CHAPTER 1

Introduction

1.1 REAL-LIFE PROBLEMS AS MOTIVATION FOR MODEL BUILDING

Data are collected in all areas of life. In research, data on several variables may be collected to investigate the interrelationships among them, or to determine factors which affect an outcome of interest. An example from medicine is the relationship between the survival time of a patient following treatment for cancer and potentially influential variables (known in this context as prognostic factors), such as age, size of the tumour, its aggressiveness (grade), and so on. Often, effects of more than 10 potentially influential variables must be considered simultaneously in a single model. Our main emphasis is on examples from the health sciences, particularly clinical epidemiology. We discuss statistical methods to develop a model which best tries to answer specific questions in the framework of regression models. Although nearly all of the examples discussed have a background in the health sciences, the methods we describe are also highly relevant in other areas where multivariable regression models with continuous variables are developed.

1.1.1 Many Candidate Models

With today's fast computers and sophisticated statistical software, it is nearly trivial to fit almost any given model to data. It is important to remember that all models are based on several more or less explicit assumptions, which may correspond more or less well to unknown biological mechanisms. Therefore, finding a good model is challenging. By 'good' we mean a model that is satisfactory and interpretable from the subject-matter point of view, robust with respect to minor variations of the current data, predictive in new data, and parsimonious. Therefore, a good model should be useful beyond the dataset on which it was created.

In fitting regression models, data analysts are frequently faced with many explanatory variables, any or all of which may to some extent affect an outcome variable. If the number of variables is large, then a smaller model seems preferable. An aim of the analysis, therefore, is the selection of a subset of 'important' variables that impact on the outcome. For this task, techniques for stepwise selection of variables are available in many statistical packages and are often used in practical applications (Miller, 2002). Despite the importance of methods of variable selection and the enormous attention paid to the topic, their properties are not well understood. All are criticized in the statistical literature.

Multivariable Model-Building Patrick Royston, Willi Sauerbrei

1.1.2 Functional Form for Continuous Predictors

A second obstacle to model building is how to deal with nonlinearity in the relation between the outcome variable and a continuous or ordered predictor. Traditionally, such predictors are entