selection with one of our data sets to build a model we present an example illustrating each of the four methods to examine the scale of a continuous covariate. The data are hypothetical and have been generated with a slightly asymmetric but quadratic-like shape. The data are available as Scale_Example and contain 500 observations of a continuous covariate, X, ranging from 20 to 70 and a binary outcome, Y, coded 0 and 1.

The first method discussed in this section is the graphical presentation of the lowess smooth of the outcome versus the covariate. This was computed in STATA and is shown in Figure 4.1. Recall that the lowess smooth provides a nonparametric description of the relationship between the logit or log-odds and the covariate. Hence, if there is any nonlinearity in the relationship it should be apparent in this plot. In fact, in this example, the departure from linearity is easily seen in Figure 4.1. The relationship is clearly asymmetric in shape. However, describing its

shape mathematically from the figure would represent a challenge that is beyond the capabilities of most readers (and even the authors) of this book. Hence, the lowess smooth, while quite useful for displaying nonlinearity in the logit does not lend itself well to modeling decisions about what the correct scale might actually be.

When faced with a complex relationship like the one shown in Figure 4.1 subject matter investigators might decide to categorize the covariate into four groups, effectively using the quartile design variables. We categorized X into four groups using cutpoints of 32, 44, and 56, which are the quartiles rounded to whole numbers. The estimated coefficients and standard errors for this logistic model are presented in Table 4.2. As described earlier, to check linearity in the logit we would plot each of the coefficients versus the midpoint of the interval, using 0.0 as the coefficient for the first quartile. Were we to present this plot it would show the log-odds ratios [each point comparing the log-odds for each quartile to the log-odds for the first quartile (i.e., the reference group)]. However, to compare the lowess smooth to the fitted model in Table 4.2 we need to plot its linear predictor (i.e., the logit, or log-odds). To plot the fitted logit values computed from the model in Table 4.2 we compute the following:

$$\begin{aligned}\text{logit}(X) &= \beta_0 + \beta_1 \times (X\_2) + \beta_2 \times (X\_3) + \beta_3 \times (X\_4) \\ &= \begin{cases} 0.754 - 2.213\,(0) - 4.451(0) - 1.992(0) & \text{if } X < 32 \\ 0.754 - 2.213\,(1) - 4.451(0) - 1.992(0) & \text{if } 32 \le X < 44 \\ 0.754 - 2.213\,(0) - 4.451(1) - 1.992(0) & \text{if } 44 \le X < 56 \\ 0.754 - 2.213\,(0) - 4.451(0) - 1.992(1) & \text{if } 56 \le X. \end{cases}\end{aligned}$$

This provides the values needed for the step function seen in Figure 4.2.

Next, we fit the model using linear splines with knots at 32, 44, and 56. The fit of the model using four linear splines in equation (4.1) is shown in Table 4.3. Due to the way the spline variables were created the coefficients estimate the slope of the logit in each interval. The magnitude of the slopes agrees with the plot in Figure 4.1, in that they become progressively less negative and then positive.

In order to compare the three approaches illustrated so far, we plot each on the same set of axes in Figure 4.2. In addition, we plot the value of the linear spline fit at each of the three knots. In order to better compare the linear spline fit to the fit from the quartile design variables, we plot the mean value of the logit from the linear spline fit within each quartile versus the midpoint of the quartile. In looking

Table 4.2 Results of Fitting the Logistic Regression Model with Quartile Design Variables (X_*j*), $n = 500$

Variable	Coeff.	Std. Err.	z	p
X_2	−2.213	0.3006	−7.36	<0.001
X_3	−4.451	0.6151	−7.24	<0.001
X_4	−1.992	0.2850	−6.99	<0.001
Constant	0.754	0.1917	3.93	<0.001

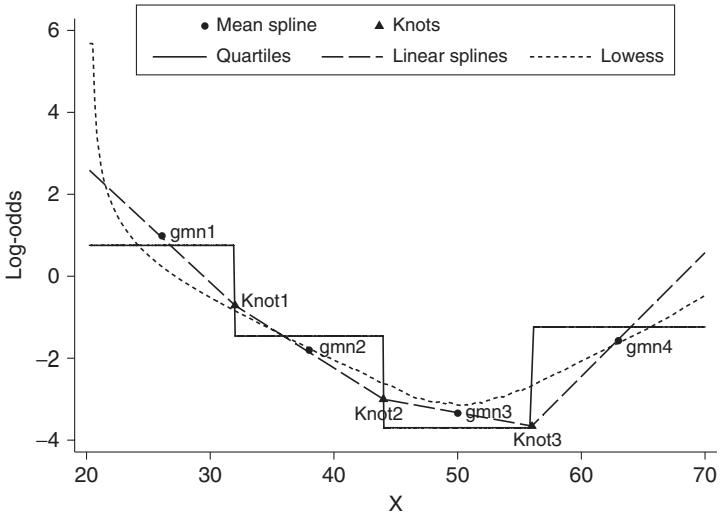

Figure 4.2 Plot of the fitted model using quartiles (—), linear splines (– –), and the lowess smooth (--). Also shown are the three knots (Knot$j$, Δ) and the mean of the linear spline fit within each quartile (gmn$j$, •), $n = 500$.

Table 4.3 Results of Fitting the Logistic Regression Model with Linear Spline Variables at Knots 32, 44, and 56, $n = 500$

Variable	Coeff.	Std. Err.	z	p
x_{l1}	−0.280	0.0552	−5.08	<0.001
x_{l2}	−0.191	0.0542	−3.52	<0.001
x_{l3}	−0.055	0.0673	−0.81	0.418
x_{l4}	0.302	0.0591	5.12	<0.001
Constant	8.263	1.5619	5.29	<0.001

at the plot several things become apparent: The fits from the linear splines and quartile design variables follow the lowess smooth to the extent that their inherent discreteness allows. The fit from the quartile design variables approximates quite closely the mean of the fit from the linear splines. So, in essence, one might say that using quartile design variables is a "poor man's" linear spline fit. Lastly, both fits are just too discrete to help suggest a model that could capture the nonlinearity seen in the lowess smooth.

In order to better explore the complicated nonlinear relationship between the logit of Y and X we display the results of using fractional polynomials in Table 4.4. The values in the column "Dev. Dif." present the difference between the deviance from the model defined by the row and that of the two-term model in the last row. This is the closed test procedure. The fact that the p-values are <0.001 in each row tells us that the two-term fractional polynomial (2, 2) is significantly different (better) than the model fit in each row. In particular, it is better than both the

Table 4.4 Results of the Fractional Polynomial Analysis

X	df	Deviance	Dev. Dif.	p	Powers
Not in model	0	592.953	206.085	<0.001	
Linear	1	521.007	134.14	<0.001	1
$m = 1$	2	452.668	65.8	<0.001	−2
$m = 2$	4	386.868			2 2

linear fit and the one-term fractional polynomial model with power −2. Hence, from a purely statistical view point we would choose the two-term model. Recall the powers (2, 2) means that this model contains X^2 and $X^2 \times \ln(X)$. The fit of this model is shown in Table 4.5.

The results in Table 4.5 indicate that the coefficients for both fractional polynomial variables are significant, but it is difficult to tell what the shape of the resulting logit as a function of the covariate X would be by simply looking at the coefficients. (Note that we divided X by 10 in calculating $Xfp1$ and $Xfp2$ so that the estimated coefficients are not excessively small.) The best and easiest way to make some judgment about shape is to examine the plot of the function. This is shown as the solid line in Figure 4.3.

Next, we fit the model with restricted cubic splines. The results are presented in Table 4.6. The first thing we note about the fit in Table 4.6 is that both estimated coefficients are significant, but are of a completely different magnitude than those for the fractional polynomial model in Table 4.5. Again, the only way to really understand the fit is via a plot. Figure 4.3 now includes the fit from the restricted cubic spline model and the lowess smooth in addition to the fractional polynomial model described earlier.

It is difficult to see from the plots in Figure 4.3 which of the two models fits better in the sense of mimicking the lowess fit. However, the deviance of the fractional polynomial model is 386.868 while that of the restricted cubic spline model is 395.128, a difference of 8.260, which suggests that the fractional polynomial model has the better fit. We also note that the fractional polynomial model appears to model the asymmetry better than the restricted cubic spline model. The knots used correspond to the quartiles and not the 10th, 50th, and 90th percentiles as suggested in Table 4.1. The fit using these knots (24, 44, and 66) had just a slightly

Table 4.5 Results of Fitting the Two-Term Fractional Polynomial (2, 2) Model, $n = 500$

Variable	Coeff.	Std. Err.	z	p
$Xfp1$	−1.883	0.1751	−10.75	<0.001
$Xfp2$	0.892	0.0843	10.58	<0.001
Constant	7.959	0.7874	10.11	<0.001

$Xfp1 = (X/10)^2$ and $Xfp2 = (X/10)^2 \times \ln(X/10)$

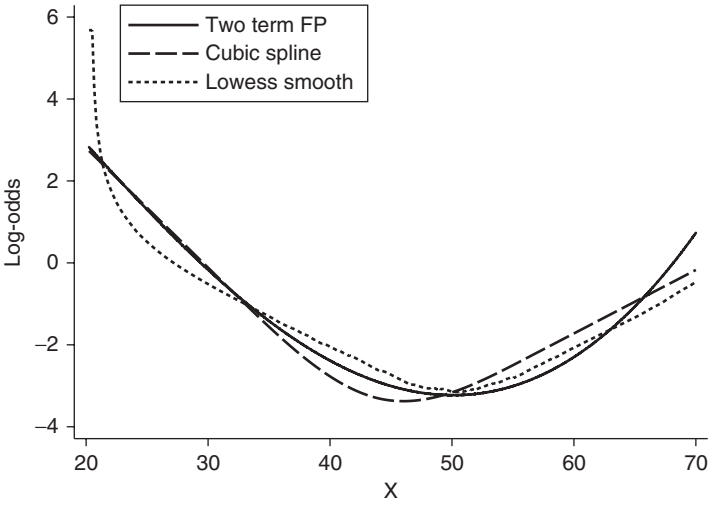

Figure 4.3 Plot of the fitted model using the two-term fractional polynomial (—), restricted cubic splines (– –), and the lowess smooth (--).

Table 4.6 Results of Fitting the Restricted Cubic Spline Model with Knots at 32, 33, and 56, $n = 500$

Variable	Coeff.	Std. Err.	z	p
x_{c1}	−0.292	0.0274	−10.67	<0.001
x_{c2}	0.298	0.0313	9.51	<0.001
Constant	8.639	0.8706	9.92	<0.001

smaller deviance, 391.646. Using four knots placed at the percentiles in Table 4.1 yields a model with effectively the same deviance as the fractional polynomial model, but at a cost of more parameters and much more complex parameterization of X. Hence our conclusion is that, based on statistical considerations, the two-term fractional polynomial model provides the better nonlinear fit from among the models explored. The phrase "statistical considerations" is an important qualifier, as the resulting shape of the logit must make clinical sense before it is used in further modeling. One other point, which we do not illustrate here, is that estimating odds ratios is considerably easier with fractional polynomial models than it is with restricted cubic spline models. Thus, if the goal is to model a nonlinear logit and to then estimate odds ratios for this covariate we highly recommend using fractional polynomials over restricted cubic splines. On the other hand if the goal is simply to model nonlinearity in the logit to control for confounding without odds ratio estimation then restricted cubic splines offer the possibility to model a quite complex relationship without actually having to specify its parametric form.

One special type of "continuous" variable that occurs reasonably often in practice is one that has many values at "zero". Consider a study in which subjects are asked

to report their lifetime use of cigarettes. All the nonsmokers report a value of zero. A one-half pack-a-day smoker for 20 years has a value of approximately 73,000 cigarettes. What makes this covariate unusual is the fact that the zero value occurs with a frequency much greater than expected for a fully continuous distribution. In addition, the nonzero values typically exhibit right skewness. Robertson et al. (1994) show that the correct way to model such a covariate is to include two terms, one that is dichotomous recording zero versus nonzero and one for the actual recorded value. Thus, the logit for such a model is

$$g(x, \boldsymbol{\beta}) = \beta_0 + \beta_1 d + \beta_2 x,$$

where $d = 0$ if $x = 0$ and $d = 1$ if $x > 0$. The advantage of this parameterization is that it allows us to model two different odds ratios. The odds ratio comparing a nonsmoker to a smoker with x^* lifetime cigarettes is

$$\text{OR}(x = x^*, x = 0) = e^{\beta_1 + \beta_2 x^*}$$

and the odds ratio for an increase of c in lifetime cigarettes is

$$\text{OR}(x = x + c, x = x) = e^{\beta_2 c}.$$

Note that during the modeling process we still need to check the scale in the logit for the positive values of the covariate. Since the distribution of x is typically skewed, fractional polynomial analysis often suggests using the one-term transformations $\ln(x)$ or $\sqrt{x}$. As noted above, odds ratios can be estimated by following the four step method discussed in Chapter 3.

4.2.2 Examples of Purposeful Selection

Example 1: The GLOW Study. For our first example of purposeful selection we use the GLOW500 data. This study is described in detail in Section 1.6.3 and the variables are described in Table 1.7. Before beginning, we remind the reader that these data are a sample from the much larger GLOW study. In particular, we over sampled fractures to obtain a modest sized data set where meaningful model building would be possible. This analysis provides a good example of an analysis designed to identify risk factors for a specified binary outcome. In this example, the outcome is fracture during the first year of follow up. Among the 500 women in this data set 125 (25%) had an incident fracture.

Step 1: The first step in purposeful selection is to fit a univariable logistic regression model for each covariate. The results of this analysis are shown in Table 4.7. Note that in this table, each row presents the results for the estimated regression coefficient(s) from a model containing only that covariate.

Table 4.7 Results of Fitting Univariable Logistic Regression Models in the GLOW Data, $n = 500$

	Coeff.	Std. Err.	$\widehat{\text{OR}}$	95% CI	G	p
AGE	0.053	0.0116	1.30[a]	1.16, 1.46	21.27	<0.001
WEIGHT	−0.0052	0.0064	0.97[b]	0.91, 1.04	0.67	0.415
HEIGHT	−0.052	0.0171	0.60[c]	0.43, 0.83	9.53	0.002
BMI	0.006	0.0172	1.03[d]	0.87, 1.22	0.11	0.738
PRIORFRAC	1.064	0.2231	2.90	1.87, 4.49	22.27	<0.001
PREMENO	0.051	0.2592	1.05	0.63, 1.75	0.04	0.845
MOMFRAC	0.661	0.2810	1.94	1.12, 3.36	5.27	0.022
ARMASSIST	0.709	0.2098	2.03	1.35, 3.07	11.41	0.001
SMOKE	−0.308	0.4358	0.74	0.31, 1.73	0.53	0.469
RATERISK						
RATERISK_2	0.546	0.2664	1.73	1.02, 2.91	11.76	0.003
RATERISK_3	0.909	0.2711	2.48	1.46, 4.22		

[a] Odds Ratio for a 5-year increase in AGE.
[b] Odds Ratio for a 5 kg increase in WEIGHT.
[c] Odds Ratio for a 10 cm increase in HEIGHT.
[d] Odds Ratio for a 5 kg/m^2 increase in BMI.

Table 4.8 Results of Fitting the Multivariable Model with All Covariates Significant at the 0.25 Level in the Univariable Analysis in the GLOW Data, $n = 500$

	Coeff.	Std. Err.	z	p	95% CI	
AGE	0.034	0.0130	2.63	0.008	0.009,	0.060
HEIGHT	−0.044	0.0183	−2.40	0.016	−0.080,	−0.008
PRIORFRAC	0.645	0.2461	2.62	0.009	0.163,	1.128
MOMFRAC	0.621	0.3070	2.02	0.043	0.020,	1.223
ARMASSIST	0.446	0.2328	1.91	0.056	−0.011,	0.902
RATERISK_2	0.422	0.2792	1.51	0.131	−0.1253,	0.969
RATERISK_3	0.707	0.2934	2.41	0.016	0.132,	1.282
Constant	2.709	3.2299	0.84	0.402	−3.621,	9.040

Step 2: We now fit our first multivariable model that contains all covariates that are significant in univariable analysis at the 25% level. The results of this fit are shown in Table 4.8. Once this model is fit we examine each covariate to ascertain its continued significance, at traditional levels, in the model. We see that the covariate with the largest p-value that is greater than 0.05 is for RATERISK2, the design/dummy variable that compares women with RATERISK = 2 to women with RATERISK = 1. The likelihood ratio test for the exclusion of self-reported risk of fracture (i.e., deleting RATERISK_2 and RATERISK_3 from the model) is $G = 5.96$, which with two degrees of freedom, yields $p = 0.051$, nearly significant at the 0.05 level.

Step 3: Next we check to see if covariate(s) removed from the model in Step 2 confound or are needed to adjust the effects of covariates remaining in the

model. In results not shown, we find that the largest percent change is 17% for the coefficient of ARMASSIST. This does not exceed our criterion of 20%. Thus, we see that while self-reported rate of risk is not a confounder it is an important covariate. No other covariates are candidates for exclusion and thus, we continue using the model in Table 4.8.

Step 4: On univariable analysis the covariates for weight (WEIGHT), body mass index (BMI), early menopause (PREMENO) and smoking (SMOKE) were not significant. When each of these covariates is added, one at a time, to the model in Table 4.8 its coefficient did not become significant. The only change of note is that the significance of BMI changed from 0.752 to 0.334. Thus the next step is to check the assumption of linearity in the logit of continuous covariates age and height.

Before moving to step 5 we consider another possible model. Since the coefficient for RATERISK_2 is not significant, one possibility is to combine levels 1 and 2, self-reported risk less than or the same as other women, into a new reference category. The advantage of this is that the new covariate is dichotomous, but we loose information about the specific log-odds of categories 1 and 2. On consultation with subject matter investigators, it was thought that combining these two categories is reasonable. Hence we fit this model and its results are shown in Table 4.9. In this model, the coefficient for the covariate RATERISK_3 now provides the estimate of the log of the odds ratio comparing the odds of fracture for individuals in level 3 to that of the combined group consisting of levels 1 and 2.

Step 5: At this point we have our preliminary main effects model and must now check for the scale of the logit for continuous covariates age and height. We presented four different methods in Section 4.2.1: the lowess smooth, quartile design variables, fractional polynomials and spline functions. In most applied settings we would always use the lowess smooth and fractional polynomials and also do so here. We also illustrate the design variable approach, as it is always an option. We reserve use of spline functions to settings where the best two-term fractional polynomial model does not seem to provide an adequate representation of the what we see in the lowess smooth.

Table 4.9 Results of Fitting the Multivariable Model after Collapsing Rate Risk into Two Categories, $n = 500$

	Coeff.	Std. Err.	z	p	95% CI	
AGE	0.033	0.0129	2.56	0.010	0.008,	0.059
HEIGHT	−0.046	0.0181	−2.55	0.011	−0.082,	−0.011
PRIORFRAC	0.664	0.2452	2.71	0.007	0.184,	1.145
MOMFRAC	0.664	0.3056	2.17	0.030	0.065,	1.263
ARMASSIST	0.473	0.2313	2.04	0.041	0.019,	0.926
RATERISK_3	0.458	0.2381	1.92	0.054	−0.009,	0.925
Constant	3.407	3.1770	1.07	0.284	−2.820,	9.633

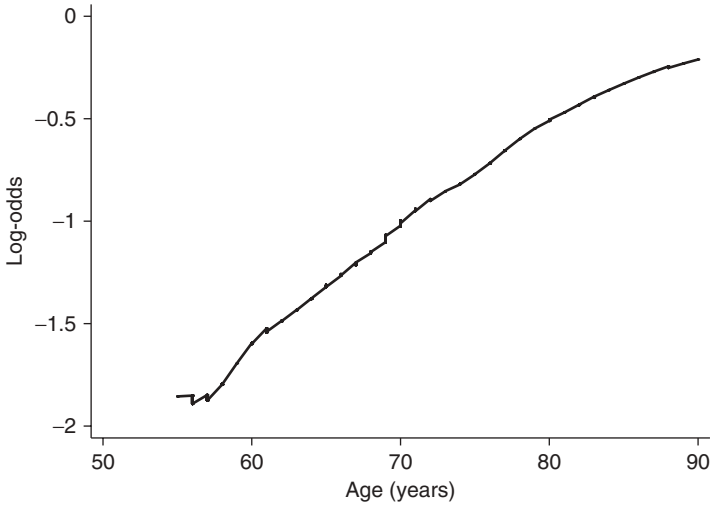

Figure 4.4 Lowess smooth on the log-odds scale of the outcome, fracture during the first year of follow-up, versus AGE, $n = 500$.

Table 4.10 Results of the Quartile Design Variable Analyses of AGE (x) from the Multivariable Model Containing the Variables Shown in the Model in Table 4.9

Quartile	1	2	3	4
Range	$x < 62$	$62 \leq x < 68$	$68 \leq x < 77$	$77 \leq x$
Midpoint	58.5	65	72.5	83.5
Coeff.	0.0	0.610	0.590	0.970
95% CI		−0.059, 1.278	−0.050, 1.229	0.311, 1.629

The lowess smooth for the outcome fracture versus age on the logit or log-odds scale is shown in Figure 4.4. Other than an inconsequential wiggle over age less than about 58, the plotted lowess smooth appears nearly linear, suggesting that there is no reason to suspect that the logit is not linear in age.

Next we examine the scale of age in the logit using quartile design variables. The results of the fit for age when it is replaced with quartile design variables in the multivariable model (Table 4.9) are shown in Table 4.10 and are plotted in Figure 4.5.

The confidence intervals for the coefficients in Table 4.10 for quartiles two and three each contain one, while that for the fourth quartile does not contain one. This suggests that the log-odds for fracture does not seem to increase significantly until after about age 72. Based on these results one might be tempted to replace age, as represented by a continuous variable, with a dichotomous variable that uses the design variable for the fourth quartile. This portrays a slightly different picture than that seen in Figure 4.4, where the lowess smoothed logit increases gradually over the entire range of age. We return to this point after performing the fractional polynomial analysis of age.

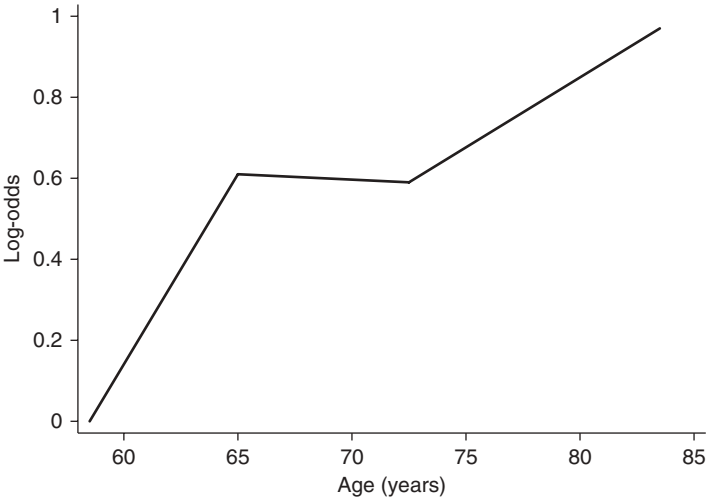

Figure 4.5 Plot of estimated logistic regression coefficients for the quartile design variables versus approximate quartile midpoints of AGE.

Table 4.11 Results of the Fractional Polynomial Analysis of AGE

	df	Deviance	Dev. dif.	p	Powers
Not in Model	0	516.421	7.468	0.113	
Linear	1	509.818	0.865	0.834	1
$m = 1$	2	509.257	0.304	0.859	-2
$m = 2$	4	508.953			3 3

The results of the fractional polynomial analysis are shown in Table 4.11. In general, when we perform a fractional polynomial analysis we proceed under the assumption that we have already decided that it is important to have the covariate in the model. Hence, we tend to ignore the results in the first row that compares the best two-term fractional polynomial model to the model not containing the covariate. The first test we look at is the one in the second row that compares the best two-term fractional polynomial model to the model treating the covariate as linear in the logit, indicated by "1" in the Powers column. In Table 4.11 the value of the likelihood ratio test is given in the "Dev. Dif." column and its p-value is in the "p" column. In this case, the test is not significant as $p = 0.834$, leading to the conclusion that the best fractional polynomial transformation is not better than the linear model. While the closed test procedure stops at this point, we always examine the results in the last two rows to see what transformations have been selected and to make sure we have not missed anything. In this case, all signs point toward treating age as linear in the logit.

The fact that the lowess smooth looks quite linear and that the supporting results from the factional polynomial analysis suggest that nothing new could be learned

about the scale of the logit in AGE from a spline variable analysis. Hence, we choose not to use it.

We remarked in discussing the plot of the quartile design variables that one might elect to dichotomize AGE at the fourth quartile. Categorization of a continuous covariate is, unfortunately, a relatively common practice in many applied fields. The temptation of its simplicity seems, in the minds of proponents, to outweigh the considerable loss of information about the covariate in such a strategy. See Royston et al. (2006) for a full discussion of the pitfalls of dichotomizing a continuous covariate. In results we do not show, but leave as an exercise, the deviance from the model using the dichotomous version of AGE is larger than that of the model in Table 4.9. Thus our decision is to treat AGE as continuous and linear in the logit.

Next we examine the continuous variable HEIGHT to determine whether it is linear in the logit. The plots of two lowess smooths on the logit scale are shown in Figure 4.6. The solid line corresponds to the smooth using all 500 subjects, while the dashed line is the smooth when one subject with a height of 199 cm is excluded. We excluded this subject to see what effect she had on the shape of the smooth. Neither smooth appears to be linear for heights less than 180 cm. The question is whether this represents a "significant" departure from linear. We examine this question using both quartile design variables and fractional polynomials (as shown in Figure 4.7).

The plot of the estimated coefficients from the quartile design variables for height shown in Figure 4.7 are based on fitting a model with $n = 500$, as the 199 cm tall woman has little effect on the coefficient in the last column of Table 4.12. The plot

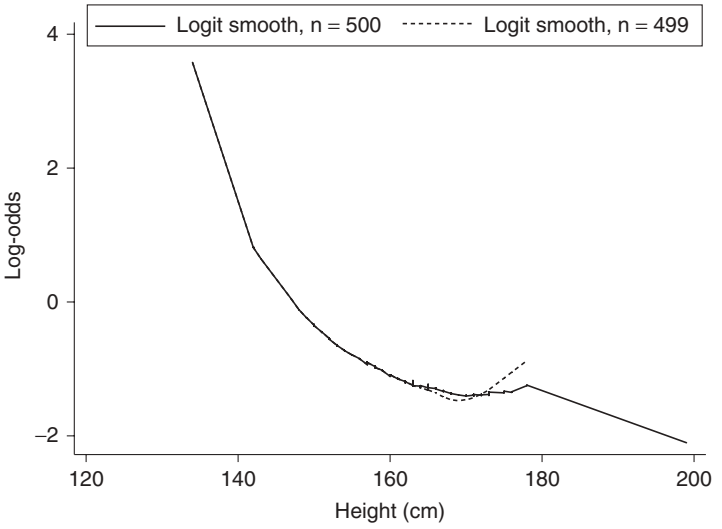

Figure 4.6 Lowess smooth on the log-odds scale of the outcome, fracture during the first year of follow up, versus HEIGHT, $n = 500$ (solid) and $n = 499$ (dashed).

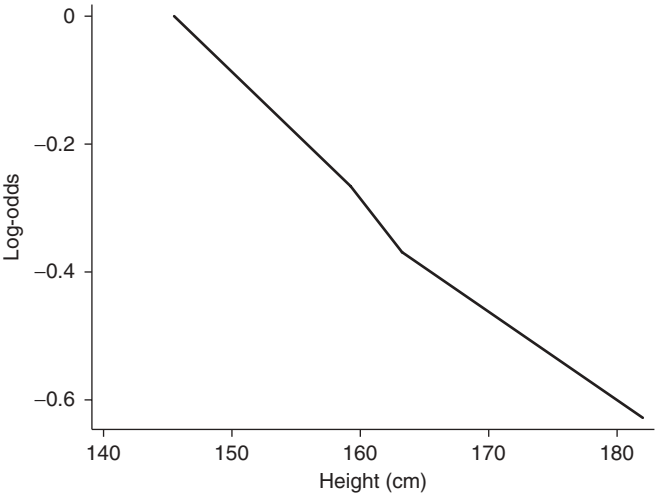

Figure 4.7 Plot of estimated logistic regression coefficients for the quartile design variables versus approximate quartile midpoints of HEIGHT.

Table 4.12 Results of the Quartile Design Variable Analyses of HEIGHT from the Multivariable Model Containing the Variables Shown in the Model in Table 4.9

Quartile	1	2	3	4
Range	$x \leq 157$	$157 < x \leq 161.5$	$161.5 < x \leq 165$	$x > 165$
Midpoint	145.5	159.25	163.25	182
Coeff.	0.0	−0.266	−0.369	−0.628
95% CI		−0.861, 0.329	−0.964, 0.226	−1.255, −0.001

Table 4.13 Results of the Fractional Polynomial Analysis of HEIGHT

	df	Deviance	Dev. Dif.	p	Powers
Not in Model	0	516.558	8.574	0.073	
Linear	1	509.818	1.834	0.608	1
m = 1	2	509.137	1.154	0.562	−2
m = 2	4	507.984			−2 −2

is strikingly linear, giving a different impression of the parametric form than what is seen in Figure 4.6.

We turn to fractional polynomials to sort out the discrepancies seen in Figure 4.6 and Figure 4.7. These results are shown in Table 4.13 where we see that the two-term fractional polynomial with powers $(-2, -2)$ is far from significantly different from the linear model. We ran the analysis excluding the 199 cm woman and the results are not appreciably different from those in Table 4.13. Hence our conclusion is to treat HEIGHT as linear in the logit. For the time being, we are going to retain

Table 4.14 Log-Likelihood, Likelihood Ratio Test (G, df = 1), and p-Value for the Addition of the Interactions to the Main Effects Model

Interaction	Log-Likelihood	G	p
Main Effects Model	−254.9089		
AGE*HEIGHT	−254.8422	0.13	0.715
AGE*PRIORFRAC	−252.3921	5.03	0.025
AGE*MOMFRAC	−254.8395	0.14	0.710
AGE*ARMASSIST	−254.8358	0.15	0.702
AGE*RATERISK3	−254.3857	1.05	0.306
HEIGHT*PRIORFRAC	−254.8024	0.21	0.645
HEIGHT*MOMFRAC	−253.7043	2.41	0.121
HEIGHT*ARMASSIST	−254.1112	1.60	0.207
HEIGHT*RATERISK3	−254.4218	0.97	0.324
PRIORFRAC*MOMFRAC	−253.5093	2.80	0.094
PRIORFRAC*ARMASSIST	−254.7962	0.23	0.635
PRIORFRAC*RATERISK3	−254.8476	0.12	0.726
MOMFRAC*ARMASSIST	−252.5179	4.78	0.029
MOMFRAC*RATERISK3	−254.6423	0.53	0.465
ARMASSIST*RATERISK3	−253.7923	2.23	0.135

the 199 cm woman in the analysis, waiting until we examine her influence using diagnostic statistics in Chapter 5. Hence our final main effects model is the one whose fit is shown in Table 4.9.

Step 6: The next step in the purposeful selection procedure is to explore possible interactions among the main effects. The subject matter investigators felt that each pair of main effects represents a plausible interaction. Hence, we fit models that individually added each of the 15 possible interactions to the main effects model. The results are summarized in Table 4.14. Three interactions are significant at the 10 percent level: Age by prior fracture (PRIORFRAC), prior fracture by mother had a fracture (MOMFRAC) and mother had a fracture by arms needed to rise from a chair (ARMASSIST). We note that prior fracture and mother having had a fracture are involved in two of the three significant interactions.

The next step is to fit a model containing the main effects and the three significant interactions. The results of this fit are shown in Table 4.15. The three degree of freedom likelihood ratio test of the interactions model in Table 4.15 versus the main effects model in Table 4.9 is $G = 11.03$ with $p = 0.012$. Thus, in aggregate, the interactions contribute to the model. However, one interaction, prior fracture by mother's fracture, is not significant with a Wald statistic $p = 0.191$. Next, we fit the model excluding this interaction and the results are shown in Table 4.16.

The estimated coefficients in the interactions model in Table 4.16 are, with one exception, significant at the five percent level. The exception is the estimated coefficient for the dichotomized self-reported risk of fracture, RATERISK3

Table 4.15 Results of Fitting the Multivariable Model with the Addition of Three Interactions, $n = 500$

	Coeff.	Std. Err.	z	p	95% CI	
AGE	0.058	0.0166	3.49	0.000	0.025,	0.091
HEIGHT	−0.049	0.0184	−2.65	0.008	−0.085,	−0.013
PRIORFRAC	4.598	1.8780	2.45	0.014	0.917,	8.278
MOMFRAC	1.472	0.4229	3.48	0.000	0.644,	2.301
ARMASSIST	0.626	0.2538	2.46	0.014	0.128,	1.123
RATERISK3	0.474	0.2410	1.97	0.049	0.002,	0.947
AGE*PRIORFRAC	−0.053	0.0259	−2.05	0.040	−0.104,	−0.002
PRIORFRAC*MOMFRAC	−0.847	0.6475	−1.31	0.191	−2.116,	0.422
MOMFRAC*ARMASSIST	−1.167	0.6168	−1.89	0.058	−2.376,	0.042
Constant	1.959	3.3272	0.59	0.556	−4.562,	8.481

Table 4.16 Results of Fitting the Multivariable Model with the Significant Interactions, $n = 500$

	Coeff.	Std. Err.	z	p	95% CI	
AGE	0.057	0.0165	3.47	0.001	0.025,	0.090
HEIGHT	−0.047	0.0183	−2.55	0.011	−0.083,	−0.011
PRIORFRAC	4.612	1.8802	2.45	0.014	0.927,	8.297
MOMFRAC	1.247	0.3930	3.17	0.002	0.476,	2.017
ARMASSIST	0.644	0.2519	2.56	0.011	0.150,	1.138
RATERISK3	0.469	0.2408	1.95	0.051	−0.003,	0.941
AGE*PRIORFRAC	−0.055	0.0259	−2.13	0.033	−0.106,	−0.004
MOMFRAC*ARMASSIST	−1.281	0.6230	−2.06	0.040	−2.502,	−0.059
Constant	1.717	3.3218	0.52	0.605	−4.793,	8.228

(1 = more, 0 = same or less) with $p = 0.051$. We elect to retain this in the model since the covariate is clinically important and its significance is nearly five percent. Hence the model in Table 4.16 is our preliminary final model. Its fit, adherence to model assumptions and assessment for influence of individual subjects is examined in Chapter 5. Following this assessment we present the results of the model in terms of odds ratios for estimates of the effect of each covariate on fracture during the first year of follow up.

In summary, our first example of model building using purposeful selection with the GLOW data illustrated: selecting variables, examining the scale in the logit for two continuous covariates and selecting and refining interactions. The resulting model in Table 4.16 is, in a sense, relatively simple in that it contains only two interactions. There was no statistical evidence of nonlinearity in the logit for the two continuous covariates.

Example 2: The Burn Injury Study. The second example is one where the goal is to obtain a model that could be used for estimating the probability of the response, as well as, to some extent, for quantifying the effect of individual risk factors. We

use the Burn Injury Study data described in Section 1.6.5 and Table 1.9. The data, BURN1000, contain information on a burn injury for 1000 subjects, 150 of whom died. As noted in Section 1.6.5 these data were sampled from a much larger data set and deaths were over sampled. Since the goal is to develop a model to estimate the probability of death from burn injury we would like a parsimonious model that would be likely to perform well in another data set. As we show later, these data illustrate some of the challenges that one can face when modeling a continuous covariate that is nonlinear in the logit. There are only six covariates and we have a large total sample size (1000) and number of outcomes (150), so rather than perform steps 1 and 2, we begin by fitting the model containing all covariates. The results of this fit are shown in Table 4.17.

In Table 4.17 the Wald test for the coefficient for GENDER is not significant with $p = 0.513$ and that of FLAME has $p = 0.100$. When we delete GENDER and refit the model the significance of the Wald test for FLAME becomes $p = 0.094$ and there is no evidence of confounding by GENDER. After consultation with an experienced burn surgeon, we decided to remove FLAME from the model for the reason that there are many different ways that flame could be involved with a burn injury and using simple yes or no coding is not precise enough to be helpful. In addition, we are striving for a model that is as parsimonious as possible. Thus our preliminary main effects model contains only four covariates: age (AGE), burn surface area (TBSA), race (RACE: 0 = non-white, 1 = white) and inhalation injury involved (INH_INJ, 0 = no, 1 = yes). The results of this fit are shown in Table 4.18.

Table 4.17 Results of Fitting a Multivariable Model to the Burn Injury Data Containing All Available Covariates, $n = 1000$

	Coeff.	Std. Err.	z	p	95% CI
AGE	0.083	0.0086	9.61	<0.001	0.066, 0.100
TBSA	0.089	0.0091	9.83	<0.001	0.072, 0.107
GENDER	−0.201	0.3078	−0.65	0.513	−0.805, 0.402
RACE	−0.701	0.3098	−2.26	0.024	−1.309, −0.094
INH_INJ	1.365	0.3618	3.77	<0.001	0.656, 2.074
FLAME	0.583	0.3545	1.64	0.100	−0.112, 1.277
Constant	−7.695	0.6912	−11.13	<0.001	−9.050, −6.341

Table 4.18 Preliminary Main Effects Model for the Burn Injury Data, $n = 1000$

	Coeff.	Std. Err.	z	p	95% CI
AGE	0.084	0.0085	9.95	<0.001	0.068, 0.101
TBSA	0.090	0.0091	9.95	<0.001	0.073, 0.108
RACE	−0.624	0.2989	−2.09	0.037	−1.209, −0.038
INH_INJ	1.523	0.3512	4.34	<0.001	0.835, 2.211
Constant	−7.595	0.6090	−12.47	<0.001	−8.788, −6.401

Table 4.19 Results of the Quartile Design Variable Analyses of the Scale of AGE

Quartile	1	2	3	4
Range	$x \leq 10.8$	$10.8 < x \leq 31.9$	$31.9 < x \leq 51.2$	$51.2 < x$
Midpoint	5.45	21.35	41.55	70.45
Coeff.	0.0	−0.483	1.139	3.770
95% CI		−1.994, 1.029	−0.066, 2.343	2.629, 4.912

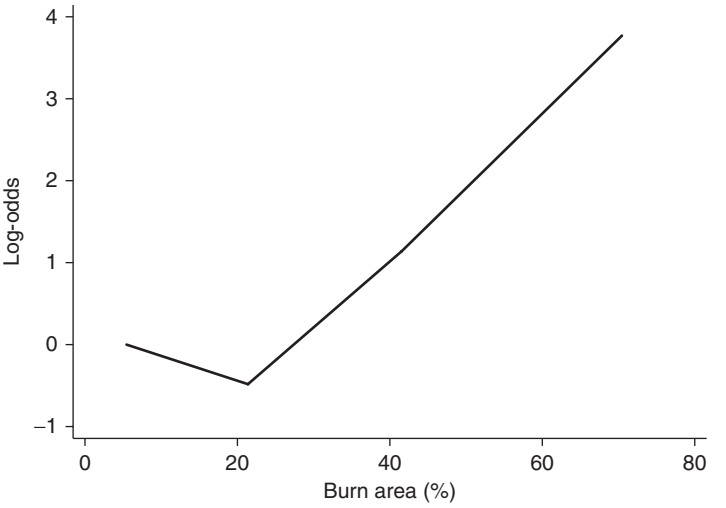

Figure 4.8 Plot of estimated logistic regression coefficients for the quartile design variables versus approximate quartile midpoints of AGE.

The next step is to examine the scale in the logit for age and burn surface area. We begin by considering age in the multivariable model in Table 4.18. The estimated coefficients for the quartile design variables are presented in Table 4.19 and plotted versus the quartile midpoints in Figure 4.8.

Only the estimated coefficient for the fourth versus the first quartile is significant. However, the plot shows that the log-odds of dying decreases and then increases, which makes clinical sense as subjects between 15 and 25, all things being equal, are known to have better outcomes than those who are younger or older. Next, we explore this in detail using the lowess smooth, fractional polynomials and restricted cubic splines.

The results of the fractional polynomial analysis of age are presented in Table 4.20. The p-values show that the two-term fractional polynomial model is better than the linear model at the 10% level but not different from the one-term fractional polynomial model. We know that a one-term fractional polynomial model is monotonic, so cannot be of the shape seen in Figure 4.8. Hence, at this point, we are going to consider both the one-term ($m = 1$) and two-term ($m = 2$)

Table 4.20 Results of the Fractional Polynomial Analysis of AGE

	df	Deviance	Dev. Dif.	p	Powers
Not in Model	0	520.362	187.147	<0.001	
Linear	1	339.785	6.569	0.087	1
$m = 1$	2	336.849	3.634	0.163	2
$m = 2$	4	333.215			3 3

fractional polynomial models as possible parameterizations of the scale of age in the logit.

Before moving on, we offer a few further comments on the models in Table 4.20. First, the model in the $m = 2$ row is the one with the numerically smallest deviance among the 36 two-term models fit. By using the "log" option in STATA we can obtain the value of the deviance for all models fit. Using this feature (output not shown) we find that there are three other two-term models [powers: (1, 0.5), (1, 1), and (2, 3)] with a deviance that differs by at most 0.7 from the best model. Note that the powers of these three models are, in a sense, no more easily interpreted than the best model's powers of (3, 3). Thus, there is no compelling reason to use any one of those as an alternative to the (3, 3) model. A natural follow up question is: If the best one-term model uses power 2, then, is the quadratic model (1, 2) an option? In this case, the deviance for the quadratic model is 335.368, which is not significantly different from the deviance for the power 2 model as $G = 1.47$ and $p = 0.225$. Also, the second best one-term model is the linear model.

Hence, by using STATA's log option we have found another model, powers 1 and 2, that may be more easily interpreted than the best fractional polynomial model. If the goal of the analysis is to estimate measures of effect for risk factors for death following a burn injury then it would make good sense to use the quadratic model as it is more easily interpreted than the power 2 model by a subject matter audience. However, our modeling goal is not effect estimation but rather estimation of the probability of death following a burn injury. For the latter goal the smaller model, power 2, may be better than the larger model, powers 1 and 2. Also, we still have additional steps in model building to perform: examining the scale of percent body surface area burned in the logit and assess the need to include interactions. In practice we would likely perform the remaining steps for both parameterizations of age. Then we would assess model adequacy and performance using the methods discussed in Chapter 5 and choose the better of the two models. This is not practical in a text so we are going to proceed with the smaller, power 2 model and leave parallel model development and evaluation, using the quadratic parameterization of age as an exercise for the reader.

Next, we try modeling age using restricted cubic splines. We found (in work we do not show here but leave as an exercise) that the best spline model is one with four knots at the percentiles in Table 4.1. The values of these four knots are: 1.1, 19, 44.37, and 78.87 years of age. The fit of this model is shown in Table 4.21, where AGESPL1, AGESPL2, and AGESPL3 are the three restricted cubic splines

Table 4.21 Fit Modeling AGE with Restricted Cubic Splines Formed from Four Knots at 1.1, 19, 44.37 and 78.87 Years, $n = 1000$

	Coeff.	Std. Err.	z	p	95% CI	
AGESPL1	−0.063	0.0608	−1.04	0.297	−0.182,	0.056
AGESPL2	0.507	0.2644	1.92	0.055	−0.011,	1.026
AGESPL3	−0.921	0.5208	−1.77	0.077	−1.941,	0.100
TBSA	0.091	0.0092	9.92	<0.001	0.073,	0.109
RACE	−0.562	0.3065	−1.83	0.067	−1.163,	0.039
INH_INJ	1.516	0.3565	4.25	<0.001	0.817,	2.215
Constant	−5.721	0.7578	−7.55	<0.001	−7.206,	−4.236

in AGE created from the four knots using an extension of the three-knot spline variable shown in equation (4.3).

In order to compare the shape of the logit in AGE for the two fractional polynomial models and the cubic spline model compared to the lowess smooth we plot all four logit functions versus age. The three parametric logit functions were scaled so that their average is the same as the average of the lowess smoothed logit. The purpose of this is to obtain a plot where the four curves are more easily compared. As an example, what we calculated to plot for the cubic spline is

$$gspl = -0.063 \times AGESPL1 + 0.507 \times AGESPL2 - 0.921 \times AGESPL3.$$

We calculated the mean of *gspl* and then added a constant to it so its mean would be equal to the mean of the lowess smooth. Similar calculations were performed using the estimated coefficient of AGE^2 to obtain $gfp1$, the mean adjusted one-term fractional polynomial model in AGE^2 and for $gfp2$, the mean adjusted two-term fractional polynomial model in AGE^3 and $AGE^3 \times \ln(AGE)$. These are shown in Figure 4.9.

We begin by comparing the four functions in the neighborhood of 20 years of age. The upper most of the four curves is the lowess smoothed logit, followed by the one-term fractional polynomial and the two-term fractional polynomial model. The lowest value results from the fit of the restricted cubic spline model. We see that the lowess smooth is nearly linear. The two fractional polynomial models are both increasing functions of age and are similar to each other, supporting $p = 0.163$ from Table 4.20. The plot of the restricted cubic spline fit has a dip, reaching its minimum at about 17 years of age and then it increases and nearly coincides with the two-term fractional polynomial model for age greater than 40. The plot of the restricted cubic spline also has the same form as the plot of the estimated coefficients from the quartile design variables in Figure 4.8.

The plots in Figure 4.9 leave us with some difficult choices. The most reasonable clinical model is the one using restricted cubic splines. However, it comes at the cost of having to use the three complex spline variables that are not easily explained, except in a figure, to clinicians. Thus, the effect of age would have to be estimated using the four-step procedure discussed in Chapter 3. The algebra necessary to

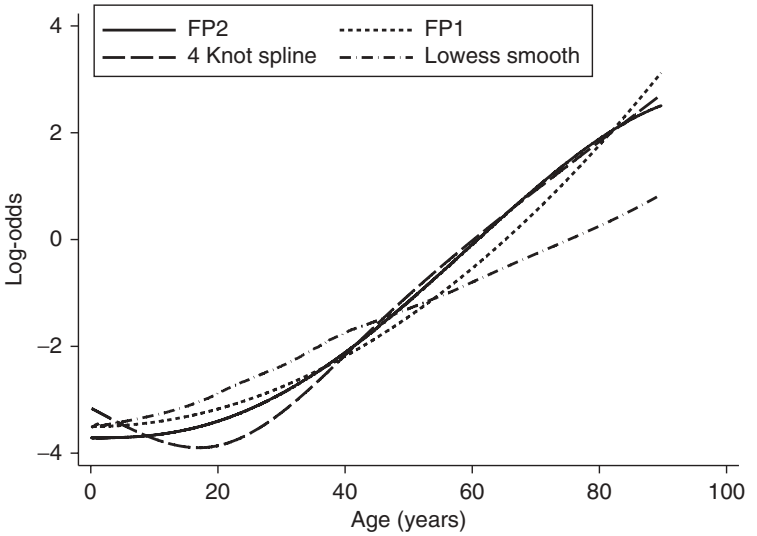

Figure 4.9 Plot of estimated logit from fits based on one (- - -) and two-term (—) fractional polynomials, restricted cubic spline (– – –), and the lowess smooth ($-\dot{c}-\dot{c}-$) of AGE. All fitted logistic regression models contain TBSA, RACE and INH_INJ.

obtain the difference in the logits is quite complicated and would yield an extremely complex equation in the spline variables and the three estimated coefficients in Table 4.21. Hence, although this is a problem that has a solution and the method for obtaining it is straightforward, the work involved is formidable. We note that once done it could be programmed. Thus, if our goal was simply to model these data we would choose to proceed with the restricted cubic splines. However, from a practical point of view, our goal is to obtain a clinically interpretable model to estimate the probability of death following a burn injury for potential use with new data. Hence our decision is to use the simple one-term fractional polynomial model as it is better than the linear model and as good as the two-term fractional polynomial model.

Before we leave consideration of the functional form in age we discuss a statistical measure that is commonly used to compare models with different numbers of parameters, the Akaike Information Criterion (AIC), Akaike (1974). This measure is defined as

$$\text{AIC} = -2 \times L + 2 \times (p + 1), \tag{4.4}$$

where L is the log-likelihood of the fitted model and p is the number of regression coefficients estimated for nonconstant covariates. Note that in Chapters 1 and 2 we defined the deviance of the fitted model as $D = -2 \times L$, thus AIC $= D + 2 \times (p + 1)$. In general, lower values of AIC are preferred to larger ones. In the current example, the deviance from the fitted one-term fractional polynomial model is $D = 336.842$. The model has five coefficients: an intercept and one each for AGE^2, TBSA, RACE, and INH_INJ. For testing purposes the transformation,

Table 4.22 Results of the Quartile Design Variable Analyses of the Scale of TBSA

Quartile	1	2	3	4
Range	$x \leq 2.5$	$2.5 < x \leq 6$	$6 < x \leq 16$	$x > 16$
Midpoint	1.3	4.25	11.0	57.0
Coeff.	0.0	0.512	1.216	3.851
95% CI		−0.729, 1.752	0.059, 2.372	2.758, 4.943

power 2, is also considered as an estimated parameter. Hence, in this case we need to add $12 = 2 \times (4 + 1 + 1)$ to the deviance not $10 = 2 \times (4 + 1)$, yielding AIC = 348.842. The value of the deviance for the spline model is $D = 331.923$. This model contains seven parameters, thus AIC = 345.923, which is smaller than the AIC for the one-term factional polynomial model. Hence, all things being equal, we would prefer the spline to the one-term fractional polynomial model. However, all things are not really equal so the considerably greater complexity of the spline model leads us to choose the one-term fractional polynomial model, even though it has a larger value of AIC. There is no statistical test to compare values of AIC.

Now that we have decided what transformation to use for age we apply the same methods to check the scale of burn area (TBSA) in the logit. At this point, we are often asked if it is better to use the transformed version of a previously examined covariate or the untransformed form. In our practice, we have not seen a set of data where using different forms gives different results. We discuss a multivariable fractional polynomial selection method in Section 4.5 that uses an iterative process using all transforms from previous iterations. So, in the current example, we follow the guidelines for purposeful selection and use AGE (untransformed) when examining the scale of burn area.

We begin by replacing TBSA in the model with the quartile-based design variables. Results for the estimated coefficients are given in Table 4.22 and plotted versus the quartile midpoint in Figure 4.10. The plot shows some departure from linearity over the first three quartiles, from 0 to 11%. Since the fourth quartile covers such a wide range we cannot see any nonlinearity in the plot beyond 11%.

The next step is to use fractional polynomials, the results of which are shown in Table 4.23. The best two-term fractional polynomial has powers −2 and 0.5. It is significantly better than the linear model with $p = 0.013$ but is not better than the one-term fractional polynomial with power 0.5, the square root ($p = 0.772$). Hence, we select the one-term transformation as best. We note that the shape of the plot in Figure 4.10 in the region less than 11% looks like a square root plot. The shape also is consistent with the burn surgeon's clinical impression of the effect of the size of burn area on mortality. The fit of this model is presented in Table 4.24.

With such straightforward and clinically plausible results from the fractional polynomial analysis we would, likely, not bother with a restricted cubic splines analysis. However, as another opportunity to demonstrate this method, we include this analysis. For TBSA, splines from three knots at the 10th (1%), 50th (6%),

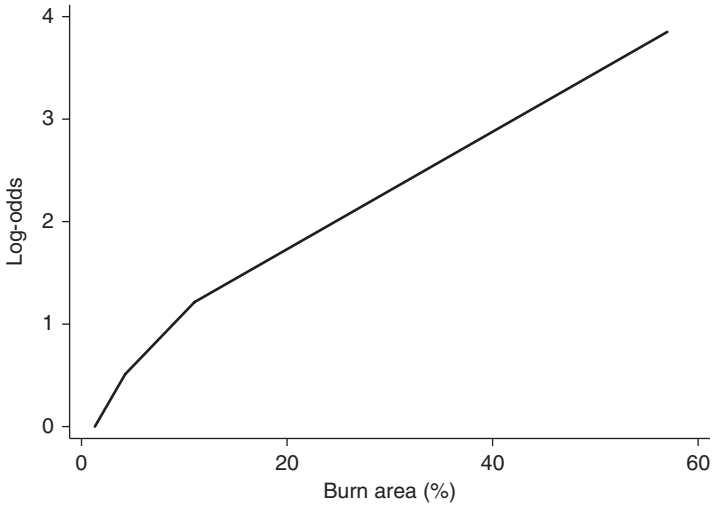

Figure 4.10 Plot of estimated logistic regression coefficients for the quartile design variables versus approximate quartile midpoints of TBSA.

Table 4.23 Results of the Fractional Polynomial Analysis of TBSA

	df	Deviance	Dev. Dif.	p	Powers
Not in Model	0	532.483	203.409	<0.001	
Linear	1	339.785	10.711	0.013	1
$m = 1$	2	329.592	0.518	0.772	.5
$m = 2$	4	329.074	—	—	−2 .5

Table 4.24 Fit of the Model Using $TBSAFP1 = \sqrt{TBSA}$, the One-Term Fractional Polynomial Transformation, $n = 1000$

	Coeff.	Std. Err.	z	p	95% CI
TBSAFP1	0.922	0.0871	10.59	<0.001	0.751, 1.092
AGE	0.085	0.0086	9.84	<0.001	0.068, 0.101
RACE	−0.623	0.3031	−2.05	0.040	−1.217, −0.029
INH_INJ	1.595	0.3463	4.60	<0.001	0.916, 2.273
Constant	−9.526	0.7544	−12.63	<0.001	−11.005, −8.048

and 90th (34.45%) percentiles of the distribution of burn area perform better (i.e., smaller deviance and AIC) than from four knots.

The results of the fit of the model using the two spline variables are shown in Table 4.25. In Figure 4.11 we plot the lowess smoothed logit and the mean adjusted logit from the one-term fractional polynomial fit,

$$gfp1 = 0.922 \times \sqrt{TBSA} - 5.468$$

Table 4.25 Fit Modeling TBSA with Restricted Cubic Splines Formed from Three Knots at 1.0, 6.0 and 34.45 Percent Burn Area, $n = 1000$

	Coeff.	Std. Err.	z	p	95% CI
TBSASPL1	0.217	0.0441	4.90	<0.001	0.130, 0.302
TBSASPL2	−0.331	0.1103	−3.00	0.003	−0.549, −0.116
AGE	0.085	0.0086	9.82	<0.001	0.068, 0.102
RACE	−0.637	0.3033	−2.10	0.036	−1.232, −0.043
INH_INJ	1.610	0.3506	4.59	<0.001	0.923, 2.297
Constant	−8.592	0.7387	−11.63	<0.001	−10.039, −7.143

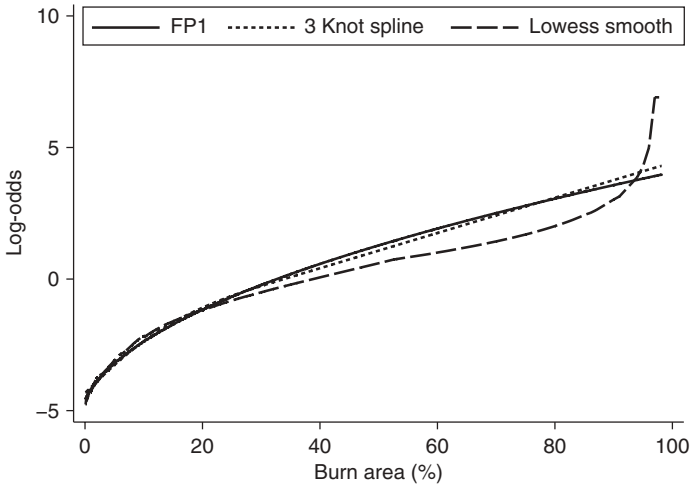

Figure 4.11 Plot of estimated logit from fits based on one-term (—) fractional polynomial, restricted cubic spline (· · ·) and the lowess smooth (– – –) of TBSA. All fitted logistic regression models contain AGE, RACE, and INH_INJ.

and the spline fit in Table 4.25,

$$gspl = 0.217 \times TBSASPL1 - 0.331 \times TBSASPL2 - 4.331,$$

where the subtracted constants are the mean adjustments.

There is virtually no difference in the plot of the logit based on the square root of TBSA and the restricted cubic spline model. We note that the lowess smoothed logit departs from these two models above 40 percent burn area. While covering a large range there are fewer than 10% of the subjects with burns this severe. The deviance for the fractional polynomial model is $D = 329.589$ and, treating the power as an estimated parameter, yields AIC $= 329.589 + 2 \times (4 + 1 + 1) = 341.589$. The deviance for the spline model is $D = 330.299$ and AIC $= 330.299 + 2 \times (5 + 1) = 342.299$. Hence we choose the model containing AGE^2, the square root of TBSA, RACE and INH_INJ as our main effects model and its fit is shown in Table 4.26 where $AGEFP1 = (AGE/10)^2$, and $TBSAFP1 = \sqrt{TBSA}$.

Table 4.26 Main Effects Model for the Burn Injury Data, $n = 1000$

	Coeff.	Std. Err.	z	p	95% CI
AGEFP1	0.087	0.0082	10.53	<0.001	0.071, 0.103
TBSAFP1	0.936	0.0874	10.71	<0.001	0.765, 1.108
RACE	−0.609	0.3096	−1.97	0.049	−1.216, −0.002
INH_INJ	1.433	0.3421	4.19	<0.001	0.763, 2.104
Constant	−7.957	0.5967	−13.34	<0.001	−9.127, −6.788

Table 4.27 Preliminary Final Model With Interaction Term for the Burn Injury Data, $n = 1000$

	Coeff.	Std. Err.	z	p	95% CI
AGEFP1	0.096	0.0096	10.02	<0.001	0.077, 0.115
TBSAFP1	0.912	0.0878	10.39	<0.001	0.740, 1.084
RACE	−0.623	0.3100	−2.01	0.045	−1.231, −0.015
INH_INJ	2.420	0.5452	4.44	<0.001	1.351, 3.488
AFP1xINH	−0.034	0.0145	−2.35	0.019	−0.063, −0.006
Constant	−8.215	0.6314	−13.01	<0.001	−9.453, −6.978

The next step in the analysis is to select interactions. With only four main effects we examined all 6 possible interactions by adding one at a time to the model in Table 4.26. Two were significant at the 10 percent level: AGEFP1 by INH_INJ and TBSAFP1 by RACE. The interaction of TBSAFP1 by RACE did not make clinical sense to the burn surgeon and thus we excluded it from further consideration. The fit of the model with the interaction is shown in Table 4.27, where AFP1xINH denotes the interaction between AGEFP1 and INH_INJ.

Before leaving this example, let us revisit the goals of the analysis. The interaction term in Table 4.27 is highly significant, demonstrating that the presence or absence of inhalation involvement with the burn injury modifies the effect of age and, likewise, age modifies the effect of inhalation involvement. Clearly, if we were interested in estimating the effect of risk factors for death we would prefer the model in Table 4.27. However, it is not clear that inclusion of the interaction would improve estimation of the probability of death. Again, simpler is sometimes better, and so, for the time being, we are going to consider both models (the ones presented in Tables 4.26 and 4.27) as possible models until we evaluate their fit and performance in Chapter 5.

4.3 OTHER METHODS FOR SELECTING COVARIATES

In the previous section we discussed purposeful selection, a method that is completely controlled by the analyst, to select a subset of covariates from a larger collection. There are other commonly used methods where selection is more automated and statistically driven. Two approaches have a long history in statistical

model building: stepwise selection and best subsets selection. A recent addition combines a version of stepwise selection with fractional polynomial modeling of continuous covariates. We consider each of these methods in this section and show how they are related to one another and compare them to purposeful selection in the context of modeling the GLOW data.

4.3.1 Stepwise Selection of Covariates

Stepwise selection of covariates has a long history in linear regression. All the major software packages have either a separate program or an option to perform this type of analysis. Currently, most, if not all, major software packages offer an option for stepwise logistic regression. At one time, stepwise regression was an extremely popular method for model building. Over the years there has been a shift away from deterministic methods for model building to methods like purposeful selection discussed in the previous section. However, we feel that stepwise methods may be useful as effective data analysis tools. In particular, there are times when the outcome being studied is relatively new and the important covariates may not be known and associations with the outcome not well understood. In these instances, most studies collect many possible covariates and screen them for significant associations. Employing a stepwise selection procedure can provide a fast and effective means to screen a large number of variables, and to fit a number of logistic regression equations simultaneously.

Any stepwise procedure for selection or deletion of variables from a model is based on a statistical algorithm that checks for the "importance" of variables, and either includes or excludes them on the basis of a fixed decision rule. The "importance" of a variable is defined in terms of a measure of the statistical significance of the coefficient, or coefficients when multiple design variables are used, for the variable. The statistics used depend on the assumptions of the model. In stepwise linear regression an F-test is used, since the model assumes that the errors are normally distributed. In logistic regression the errors are assumed to follow a binomial distribution, and significance can be assessed using any one of the three equivalent tests discussed in Chapters 1 and 2: likelihood ratio, score, and Wald test. A particular software package may or may not offer the user a choice of which of the three tests to use. We use the likelihood ratio test, in what follows, to describe the methods. The other two tests could be used equally well. In practice, we have not seen important differences in models identified when the three tests are used on the same set of data. Given a choice we prefer to use the likelihood ratio test but use of one of the other tests by a statistical package does not present a problem or disadvantage.

We discussed in Chapter 3 that a polychotomous variable with k levels is appropriately modeled through its $k-1$ design variables. Since the magnitude of the likelihood ratio test, G, depends on its degrees of freedom, any procedure based on the likelihood ratio test, or one of the other two tests, must account for possible differences in degrees of freedom between variables. This is done by assessing significance through the p-value for G.

We describe and illustrate the algorithm for forward selection followed by backward elimination in stepwise logistic regression. Any variants of this algorithm are simple modifications of this procedure. The method is described by considering the statistical computations that the computer must perform at each step.

Step (0): Assume that we have available a total of p possible independent variables, all of which are judged to be of plausible "clinical" importance in studying the outcome variable. Step (0) begins by fitting the "intercept only model" and evaluating its log-likelihood, L_0. Next, each of the p possible univariable logistic regression models is fit and its corresponding log-likelihood computed. Let the value of the log-likelihood for the model containing variable x_j at step zero be denoted by $L_j^{(0)}$. The subscript j refers to the variable that has been added to the model, and the superscript (0) refers to the step. This notation is used throughout the discussion of stepwise logistic regression to keep track of both step number and variables in the model.

Let the value of the likelihood ratio test for the model containing x_j versus the intercept only model, be denoted by $G_j^{(0)} = -2(L_0 - L_j^{(0)})$, and its p-value be denoted by $p_j^{(0)}$. This p-value is equal to the probability $\Pr[\chi^2(\nu) > G_j^{(0)}] = p_j^{(0)}$, where $\nu = 1$ if x_j is continuous or dichotomous, and $\nu = k - 1$ if x_j is polychotomous with k categories.

The "most important" variable is the one with the smallest p-value. If we denote this variable by x_{e_1}, then $p_{e_1}^{(0)} = \min(p_j^{(0)})$, where "min" stands for selecting the minimum of the quantities enclosed in the brackets. The subscript "e_1" is used to denote that the variable is a candidate for entry at step 1. For example, if variable x_2 had the smallest p-value, then $p_2^{(0)} = \min(p_j^{(0)})$, and $e_1 = 2$. The fact that x_{e_1} is the most important variable does not guarantee that it is "statistically significant". For example, if $p_{e_1}^{(0)} = 0.83$, we would probably conclude that there is little point in continuing this analysis because the most important variable is not related to the outcome. On the other hand, if $p_{e_1}^{(0)} = 0.003$, we would examine the logistic regression containing this variable and then determine whether there are other variables that are important given that x_{e1} is in the model.

A crucial factor when using stepwise logistic regression is the choice of an "alpha" level to judge the importance of variables. Let p_{E} denote our choice where the "E" stands for entry. The choice for p_{E} determines how many variables eventually are included in the model. Bendel and Afifi (1977) studied the choice of p_{E} for stepwise linear regression, and Costanza and Afifi (1979) studied the choice for stepwise discriminant analysis. Lee and Koval (1997) examined the issue of significance level for forward stepwise logistic regression. The results of this research have shown that the choice of $p_{\mathrm{E}} = 0.05$ is too stringent, often excluding important variables from the model. Choosing a value for p_{E} in the range from 0.15 to 0.20 is highly recommended.

Sometimes the goal of the analysis may be to provide a more complete set of possible predictors of the response variable. In these cases, use of $p_{\mathrm{E}} = 0.25$ (or even larger) might be a reasonable choice. Whatever the choice for p_{E}, a variable is judged important enough to include in the model if the p-value for G

is less than p_{E}. Thus, the program proceeds to step (1) if $p_{e_1}^{(0)} < p_{\mathrm{E}}$; otherwise, it stops.

Step (1): This step begins with a fit of the logistic regression model containing x_{e_1}. Let $L_{e_1}^{(1)}$ denote the log-likelihood of this model. To determine whether any of the remaining $p-1$ variables are important once the variable x_{e_1} is in the model, we fit the $p-1$ logistic regression models containing x_{e_1} and x_j, $j = 1, 2, 3, \ldots, p$ and $j \neq e_1$. For the model containing x_{e_1} and x_j let the log-likelihood be denoted by $L_{e_1 j}^{(1)}$, and let the likelihood ratio chi-square statistic of this model versus the model containing only x_{e_1} be denoted by $G_j^{(1)} = -2(L_{e_1}^{(1)} - L_{e_1 j}^{(1)})$. The p-value for this statistic is denoted by $p_j^{(1)}$. Let the variable with the smallest p-value at step (1) be x_{e_2} where $p_{e_2}^{(1)} = \min(p_j^{(1)})$. If this value is less than p_{E} then we proceed to Step (2); otherwise we stop.

Step (2): The step begins with a fit of the model containing both x_{e_1} and x_{e_2}. It is possible that once x_{e_2} has been added to the model, x_{e_1} is no longer important. Thus, Step (2) includes a check for backward elimination. In general, this check is done by fitting models that delete one of the variables added in the previous steps to assess the continued importance of the variable removed. At Step (2) let $L_{-e_j}^{(2)}$ denote the log-likelihood of the model with x_{e_j} removed. In similar fashion let the likelihood ratio test of this model versus the full model at Step (2) be $G_{-e_j}^{(2)} = -2(L_{-e_j}^{(2)} - L_{e_1 e_2}^{(2)})$ and $p_{-e_j}^{(2)}$ be its p-value.

To ascertain whether a variable should be deleted from the model the program selects that variable, which when removed, yields the maximum p-value. Denoting this variable as x_{r_2}, then $p_{r_2}^{(2)} = \max(p_{-e_1}^{(2)}, p_{-e_2}^{(2)})$. To decide whether x_{r_2} should be removed, the program compares $p_{r_2}^{(2)}$ to a second pre-chosen "alpha" level p_{R}, which indicates some minimal level of continued contribution to the model where "R" stands for remove. Whatever value we choose for p_{R}, it must exceed the value of p_{E} to guard against the possibility of having the program enter and remove the same variable at successive steps.

If we do not wish to exclude many variables once they have entered, then we might use $p_{\mathrm{R}} = 0.9$. A more stringent value would be used if a continued "significant" contribution were required. For example, if we used $p_{\mathrm{E}} = 0.15$, then we might choose $p_{\mathrm{R}} = 0.20$. If the maximum p-value to remove $p_{r_2}^{(2)}$, exceeds p_{R}, then x_{r_2} is removed from the model. If $p_{r_2}^{(2)}$ is less than p_{R} then x_{r_2} remains in the model. In either case the program proceeds to the variable selection phase.

At the forward selection step each of the $p-2$ logistic regression models are fit containing x_{e_1}, x_{e_2} and x_j for $j = 1, 2, 3, \ldots p$, $j \neq e_1, e_2$. The program evaluates the log-likelihood for each model, computes the likelihood ratio test versus the model containing only x_{e_1} and x_{e_2} and determines the corresponding p-value. Let x_{e_3} denote the variable with the minimum p-value, that is, $p_{e_3}^{(2)} = \min(p_j^{(2)})$. If this p-value is smaller than p_{E}, $p_{e_3}^{(2)} < p_{\mathrm{E}}$, then the program proceeds to Step (3); otherwise, it stops.

Step (3): The procedure for Step (3) is identical to that of Step (2). The program fits the model including the variable selected during the previous step, performs a

check for backward elimination followed by forward selection. The process continues in this manner until the program stops at Step (S).

Step (S): This step occurs when: (i) all p variables have entered the model or (ii) all variables in the model have p-values to remove that are less than p_R, and the variables not included in the model have p-values to enter that exceed p_E. The model at this step contains those variables that are important relative to the criteria of p_E and p_R. These may or may not be the variables reported in a final model. For instance, if the chosen values of p_E and p_R correspond to our preferred levels for statistical significance, then the model at step S may well contain the significant variables. However, if we have used values for p_E and p_R that are less stringent, then we should select the variables for a final model from a table that summarizes the results of the stepwise procedure.

There are two methods that may be used to select variables from a summary table; these are comparable to methods commonly used in stepwise linear regression. The first method is based on the p-value for entry at each step, while the second is based on a likelihood ratio test of the model at the current step versus the model at the last step. In most cases we prefer to use the first method as it can be performed with the output provided by statistical packages.

Let "q" denote an arbitrary step in the procedure. In the first method we compare $p_{e_q}^{(q-1)}$ to a pre-chosen significance level such as $\alpha = 0.15$. If the value $p_{e_q}^{(q-1)}$ is less than α, then we move to Step (q). We stop at the step when $p_{e_q}^{(q-1)}$ exceeds α. We consider the model at the previous step for further analysis. In this method the criterion for entry is based on a test of the significance of the coefficient for x_{e_q} conditional on $x_{e_1}, x_{e_2}, \ldots, x_{e_{q-1}}$ being in the model. The degrees of freedom for the test are 1 or $k - 1$ depending on whether x_{e_q} is continuous or polychotomous with k categories.

In the second method, rather than comparing the model at the current step [Step (q)] to the model at the previous step [Step $(q-1)$] we compare it to the model at the last step [Step (S)]. We evaluate the p-value for the likelihood ratio test of these two models and proceed in this fashion until this p-value exceeds α. This tests that the coefficients for the variables added to the model from Step (q) to Step (S) are all equal to zero. At any given step it has more degrees of freedom than the test employed in the first method. For this reason the second method, on occasion, may select a larger number of variables than the first method, but only when rather liberal, large, values are used for the entry and removal criteria.

It is well known that the p-values calculated in stepwise selection procedures are not p-values in the traditional hypothesis testing context. Instead, they should be thought of as indicators of relative importance among variables. We recommend that one err in the direction of selecting a relatively rich model following stepwise selection. The variables so identified should then be subjected to the more intensive analysis described in the previous section.

A common modification of the stepwise selection procedure just described is to begin with a model at step zero that contains known important covariates. Selection is then performed from among the other available variables. One instance when

this approach may be useful is to select interactions from among those possible from a main effects model.

Freedman (1983) urges caution when considering a model with many variables, noting that significant linear regressions may be obtained from "noise" variables, completely unrelated to the outcome variable. Flack and Chang (1987) have shown similar results regarding the frequency of selection of "noise" variables. Thus, a thorough analysis that examines statistical and clinical significance is especially important following any stepwise method.

Other versions of stepwise selection are possible. One might choose to use the previously described method but only enter variables, allowing no option for removal at each step. This is called forward selection. Another popular method is to begin at Step (0) with all p variables in the model and then proceed to sequentially eliminate nonstatistically significant variables. This is called backward elimination. We illustrate backward elimination in Section 4.3.2 as a way to approximate best subset selection.

As an example, we apply the stepwise variable selection procedure to the GLOW data analyzed using purposeful selection in Section 4.2. The reader is reminded that this procedure should be viewed as a first step in the model building process—basic variable selection. Subsequent steps such as determination of scale, as described in Section 4.2, would follow. The calculations were performed in SAS, which uses the Score Test for entry and the Wald test for removal of variables. The results are presented in Table 4.28 in terms of the p-values to enter and remove calculated at each step. The order of the variables given column-wise in the table is the order in which they were selected. In each column the values below the horizontal line are p_E values and values above the horizontal line are p_R values. The program was run using $p_E = 0.15$ and $p_R = 0.20$.

We choose to use SAS as it has the option to display the step-by-step details required for Table 4.28. One disadvantage of SAS is that it does not allow one to group the design variables formed from a categorical covariate with more than two levels for entry or removal. STATA does have this feature but has not provided step-by-step detail. However the models selected by both SAS and STATA are the same at Step (S).

Step (0): At Step (0) the program selects as a candidate for entry at Step (1) the variable with the smallest p-value in the first column of Table 4.28. The variable is history of prior fracture (PRIORFRAC). As seen in the table the p-values of both PRIORFRAC and AGE are <0.0001, but the value of the Score test (not shown) for PRIORFRAC is 23.8 while that for AGE is 21.6, each with one degree of freedom. Hence PRIORFRAC was selected for entry at Step (1).

Step (1): The program begins by fitting the model containing PRIORFRAC. The program does not remove the variable just entered since we choose the criterion such that $p_R > p_E$. This is true for the variable entered at any step—not just the first step. The variable with the smallest p-value to enter

Table 4.28 Results of Applying Stepwise Variable Selection Using the Score Test to Select Variables and the Wald Test for Removal of Variables in the GLOW Data

Variable/Step	0	1	2	3	4	5	6	7
PRIORFRAC	<0.001	<0.001	<0.001	0.002	0.003	0.003	0.007	0.009
AGE	<0.001	<0.001	<0.001*	<0.001	0.001	0.002	0.010	0.009
RATERISK3	0.006	0.046	0.017	0.018*	0.016	0.028	0.054	0.016
HEIGHT	0.002	0.009	0.0336	0.032	0.033*	0.022	0.011	0.016
MOMFRAC	0.017	0.021	0.027	0.051	0.032	0.034*	0.030	0.043
ARMASSIST	0.001	0.011	0.053	0.099	0.046	0.040	0.041*	0.056
RATERISK2	0.749	0.607	0.649	0.045	0.065	0.094	0.129	0.131*
BMI	0.738	0.745	0.217	0.110	0.128	0.091	0.342	0.333
WEIGHT	0.418	0.482	0.770	0.545	0.166	0.120	0.420	0.412
SMOKE	0.479	0.320	0.533	0.525	0.501	0.512	0.437	0.453
PREMENO	0.845	0.866	0.361	0.389	0.439	0.413	0.586	0.669

At each step the p-values to enter are presented below the horizontal line, and the p-value to remove are presented above the horizontal line in each column. The asterisk denotes the maximum p-value to remove at each step.

at step (1) is age at entry in the study (AGE) with $p < 0.001$, which is less than 0.15 so the program moves to Step (2).

Step (2): The p-values to remove appear above the solid line in each column of Table 4.28. We denote the largest p-value to remove with an "*". The model containing both PRIORFRAC and AGE is fit and we see that both p-values to remove are <0.001. Since neither exceeds 0.20, the program moves to the variable selection phase. The smallest p-value to enter among the remaining variables not in the model is $p = 0.017$, for the design variable comparing level 3 to level 1 of self-reported risk of fracture. Since the value is less than 0.15 the program proceeds to Step (3).

Step (3): At Step (3) Table 4.28 shows that the largest p-value to remove is for the variable that just entered the model, RATERISK3 and, since this does not exceed 0.20, the program moves to the variable selection phase. The smallest p-value to enter among the remaining variables not in the model is for height at enrollment in the study (HEIGHT) with $p = 0.032$. This value is less than 0.15 so the program proceeds to Step (4).

Step (4): At Step (4) the program finds that the maximum p-value to remove is HEIGHT, which just entered the model. Hence it is not removed from the model. In the selection phase the program finds that the minimum p-value for entry is 0.032 for the variable mother had a fracture, MOMFRAC. Since this value is less than 0.15, the program proceeds to Step (5).

Step (5): At Step (5) the largest p-value to remove is for MOMFRAC, which just entered the model, so it is not removed. Next the program selects for entry the variable "arms are needed to stand from a chair" (ARMASSIST).

Since the p-value for entry of 0.040 is less than p_E it enters the model at Step (6).

Step (6): At Step (6) the variable with the largest p-value to remove is, again, the variable that just entered, so none are removed. The variable with the smallest p-value to enter is the design variable for self-reported rate of risk at level 2 versus level 1, RATERISK2, with $p = 0.129$. Since this value is less than 0.15 it enters the model at Step (7).

Step (7): The variable with the largest p-value to remove is RATERISK2, which just entered the model so no variables are removed. At the selection for entry phase we see that body mass index, BMI, has the smallest p-value, but its value, 0.333, exceeds the criterion for entry of 0.15. The program stops at Step (7) as no variables can be removed and none can enter using our chosen criteria of $p_E = 0.15$ and $p_R = 0.20$.

If, for some reason, we wanted to see every variable enter the model then we would have to rerun the program with much larger values. For example, at Step (7) we see that the largest p-value for entry is 0.669 for early menopause, PREMENO. So choosing $p_E = 0.80$ and $p_R = 0.85$ would probably allow all variables to enter the model. Having said this, it would be highly unusual to choose a p-value for entry that exceeds 0.50. The idea behind letting in variables that are unlikely to be in the final model is to check for the possible confounding effect of marginally significant variables. We know from practical experience that it is rare for a variable to be a confounder if its estimated coefficient(s) in a multivariable model are significant at 0.15 or higher.

Before moving on, we note that the model selected by stepwise methods in Table 4.28 contains the same seven covariates identified by purposeful selection in Table 4.8. This is often, though not always, the case. The purposeful selection model was further simplified by excluding RATERISK2, since it was not a confounder and subject matter experts felt it was reasonable to pool "same risk" and "less risk" into a single reference category, thus using only RATERISK3. For the time being we are going to use the model at Step (7) that includes both design variables.

We noted that there are two methods to select the final model from a table summarizing the steps. In our example, the program was run with $p_E = 0.15$, a value that, we believe, selects variables with significant coefficients; thus, it is not necessary to go to the summary table to select the variables to be used in a final model. The second method is based on comparing the model at each step to the last step that–in work not shown–also selects the model at Step (7). We leave performing stepwise selection on the GLOW data using $p_E = 0.80$ and $p_R = 0.85$, with final model selection based on the second method as an exercise.

At the conclusion of the stepwise selection process we have only identified a collection of variables that seem to be statistically important. If there were known clinically important variables then these should have been added before proceeding with stepwise selection of other covariates. If at the end of stepwise election there are continuous covariates in the model, then at this point, one should determine

their appropriate scale in the logit. The model contains age and height, both of which were shown to be linear in the logit in Section 4.2.

Once the scale of the continuous covariates has been examined, and corrected if necessary, we may consider applying stepwise selection to identify interactions. The candidate interaction terms are those that seem clinically reasonable given the main effect variables in the model. We begin at Step (0) with the main effects model, including any clinically significant covariates, and sequentially select from among the possible interactions. We can use either method 1 or method 2 to select the significant interactions. The final model contains previously identified main effects and significant interaction terms.

The same software may be used for stepwise selection of interactions as was used for the selection of main effects. The difference is that all main effect variables are forced into the model at Step (0) and selection is restricted to interactions. In total there are 15 possible interactions listed in Table 4.29 where they are inverse rank ordered by the p-values at the last step. We again use SAS to select the interactions stepwise. We remind the reader that, in SAS, selection for entry is based on the Score test, and the test for removal is based on the Wald test.

Before proceeding with stepwise selection of interactions we decided to remove RATERISK2 from the model and keep RATERISK3. Thus, we have chosen to use the recoded version of the self-reported risk variable from purposeful selection in the previous section. There are 15 interactions that can be formed from the six main effects and subject matter experts considered all 15 to be clinically reasonable.

Table 4.29 Results of Applying Stepwise Variable Selection to Interactions from the Main Effects Model from the GLOW Study Using the Score Test to Select Variables and the Wald Test to Remove Variables

Variable/Step	0	1	2
AGE*PRIORFRAC	0.024	0.025*	0.033
MOMFRAC*ARMASSIST	0.028	0.038	0.040*
HEIGHT*MOMFRAC	0.112	0.110	0.162
ARMASSIST*RATERISK3	0.135	0.123	0.174
PRIORFRAC*MOMFRAC	0.092	0.123	0.188
HEIGHT*ARMASSIST	0.206	0.184	0.252
HEIGHT*RATERISK3	0.319	0.308	0.0386
PRIORFRAC*ARMASSIST	0.636	0.399	0.423
AGE*RATERISK3	0.304	0.446	0.435
AGE*MOMFRAC	0.708	0.753	0.463
MOMFRAC*RATERISK3	0.465	0.468	0.580
AGE*HEIGHT	0.716	0.803	0.795
HEIGHT*PRIORFRAC	0.644	0.815	0.904
PRIORFRAC*RATERISK3	0.726	0.843	0.959
AGE*ARMASSIST	0.702	0.968	0.999

At each step the p-values to enter are presented below the horizontal line, and the p-value to remove are presented above the horizontal line in each column. The asterisk denotes the maximum p-value to remove at each step.

The results in Table 4.29 show that only two of the interactions were selected. At Step (1) the interaction of age and history of prior fracture (PRIORFRAC) entered and at Step (2) the interaction of mother having had a fracture (MOMFRAC) and need arms to rise from a chair (ARMASSIST) entered. The most significant interaction among those not selected at Step (2) is that of HEIGHT and MOMFRAC with $p = 0.162$, which is not less than the criterion for entry of 0.15 and hence does not enter the model.

It is worthwhile to point out that the p-values for Step (0) in Table 4.29, which are based on the Score test, are quite similar to those in the last column of Table 4.14 that are based on the likelihood ratio test.

Adding the two selected interactions to the main effects (all of which were selected stepwise) yields the same model obtained by purposeful selection shown in Table 4.16. This may not always be the case. In our experience, models obtained by these two approaches rarely differ by more than a couple of variables. In a situation where different approaches yield different models, we recommend proceeding with a combined larger model via purposeful selection using both confounding and statistical significance as criteria for model simplification.

Since the stepwise model is shown in Table 4.16 we do not repeat the results in another table in this section.

In conclusion, we emphasize that stepwise selection identifies variables as candidates for a model solely on statistical grounds. Thus, following stepwise selection of main effects all variables should be carefully scrutinized for clinical plausibility. In general, interactions must attain statistical significance to alter the point and interval estimates from a main effects model. Thus, stepwise selection of interactions using statistical significance can provide a valuable contribution to model identification, especially when there are large numbers of clinically plausible interactions generated from the main effects.

4.3.2 Best Subsets Logistic Regression

An alternative to stepwise selection of variables for a model is best subset selection. This approach to model building has been available for linear regression for many years and makes use of the branch and bound algorithm of Furnival and Wilson (1974). Typical software implementing this method for linear regression identifies a specified number of "best" models containing one, two, three variables, and so on, up to the single model containing all p variables. Lawless and Singhal (1978, 1987a, 1987b) proposed an extension that may be used with any nonnormal errors model. The crux of their method involves application of the Furnival-Wilson algorithm to a linear approximation of the cross-product sum-of-squares matrix that yields approximations to the maximum likelihood estimates. Selected models are then compared to the model containing all variables using a likelihood ratio test. Hosmer et al. (1989) show that, for logistic regression, the full generality of the Lawless and Singhal approach is not needed. Best subsets logistic regression may be performed in a straightforward manner using any program capable of best subsets linear regression. Also, some packages, including SAS, have implemented

the Lawless and Singhal method in their logistic regression modules. The advantage of these two approaches is that one may examine, and hence compare, several different models selected by some criterion. If, however, one is merely interested in obtaining the best model from the best subsets method, then a quick route to this end is to employ a method described in Royston and Sauerbrei (2008, Chapter 2). They discuss results showing that the model selected using stepwise backward elimination with $p_R = 0.157$ yields a model that agrees, in content, quite closely with the best of the best subset selected models using a criterion such as AIC from equation (4.4). The disadvantage of this quicker approach is that one is not able to see the content of other best models. We illustrate best subsets selection using the GLOW data. An important caveat to using best subsets selection is that, as described, it only identifies a collection of main effects. As described in the previous two sections, there is considerable work remaining in the model building process after main effects are selected.

Applying best subsets linear regression software to perform best subsets logistic regression is most easily explained using vector and matrix notation. In this regard, we let $\mathbf{X}$ denote the $n \times (p+1)$ matrix containing the values of all p independent variables for each subject, with the first column containing 1 to represent the constant term. Here the p variables may represent the total number of variables, or those selected at the univariable stage of model building. We let $\mathbf{V}$ denote an $n \times n$ diagonal matrix with general element $v_i = \hat{\pi}_i(1-\hat{\pi}_i)$ where $\hat{\pi}_i$ is the estimated logistic probability computed using the maximum likelihood estimate, $\hat{\boldsymbol{\beta}}$, and the data for the i^{th} case, $\mathbf{x}_i$.

For the sake of clarity of presentation in this section, we repeat the expression for $\mathbf{X}$ and $\mathbf{V}$ given in Chapter 2. They are as follows:

$$\mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}$$

and

$$\mathbf{V} = \begin{bmatrix} \hat{\pi}_1\left(1-\hat{\pi}_1\right) & 0 & \cdots & 0 \\ 0 & \hat{\pi}_2(1-\hat{\pi}_2) & \cdots & 0 \\ \vdots & 0 & \ddots & \vdots \\ 0 & \cdots & 0 & \hat{\pi}_n(1-\hat{\pi}_n) \end{bmatrix}.$$

As noted in Chapter 2, the maximum likelihood estimate is determined iteratively. It may be shown [see Pregibon (1981)] that $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{V}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}\mathbf{z}$, where $\mathbf{z} = \mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{V}^{-1}\mathbf{r}$ and r is the vector of residuals, $\mathbf{r} = (\mathbf{y} - \hat{\pi})$. This representation of $\hat{\boldsymbol{\beta}}$ provides the basis for use of linear regression software. It is easy to verify that any linear regression package that allows weights produces coefficient estimates identical to $\hat{\boldsymbol{\beta}}$ when used with z_i as the dependent variable and case weights, v_i, equal to the diagonal elements of $\mathbf{V}$.

If we wanted to replicate the results of the maximum likelihood fit from a logistic regression package using a linear regression package, for each case we would first calculate the value of a dependent variable as follows:

$$\begin{aligned} z_i &= \hat{\beta}_0 + \sum_{j=1}^{p} \hat{\beta}_j x_{ij} + \frac{(y_i - \hat{\pi}_i)}{\hat{\pi}_i(1-\hat{\pi}_i)} \\ &= \ln\left(\frac{\hat{\pi}_i}{1-\hat{\pi}_i}\right) + \frac{(y_i - \hat{\pi}_i)}{\hat{\pi}_i(1-\hat{\pi}_i)} \end{aligned} \tag{4.5}$$

and a case weight

$$v_i = \hat{\pi}_i(1-\hat{\pi}_i). \tag{4.6}$$

Note that all we need is access to the fitted values, $\hat{\pi}_i$, to compute the values of z_i and v_i. Next, we would run a linear regression program using the values of z_i for the dependent variable, the values of $\mathbf{x}_i$ for our vector of independent variables, and the values of v_i for our case weights.

Proceeding further with the linear regression, it can be shown that the residuals from this fit are

$$(z_i - \hat{z}_i) = \frac{(y_i - \hat{\pi}_i)}{\hat{\pi}_i(1-\hat{\pi}_i)}$$

and the weighted residual sum-of-squares produced by the program is

$$\sum_{i=1}^{n} v_i(z_i - \hat{z}_i)^2 = \sum_{i=1}^{n} \frac{(y_i - \hat{\pi}_i)^2}{\hat{\pi}_i(1-\hat{\pi}_i)},$$

which is X^2, the Pearson chi-square statistic from a maximum likelihood logistic regression program. It follows that the mean residual sum-of-squares is $s^2 = X^2/(n-p-1)$. The estimates of the standard error of the estimated coefficients produced by the linear regression program are s times the square root of the diagonal elements of the matrix $(\mathbf{X}'\mathbf{V}\mathbf{X})^{-1}$. Thus, to obtain the correct values given in equation (2.5) we would have to divide the estimates of the standard error produced by the linear regression program by s, the square root of the mean square error (or standard error of the estimate).

The ability to duplicate the maximum likelihood fit in a linear regression package forms the foundation of the suggested method for performing best subsets logistic regression. In particular, Hosmer et al. (1989) show that use of any best subsets linear regression program with values of z_i in equation (4.5) for the dependent variable, case weights v_i shown in equation (4.6), and covariates $\mathbf{x}_i$, produces for any subset of q variables, the approximate coefficient estimates of Lawless and Singhal (1978). Hence, we may use any best subsets linear regression program to execute the computations for best subsets logistic regression. One practical difficulty is that there is not much software available that actually implements the traditional best subsets linear regression. A recent user-supplied contribution to the STATA package by Lindsey and Sheather (2010) does perform this analysis but only provides the content of the best model of each size.

The subsets of variables selected for "best" models depend on the criterion chosen for "best." In best subsets linear regression a number of different criteria have been used to select variables. Two are based on the concept of the proportion of the total variation explained by the model. These are R^2, the ratio of the regression sum-of-squares to the total sum-of-squares, and adjusted R^2 (or AR^2), the ratio of the regression mean squares to the total mean squares. Since the adjusted R^2 is based on mean squares rather than sums-of-squares, it provides a correction for the number of variables in the model. This is important, as we must be able to compare models containing different variables and different numbers of variables. If we use R^2, the best model is always the model containing all p variables, a result that is not at all helpful. An obvious extension for best subsets logistic regression is to base the R^2 measures, in a manner similar to that shown in Chapter 5, on deviance rather than Pearson chi-square. However, we do not recommend the use of the R^2 measures for best subsets logistic regression. Instead, we prefer to use C_q, a measure developed by Mallows (1973) or the Akaike Information Criterion (AIC) developed by Akaike (1974) and defined in equation (4.4).

Mallows' C_q is a measure of predictive squared error. We note that the measure is denoted as C_p by other authors. We chose to use "q" instead of "p" in this text since we use p to refer to the total number of possible variables, while q refers to some subset of variables.

A summary of the development of the criterion C_q in linear regression may be found in many texts on this subject, for example, Ryan (1997). Hosmer et al. (1989) show that when best subsets logistic regression is performed via a best subsets linear regression package in the manner described previously in this section, Mallows' C_q has the same intuitive appeal as it does in linear regression. In particular they show that for a subset of q of the p variables

$$C_q = \frac{X^2 + \lambda^*}{X^2/(n - p - 1)} + 2(q + 1) - n,$$

where

$$X^2 = \sum\{(y_i - \hat{\pi}_i)^2/[\hat{\pi}_i(1 - \hat{\pi}_i)]\},$$

the Pearson chi-square statistic for the model with p variables and λ^* is the multivariable Wald test statistic for the hypothesis that the coefficients for the $p - q$ variables not in the model are equal to zero. Under the assumption that the model fit is the correct one, the approximate expected values of X^2 and λ^* are $(n - p - 1)$ and $p - q$, respectively. Substitution of these approximate expected values into the expression for C_q yields $C_q = q + 1$. Hence, models with C_q near $q + 1$ are candidates for a best model. The best subsets linear regression program selects as best that subset with the smallest value of C_q.

The Akaike Information Criterion (AIC) does not have a reference standard based on the number of variables, in or out of the model. The best model is simply the one with the smallest value of

$$\text{AIC}_q = -2 \times L_q + 2 \times (q + 1). \tag{4.7}$$

We modified the definition in equation (4.6) by adding the subscript "q" to denote the fact that AIC is being computed over models of different sizes.

Some programs, for example, SAS's PROC LOGISTIC, provide a best subsets selection of covariates based on the Score test for the variables in the model. For example, the best two variable model is the one with the largest Score test among all two variable models. The output lists the covariates and Score test for a user specified number of best models of each size. The difficulty one faces when presented with this output is that the Score test increases with the number of variables in the model. Hosmer et al. (2008) show how an approximation to Mallows' C_q can be obtained from Score test output in a survival time analysis. A similar approximation can be obtained from C_q for logistic regression. First, we assume that the Pearson chi-square statistic is equal to its mean, that is $X^2 \approx (n - p - 1)$. Next we assume that the Wald statistic for the $p - q$ excluded covariates may be approximated by the difference between the values of the Score test for all p covariates and the Score test for q covariates, namely $\lambda_q^* \approx S_p - S_q$. This results in the following approximation

$$\begin{aligned} C_q &= \frac{X^2 + \lambda^*}{X^2/(n-p-1)} + 2(q+1) - n \\ &\approx \frac{(n-p-1) + (S_p - S_q)}{1} + 2(q+1) - n \\ &\approx S_p - S_q + 2q - p + 1. \end{aligned} \tag{4.8}$$

The value of S_p is the Score test for the model containing all p covariates and is obtained from the computer output. The value of S_q is the Score test for the particular subset of q covariates and its value is also obtained from the output. Use of the best subsets linear regression package should help identify, in the same way its application in linear regression does, a core of important covariates from the p possible covariates. After identifying the important variables, we suggest that further modeling proceed in the manner described in Section 4.2 for purposeful selection of covariates. Users should not be lured into accepting the variables suggested by a best subset strategy without considerable critical evaluation.

We illustrate best subsets selection using the Score test method implemented in SAS with the GLOW data. The variables used were the 10 indicated in Table 1.7, with the exception of the fracture risk score, since it is a composite formed from many individual covariates. Self-reported rate of risk is modeled using two design variables RATERISK2 and RATERISK3. In Table 4.30 we present the results of the five best models selected using C_q in (4.8) as the criterion. In addition to the variables selected, we show the values of C_q and the values of S_q for each model and the value of AIC_q from (4.7).

Using only the summary statistics, we would select Model 1 as the best model since it has the smallest values of both C_q and AIC_q. It is interesting to note that this model is different from the model selected by purposeful selection (Model 5), and stepwise (Model 4), in that height is not in the model, but weight and BMI are included. The differences in the values of both C_q and AIC_q over the five models

Table 4.30 Five Best Models Identified Using the Score Test Approximation of Mallow's C_q, Table Lists Model Covariates, Approximate C_q, S_q, and AIC$_q$ ($S_{11} = 59.1672$)

Model	Model Covariates	S_q	C_q	AIC$_q$
1	PRIORFRAC, AGE, WEIGHT, BMI, MOMFRAC, ARMASSIST, RATERISK2, RATERISK3	57.4602	7.707	523.1954
2	PRIORFRAC, AGE, WEIGHT, BMI, MOMFRAC, ARMASSIST, RATERISK3	55.4424	7.724	523.5289
3	PRIORFRAC, AGE, WEIGHT, BMI, MOMFRAC, RATERISK2, RATERISK3	55.2662	7.901	523.1987
4[a]	PRIORFRAC, AGE, HEIGHT, MOMFRAC, ARMASSIST, RATERISK2, RATERISK3	55.2657	7.902	523.5004
5[b]	PRIORFRAC, AGE, HEIGHT, MOMFRAC, ARMASSIST, RATERISK3	53.2400	7.927	523.8178

[a]Main effects model identified by stepwise selection.
[b]Main effects model identified by purposeful selection.

are negligible. Thus choice among the five models comes down, as it should, to subject matter considerations.

Note that all five models contain PRIORFRAC, AGE, and MOMFRAC. Four of the five contain ARMASSIST. Three models contain WEIGHT and BMI and two contain HEIGHT. Three models contain both RATERISK2 and RATERISK3 and two contain only RATERISK3. Hence, we conclude that the core of important variables in these five models is PRIORFRAC, AGE, MOMFRAC, ARMASSIST, RATERISK2, and RATERISK3, with body composition modeled either by WEIGHT and BMI or by HEIGHT.

In using purposeful selection in Section 4.2 we found that the estimated coefficient for RATERISK2 was not significant and, in consultation with experts, decided to only use RATERISK3, which is a design variable for level 3 versus 1 and 2. Now the choice is between model 2 and model 5. Further analysis showed that the estimated coefficient for ARMASSIST is not significant, $p = 0.125$, in model 2. Deleting it yields a sixth best model (not shown) with $C_q = 8.182$ and AIC$_q = 523.87$. Thus, the choice is now between two models, each with six covariates. The more important difference between the two models is that one contains HEIGHT and the other contains WEIGHT and BMI. We leave further comparison of these two models as an exercise.

In practice, once we have finalized the main effects model, we could employ best subsets selection to decide on possible interactions. We leave this as an exercise.

Application of the backwards elimination approach described by Royston and Sauerbrei (2008, Sections 2.6.3 and 2.9.3) with $p_R = 0.157$ to the GLOW data yields the same best subsets model, Model 1 in table 4.30. This is not always going to be the case, but this easy to use approach should always identify a reasonable set of model covariates for further evaluation.

The advantage of the proposed method of best subsets logistic regression is that more models can be quickly screened than is possible with the other variable selection methods. There is, however, one potential disadvantage with the best subsets approach: we must be able to fit the model containing all of the possible covariates. In analyses that include a large number of variables this may not be possible. Numerical problems can occur when we overfit a logistic regression model. If the model has many variables, we run the risk that the data are too thin to be able to estimate all the parameters. If the full model proves to be too rich, then some selective weeding out of obviously unimportant variables with univariable tests may remedy this problem. Another approach is to perform the best subsets analysis using several smaller "full" models. Numerical problems are discussed in more detail in the next section.

In summary, the ability to use weighted least squares best subsets linear regression software or the Score test approximation method to identify variables for logistic regression should be kept in mind as a possible aid to variable selection. As is the case with any statistical selection method, the clinical basis of all variables should be addressed before any model is accepted as the final model.

4.3.3 Selecting Covariates and Checking their Scale Using Multivariable Fractional Polynomials

Sauerbrei et al. (2006) describe software for SAS, STATA and R that implements a multivariable fractional polynomial method. Royston and Sauerbrei (2008, Chapter 6) describe the method in detail and it is now available in distributed STATA. The method combines elements of backward elimination of nonsignificant covariates with an iterative examination of the scale of all continuous covariates and can be used with either the closed or sequential test procedures described in Section 4.2.

The multivariable fractional polynomial procedure requires that two significance levels be specified: the first, α_1, for the test for exclusion from or addition to, the model and the second, α_2, to assess the significance of the fractional polynomial transforms of a continuous covariate. We use the same notation as Royston and Sauerbrei (2008) to denote the method and its significance levels, namely mfp(α_1, α_2).

The method begins, cycle 1, by fitting a multivariable model that contains the user-specified covariates. This initial collection, ideally, would include all study covariates. However, we may have a setting where this is not possible, for any one of a number of numerical problems. If this occurs, a reasonable solution is to choose a subset of covariates that includes the clinically important covariates and those significant at, say, the 25 percent level on univariable analysis. This is, basically, the starting point of purposeful selection.

The initial fit at cycle 1 includes all covariates as linear terms in the logit. In subsequent fits, each covariate is modeled according to a specified number of degrees of freedom. All dichotomous and design variables have one degree of freedom, meaning they are not candidates for fractional polynomial transformation. Continuous covariates may be forced to be modeled linearly by specifying one degree

of freedom, or may be candidates for a one- or two-term fractional polynomial by specifying 2 or 4 degrees of freedom, respectively.

Following the initial multivariable linear fit, variables are considered in descending order of their Wald statistics. For covariates modeled with one degree of freedom, a partial likelihood ratio test is used to assess their contribution to the model, and its significance relative to the chosen level of significance, α_1, is noted. Continuous covariates are modeled using either the closed or sequential test method, noting whether the covariate should be removed using α_1, kept linear, or transformed using α_2. In keeping with our approach to stepwise selection and best subsets we set the level of significance for staying in the model at $\alpha_1 = 0.15$. We use the five percent level of significance, $\alpha_2 = 0.05$, for testing the need to transform. In the example, we use the closed test procedure, which is the default method in STATA. This completes the first cycle.

The second cycle begins with a fit of a multivariable model containing the significant covariates from cycle one (i.e., the model with significant continuous covariates, that may be transformed and significant dichotomous covariates). All covariates, examined in descending order of significance, are considered again for possible transformation, inclusion or exclusion from the model. Continuous covariates with a significant fractional polynomial transformation are entered transformed, which becomes their null model. The point of this step is twofold: (1) does the transformation "linearize" the covariate in the logit? and (2) does the transformation affect scaling of other covariates? Each covariate's level of significance is noted as well as the need to transform. This completes the second cycle.

The procedure stops when the results of two consecutive cycles are the same. The minimum number is two. More than two cycles occur if additional transformations of continuous covariates are suggested in cycle two and beyond, or if the level of significance of the partial likelihood ratio test for contribution to the model, changes the decision to include or exclude a covariate.

We use mfp(0.15, 0.05) on the GLOW500 data from the GLOW Study with the same 10 covariates used in the previous three sections and model self-reported risk of fracture with two design variables. We note that in STATA one may consider design variables formed from a categorical covariate with more than two levels as a group or separately. In the example, we consider the two design variables for self-reported risk of fracture separately, as that is how they were modeled in stepwise and best subsets. The method took two cycles to converge. We present the results from cycle 1 in Table 4.31, and cycle 2 in Table 4.32.

The cycle begins by fitting the model containing all 11 covariates. In Table 4.31, the first covariate processed is having had a prior fracture, PRIORFRAC, so we know it had the largest Wald statistic. Because PRIORFRAC is dichotomous the first test, line* 1, compares the 10 covariate model not containing PRIORFRAC to the 11 covariate model containing PRIORFRAC. This is indicated in the last two

*The STATA output does not include line numbers. We included them in Table 4.31 and Table 4.32 to help in discussing the results.

Table 4.31 Results from the Cycle 1 Fit of MFP Applied to the GLOW500 Data

Line	Variable	Model	(vs.)	Deviance	G	p	Powers	(vs.)	
1	PRIORFRAC	null	lin.	511.004	7.167	0.007*	.		1
2		Final		503.837			1		
3	AGE	null	FP2	510.869	7.629	0.106*	.	3	3
4		lin.		503.837	0.598	0.897	1		
5		Final		503.837			1		
6	RATERISK3	null	lin.	510.335	6.498	0.011*	.		1
7		Final		503.837			1		
8	MOMFRAC	null	lin.	507.944	4.107	0.043*	.		1
9		Final		503.837			1		
10	RATERISK2	null	lin.	506.123	2.286	0.131*	.		1
11		Final		503.837			1		
12	BMI	null	FP2	506.098	4.524	0.340	.	−2	1
13		Final		506.098			.		
14	ARMASSIST	null	lin.	508.155	2.058	0.151	.		1
15		Final		508.155			.		
16	WEIGHT	null	FP2	510.209	6.103	0.192	.	−2	−2
17		Final		510.209			.		
18	HEIGHT	null	FP2	514.905	6.689	0.153	.	−2	−2
19		Final		514.905			.		
20	SMOKE	null	lin.	515.296	0.391	0.532	.		1
21		Final		515.296			.		
22	PREMENO	null	lin.	515.844	0.547	0.459	.		1
23		Final		515.844			.		

*$p <$ chosen significance level for inclusion.
†$p <$ chosen significance level for transformation.

columns of line 1 where "." denotes that the covariate is not in the model and "1" denotes that it is modeled linearly in the logit. The value in the Deviance column, 511.004, in line 1 is for the model that excludes PRIORFRAC. The value in the G column of line 1, 7.167, is the difference between 511.004 and the Deviance for the model containing PRIORFRAC. The value in the p column in line 1 is the significance level using one degree of freedom, $\Pr[\chi^2(1) \geq 7.167] = 0.007$. The "*" denotes that the test is significant at the specified significance level for inclusion in the model, $\alpha_1 = 0.15$. Because the test is significant and since PRIORFRAC is dichotomous the final model in line 2 is the one that includes PRIORFRAC. Hence, in this case, the value of the Deviance in line 1 is the sum of the Deviance in line 2 and G in line 1 and the "1" in the "Powers" column means it enters as a single-term (i.e., linear in the logit).

The second covariate processed is age, AGE, as it had the second largest Wald statistic. This variable is continuous and, as such, it is first modeled using the best two-term fractional polynomial transformation with the powers shown in the last column of line 3, (3, 3), that is $[\text{AGE}^3, \text{AGE}^3 \times \ln(AGE)]$. The partial likelihood ratio test comparing this best two-term fractional polynomial modeling of age to the 10 covariate model that excludes age is, from line 3, $G = 7.269$ which, with 4

Table 4.32 Results from the Cycle 2 Fit of MFP Applied to the GLOW500 Data

Line	Variable	Model	(vs.)	Deviance	G	p	Powers	(vs.)
1	PRIORFRAC	null	lin.	524.264	8.42	0.004*	.	1
2		Final		515.844			1	
3	AGE	null	FP2	529.003	13.744	0.008*	.	3 3
4		lin.		515.844	0.584	0.9	1	
5		Final		515.844			1	
6	RATERISK3	null	lin.	523.740	7.897	0.005*	.	1
7		Final		515.844			1	
8	MOMFRAC	null	lin.	518.899	3.055	0.080*	.	1
9		Final		515.844			1	
10	RATERISK2	null	lin.	519.360	3.517	0.061*	.	1
11		Final		515.844			1	
12	BMI	null	FP2	515.844	4.433	0.351	.	−2 −2
13		Final		515.844			.	
14	ARMASSIST	null	lin.	515.844	2.41	0.121*	.	1
15		Final		513.434			1	
16	WEIGHT	null	FP2	513.434	3.611	0.461	.	−2 −2
17		Final		513.434			.	
18	HEIGHT	null	FP2	513.434	7.749	0.101*	.	−2 −2
19		lin.		507.500	1.816	0.612	1	
20		Final		507.500			1	
21	SMOKE	null	lin.	507.500	0.587	0.444	.	1
22		Final		507.500			.	
23	PREMENO	null	lin.	507.500	0.181	0.67	.	1

*$p <$ chosen significance level for inclusion.
†$p <$ chosen significance level for transformation.

degrees of freedom, yields $\Pr[\chi^2(4) \geq 7.629] = 0.106$. Since this is significant at the 0.15 level the two-term fractional polynomial model is compared to the linear model in line 4. The partial likelihood ratio test in line 4 is $G = 0.598$, which with 3 degrees of freedom, yields $p = 0.897$. Since two different parameterizations of age are being compared, the p-value is compared to $\alpha_2 = 0.05$ and the test is not significant. Hence, there is no further modeling of age with the final model, age linear, given in line 5. Had the two-term fractional polynomial model been significantly different from the linear model the best one-term fractional polynomial model would have been found and compared to the two-term model, again at the α_2 level of significance.

The next three variables examined are the dichotomous covariates RATERISK3, MOMFRAC and RATERISK2. Each is significant at the $\alpha_1 = 0.15$ level and thus will be retained in the model fit at cycle 2.

The covariate BMI is next examined in line 12. The partial likelihood ratio test comparing the best two-term fractional polynomial model, powers $(-2, 1)$, with the model that excludes BMI is $G = 4.524$ which, with four degrees of freedom, results in $p = 0.340$. This is not significant at the $\alpha_1 = 0.15$ level, so BMI is not

included in the model fit in cycle 2. The remaining five covariates, ARMASSIST, WEIGHT, HEIGHT, SMOKE, and PREMENO, are individually not significant at the 0.15 level. However, the significance of the partial likelihood ratio tests for ARMASSIST and HEIGHT have p-values that are close to the threshold of 0.15. It is possible that these two could be selected for inclusion in cycle 2 when a smaller model is fit.

The model fit at cycle two contains the first five covariates in Table 4.31, namely PRIORFRAC, AGE, RATERISK3, MOMFRAC and RATERISK2. The results of cycle 2 are shown in Table 4.32.

The results in the first 13 lines of Table 4.32 are similar to those in Table 4.31 for these covariates. The difference between these tables is that the partial likelihood ratio tests in Table 4.32 are based now, not on the full 11 covariate model, but on a five covariate model. In line 14 we see that ARMASSIST contributes to the model at the 0.15 level with $p = 0.121$. In line 18 we see that HEIGHT also is significant ($p = 0.101$). Hence at the next cycle a seven covariate model is fit: the five in lines 1–10 plus ARMASSIST and HEIGHT. The decisions based on this fit are similar to those in Table 4.31. Hence the procedure converges at cycle 2.

We note that application of mfp(0.15, 0.05) to the GLOW500 data yields exactly the same model identified by purposeful selection and stepwise selection. The model obtained using best subsets was similar but selected BMI and WEIGHT in place of HEIGHT. Much of the congruence between the various methods can be attributed to the fact that, in this example, none of the continuous covariates had significant fractional polynomial transformations.

To provide an example when continuous covariates are transformed we apply mfp(0.15, 0.05) to the Burn Study data analyzed in Section 4.2. The covariates modeled (see Table 1.9) are total burn surface (TBSA), age (AGE), burn involving an inhalation injury (INH_INJ), race (RACE, 0 = non-white, 1 = white), burn involving a flame (FLAME) and gender (GENDER, 0 = female, 1 = male). The procedure converged in two cycles and we show the results from cycle 2 in Table 4.33.

The results for TBSA and AGE in Table 4.33 provide good examples of when fractional polynomial transformations are found to be significant with the mfp method.

The first variable processed is TBSA. The results in line 1 show that the two-term fractional polynomial model, powers $(-2, 0.5)$, is significant when compared to the model not containing TBSA with $p < 0.001$. Hence the procedure now compares the two-term fractional polynomial model to the model linear in TBSA in line 2. With $p = 0.001$, the test is significant at the $\alpha_2 = 0.05$ level, as indicated by the "+". Next, the two-term model is compared to the best one-term fractional polynomial model [power (0.5)]. The significance level, computed with 2 degrees of freedom is $p = 0.520$. Since this is not significant at the 0.05 level the process stops and the one-term fractional polynomial model is the final model for TBSA, shown in line 4.

The results for age in lines 5–8 are similar to those for TBSA in that the final model is the one-term fractional polynomial model with power (2). The results for

Table 4.33 Results from the Cycle 2 Fit of MFP Applied to the Burn Data

Line	Variable	Model	(vs.)	Deviance	G	p	Powers	(vs.)
1	TBSA	null	FP2	528.892	208.263	<0.001*	.	−2 .5
2		lin.		336.842	16.213	0.001†	1	
3		FP1		321.935	1.306	0.52	0.5	
4		Final		321.935			0.5	
5	AGE	null	FP2	505.022	184.862	<0.001*	.	1 1
6		lin.		329.589	9.429	0.024†	1	
7		FP1		321.935	1.775	0.412	2	
8		Final		321.935			2	
9	INH_INJ	null	lin.	339.521	17.586	0.000*	.	1
10		Final		321.935			1	
11	RACE	null	lin.	325.869	3.934	0.047*	.	1
12		Final		321.935			1	
13	FLAME	null	lin.	321.935	1.838	0.175	.	1
14		Final		321.935			.	
15	GENDER	null	lin.	321.935	0.129	0.719	.	1
16		Final		321.935			.	

*$p <$ chosen significance level for inclusion.
†$p <$ chosen significance level for transformation.

inhalation injury in lines 9 and 10 show it is significant as is race in lines 11 and 12. The last two covariates processed, FLAME and GENDER, do not contribute to the model with significance levels of $p = 0.175$ and $p = 0.719$ respectively. As noted the mfp procedure converged at two cycles. The resulting model with four covariates, $\sqrt{TBSA}$, AGE^2, INH_INJ and RACE, is the same model initially obtained using purposeful selection in Section 4.2. As we noted there, we added AGE to the model for purposes of ease of interpretation, even though its coefficient was not significant when added to the model containing AGE^2.

The mfp(α_1, α_2) method is clearly an extremely powerful analytic modeling tool, which on the surface, would appear to relieve the analyst of having to think too hard about model content. This is not the case, of course. We recommend that, if one uses this approach then its model be considered as a suggestion for a possible main effects model, much in the way that stepwise and best subsets identify possible models. The model needs a thorough evaluation to be sure all covariates and transformations make clinical sense, that transformations are not caused by a few extreme observations and, importantly, that excluded covariates are not confounders of model covariate estimates of effect. We highly recommend that you spend time with Royston and Sauerbrei (2008, Chapter 6), Sauerbrei et al. (2006) and the host of other excellent papers cited that describe in detail, the development and use of both fractional polynomials and the mfp(α_1, α_2) procedure.

In summary, stepwise, best subsets and multivariable fractional polynomials have their place as covariate selection methods, but it is always the responsibility of the user to choose the content and form of the final model.

4.4 NUMERICAL PROBLEMS

In previous chapters we have occasionally mentioned various numerical problems that can occur when fitting a logistic regression model. These problems are caused by certain structures in the data coupled with the lack of appropriate checks in some logistic regression software. The goal of this section is to illustrate these structures in certain simple situations and illustrate what can happen when the logistic regression model is fit to such data. The issue here is not one of model correctness or specification, but the effect certain data patterns have on the computation of parameter estimates. Some of these problems are due to "thin" data, namely not enough outcomes, usually $y = 1$, and/or small frequencies for a categorical covariate. In some settings use of exact logistic regression methods, discussed in Section 10.3 can provide correctly estimated coefficients and standard errors. In this section we present results from running various example data in several different packages.

For some of the examples we do not state which package produced the results. The reason is that packages are revised and the results we get in one version with these ill conditioned data might well change in the next release. Also different packages might provide different output from the same ill conditioned data. The point of the examples is to learn the numerical signs and symptoms that indicate a numerical problem in the data.

Perhaps the simplest and thus most obvious situation is when we have a frequency of zero in a contingency table. An example of such a contingency table is given in Table 4.34. The estimated odds ratios and log-odds ratios using the first level of the covariate as the reference group are given in the first two rows below the table. The point estimate of the odds ratios for level 3 versus level 1 is infinite since all subjects at level 3 responded. The results of fitting a logistic regression model to these data are given in the last two rows. The estimated coefficient in the first column is the intercept coefficient. The particular package used does not really matter as many, but not all, packages produce similar output. One program that does identify the problem is STATA. It provides an error message that $x = 3$ perfectly predicts the outcome and the design variable for $x = 3$ is not included

Table 4.34 A Contingency Table with a Zero Cell Count and the Results of Fitting a Logistic Regression Model to these Data

Outcome / x	1	2	3	Total
1	7	12	20	39
0	13	8	0	21
Total	20	20	20	60
$\widehat{\text{OR}}$	1	2.79	inf	
$\ln(\widehat{\text{OR}})$	0	1.03	inf	
$\hat{\beta}$	−0.62	1.03	11.7	
$\widehat{\text{SE}}$	0.47	0.65	34.9	

in the fit of the model. Other programs may or may not provide some sort of error message indicating that convergence was not obtained or that the maximum number of iterations was used. What is rather obvious, and the tip-off that there is a problem with the model, is the large estimated coefficient for the second design variable and especially its large estimated standard error.

A common practice to avoid having an undefined point estimate is to add one-half to each of the cell counts. Adding one-half may allow us to move forward with the analysis of a single contingency table, but such a simplistic remedy is rarely satisfactory with a more complex data set.

As a slightly more complex example we consider the stratified 2 by 2 tables shown in Table 4.35. The stratum-specific point estimates of the odds ratios are provided below each 2 by 2 table. The results of fitting a series of logistic regression models are provided in Table 4.36.

In the case of the data shown in Table 4.35 we do not encounter problems until we include the stratum z, by risk factor x, and interaction terms, $x \times z\_2$ and $x \times z\_3$ in the model. The addition of the interaction terms results in a model that is equivalent to fitting a model with a single categorical variable with six levels, one for each column in Table 4.35. Thus, in a sense, the problem encountered when we include the interaction is the same one illustrated in Table 4.34. As was the case when fitting a model to the data in Table 4.34, the presence of a zero cell count is manifested by an unbelievably large estimated coefficient and estimated standard error.

The presence of a zero cell count should be detected during the univariable screening of the data. Knowing that the zero cell count is going to cause problems

Table 4.35 Stratified 2 by 2 Contingency Tables with a Zero Cell Count Within One Stratum

Stratum (z)	1		2		3	
Outcome / x	1	0	1	0	1	0
1	5	2	10	2	15	1
0	5	8	2	6	0	4
Total	10	10	12	8	15	5
$\widehat{\text{OR}}$	4	15	inf			

Table 4.36 Results of Fitting Logistic Regression Models to the Data in Table 4.35

Model	1		2	
Variable	Coeff.	Std. Err.	Coeff.	Std. Err.
x	2.77	0.72	1.39	1.01
z_2	1.19	0.81	0.29	1.14
z_3	2.04	0.89	0.00	1.37
$x \times z\_2$			1.32	1.51
$x \times z\_3$			11.54	50.22
Constant	−2.32	0.77	−1.39	0.79

in the modeling stage of the analysis we could collapse the categories of the variable in a meaningful way to eliminate it, eliminate the category completely, or if the variable is at least ordinal scale, treat it as continuous.

The type of zero cell count illustrated in Table 4.35 results from spreading the data over too many cells. This problem is not likely to occur until we begin to include interactions in the model. When it does occur, we should examine the three-way contingency table equivalent to the one shown in Table 4.35. The unstable results prevent us from determining whether, in fact, the interaction is important. To assess the interaction we first need to eliminate the zero cell count. One way to do this is by collapsing categories of the stratification variable. For example, in Table 4.35 we might decide that values of $z = 2$ and $z = 3$ are similar enough to pool them. The stratified analysis would then have two 2 by 2 tables the second of which results from pooling the tables for $z = 2$ and $z = 3$. A second approach is to define a new variable equal to the combination of the stratification variable and the risk factor and to pool over levels of this variable and model it as a main effect variable. Using Table 4.35 as an example, we would have a variable with six levels corresponding to the six columns in the table. We could collapse levels five and six together. Another pooling strategy would be to pool levels three and five, and four and six. This pooling strategy is equivalent to collapsing over levels of the stratification variable. The net effect is the loss of degrees of freedom commensurate with the amount of pooling. Twice the difference in the log-likelihood for the main effects only model, and the model with the modified interaction term added, provides a statistic for the significance of the coefficients for the modified interaction term.

The fitted models shown in Tables 4.34 and 4.36 resulted in large estimated coefficients and estimated standard errors. In some examples we have encountered, the magnitude of the estimated coefficient was not large enough to suspect a numerical problem, but the estimated standard error always was. Hence, we believe that the best indicator of a numerical problem in logistic regression is the estimated standard error. In general, any time that the estimated standard error of an estimated coefficient is large relative to the point estimate, we should suspect the presence of one of the data structures described in this section.

A second type of numerical problem occurs when a collection of the covariates completely separates the outcome groups or, in the terminology of discriminant analysis, the covariates discriminate perfectly. For example, suppose that the age of every subject with the outcome present was greater than 50 and the age of all subjects with the outcome absent was less than 49. Thus, if we know the age of a subject we know with certainty the value of the outcome variable. In this situation there is no overlap in the distribution of the covariates between the two outcome groups. This type of data has been shown by Bryson and Johnson (1981) to have the property of monotone likelihood. The net result is that the maximum likelihood estimates do not exist [see Albert and Anderson (1984); Santner and Duffy (1986)]. In order to have finite maximum likelihood estimates we must have some overlap in the distribution of the covariates in the model.

Table 4.37 Estimated Slope ($\hat{\beta}_x$), Constant ($\hat{\beta}_0$), and Estimated Standard Errors ($\widehat{\text{SE}}$) when the Data Have Complete Separation, Quasicomplete Separation, and Overlap

Estimates/x_6	5.5	6.0	6.05	6.10	6.15	6.20	8.0
$\hat{\beta}_x$	20.3	7.5	3.7	3.0	2.6	2.3	0.2
$\widehat{\text{SE}}$	36.0	42.4	6.3	4.4	3.6	3.0	0.7
$\hat{\beta}_0$	−116.6	−44.0	−22.2	−17.9	−15.3	−13.5	−0.1
$\widehat{\text{SE}}$	208.1	254.3	38.2	27.1	22.1	189.1	5.8

A simple example illustrates the problem of complete separation and the results of fitting logistic regression models to such data. Suppose we have the following 12 pairs of covariate and outcome, (x, y) : (1,0), (2,0), (3,0), (4,0), (5,0), ($x_6 = 5.5$, or 6.0, or 6.05, or 6.1, or 6.2, or 8.0, $y_6 = 0$), (6,1), (7,1), (8,1), (9,1), (10,1), (11,1). The results of fitting logistic regression models when x_6 takes on one of the values 5.5, 6.0, 6.05, 6.1, 6.2, or 8, using SAS version 9.2 are given in Table 4.37. When we use $x_6 = 5.5$ we have complete separation and all estimated parameters are huge, since the maximum likelihood estimates do not exist. SAS provides a warning but at the same time provides the values of the estimates at the last iteration, leaving the ultimate decision about how to handle the output to the user. Similar behavior occurs when the value of $x_6 = 6.0$ is used. SAS notes this fact and again provides estimates. When overlap is at a single or a few tied values the configuration was termed by Albert and Anderson (1984) as quasi complete separation. As the value of x_6 takes on values greater than 6.0 the overlap becomes greater and the estimated parameters and standard errors begin to attain more reasonable values. The sensitivity of the fit to the overlap depends on the sample size and the range of the covariate. The tip-off that something is amiss is, as in the case of the zero cell count, the very large estimated coefficients and especially the large estimated standard errors. Other programs, including STATA, do not provide output when there is complete or quasicomplete separation, for example, $x_6 = 5.5$ or $x_6 = 6$. In the remaining cases STATA and SAS produce similar results.

The occurrence of complete separation in practice depends on the sample size, the number of subjects with the outcome present, and the number of variables included in the model. For example, suppose we have a sample of 25 subjects and only five have the outcome present. The chance that the main effects model demonstrates complete separation increases with the number of variables we include in the model. Thus, the modeling strategy that includes all variables in the model is particularly sensitive to complete separation. Albert and Anderson (1984) and Santner and Duffy (1986) provide rather complicated diagnostic procedures for determining whether a set of data displays complete or quasicomplete separation. Albert and Anderson (1984) recommend that in the absence of their diagnostic, if one looks at the estimated standard errors and if these tend to increase substantially with each iteration of the fit, then one can suspect the presence of complete separation. As

Table 4.38 Data Displaying Near Collinearity Among the Independent Variables and Constant

Subject	x_1	x_2	x_3	y
1	0.225	0.231	1.026	0
2	0.487	0.489	1.022	1
3	−1.080	−1.070	1.074	0
4	−0.870	−0.870	1.091	0
5	−0.580	−0.570	1.095	0
6	−0.640	−0.640	1.010	0
7	1.614	1.619	1.087	0
8	0.352	0.355	1.095	1
9	−1.025	−1.018	1.008	0
10	0.929	0.937	1.057	1

noted in Chapter 3 the easiest way to address complete separation is to use some careful univariable analyses. The occurrence of complete separation is not likely to be of great clinical importance as it is usually a numerical coincidence rather than describing some important clinical phenomenon. It is a problem we must work around.

As is the case in linear regression, model fitting via logistic regression is also sensitive to collinearities among the independent variables in the model. Most software packages have some sort of diagnostic check, like the tolerance test employed in linear regression. Nevertheless it is possible for variables to pass these tests and have the program run, but yield output that is clearly nonsense. As a simple example, we fit logistic regression models using STATA to the data displayed in Table 4.20. In the table $x_1 \sim N(0, 1)$ and the outcome variable was generated by comparing a $U(0, 1)$ variate, u, to the true probability $\pi(x_1) = e^{x_1}/(1 + e^{x_1})$ as follows: if $u < \pi(x_1)$ then $y = 1$, otherwise $y = 0$. The notation $N(0, 1)$ indicates a random variable following the standard normal (mean = 0, variance = 1) distribution and $U(a, b)$ indicates a random variable following the uniform distribution on the interval $[a, b]$. The other variables were generated from x_1 and the constant as follows: $x_2 = x_1 + U(0, 0.1)$ and $x_3 = 1 + U(0, 0.01)$. Thus, x_1 and x_2 are highly correlated and x_3 is nearly collinear with the constant term. The results of fitting logistic regression models to various subsets of the variables shown in Table 4.38 are presented in Table 4.39.

The model that includes the highly correlated variables x_1 and x_2 has both very large estimated slope coefficients and estimated standard errors. For the model containing x_3 we see that the estimated coefficients are of reasonable magnitude but the estimated standard errors are much larger than we would expect. The model containing all variables is a composite of the results of the other models. In all cases the tip-off for a problem comes from the aberrantly large estimated standard errors.

In a more complicated data set, an analysis of the associations among the covariates using a collinearity analysis similar to that performed in linear regression should be helpful in identifying the dependencies among the covariates. Belsley

Table 4.39 Estimated Coefficients and Standard Errors from Fitting Logistic Regression Models to the Data in Table 4.38

Var.	Coeff.	Std. Err.	Coeff.	Std. Err.	Coeff.	Std. Err.	Coeff.	Std. Err.
x_1	1.4	1.0	104.2	256.2			79.8	272.6
x_2			−103.4	256.0			−78.3	272.5
x_3					1.8	20.0	−11.1	206.6
Cons.	−1.0	0.8	−0.3	1.3	−2.7	21.1	11.4	27.8

et al. (1980) discuss a number of methods that are implemented in many linear regression packages. One would normally not employ such an in-depth investigation of the covariates unless there was evidence of degradation in the fit similar to that shown in Table 4.39. An alternative is to use the ridge regression methods proposed by Schaefer (1986).

In general, the numerical problems of a zero cell count, complete separation, and collinearity, are manifested by extraordinarily large estimated standard errors and sometimes by a large estimated coefficient as well. New users and those without much computer experience are especially cautioned to look at their results carefully for evidence of numerical problems. In many settings all is not lost. Heinze and Schemper (2002) and Heinze (2006) discuss and illustrate the use of methods that can produce valid parameter estimates and confidence intervals with data containing zero frequency cells and/or separation. These methods include exact logistic regression and penalized likelihood methods, which we discuss and illustrate in Section 10.3.

EXERCISES

1. Show algebraically and with a numerical example of your choice that the restricted cubic spline functions in equation (4.3) meet at the three knots.

2. In the modeling of the GLOW500 data using purposeful selection age was modeled as linear in the logit. We noted that the estimated coefficients for the quartile design variables for age in Table 4.10 suggested an alternative parameterization: using the design variable for the fourth quartile AGE_4. This parameterization of age was not pursued further. Proceed with purposeful selection using AGE_4. To save time, assume that your main effects model is the one in Table 4.9 but with AGE replaced by AGE_4. Compare your model to the one in Table 4.15 that resulted when age was modeled as linear in the logit. Which model do you think is the better one for estimating risk factors for fracture?

3. In the modeling of the Burn Injury data questions came up as to how to model age. There were essentially three choices: linear (power 1), quadratic (powers 1 and 2) and the best fractional polynomial model (power 2). In the text we

proceeded with power 2. Perform selection of interactions for the other two parameterizations and save your work for an exercise on model evaluation in Chapter 5.

4. Demonstrate best subset selection of interactions by beginning with the main effects model from the GLOW500 data.

5. The restricted cubic spline analysis for age in the Burn Injury Study shown in Table 4.21 used four knots at the 5th, 35th, 65th, and 95th percentiles (see Table 4.1). Verify that spline functions formed from these four knots provide a better model than using three or five knots placed at the respective percentiles in Table 4.1.

6. Consider the data from the Myopia Study described in Section 1.6.6 whose variables are described in Table 1.10. The binary outcome variable is MYOPIC ($0 = \text{Yes},\ 1 = \text{No}$). Consider as independent variables all others in Table 1.10 except spherical equivalent refraction (SPHEQ) as it is used to define the outcome variable, the composite of near-work hours (DIOPTERHR) and study year (STUDYYEAR).

(a) Use purposeful selection to obtain what you feel is the best model for estimating the effect of the risk factors on myopia. This analysis must include identification of the scale in the logit of all continuous covariates and selection of interactions. Assume that all possible interactions among your main effects are clinically reasonable.

(b) Repeat problem 6(a) using stepwise selection of covariates (main effects and then interactions among main effects forcing in the main effects).

(c) Repeat problem 6(a) using best subset selection of covariates with Mallows' C_q (main effects and then interactions among main effects forcing in the main effects).

(d) Repeat problem 6(a) using multivariable fractional polynomial selection of main effects followed by purposeful selection of interactions.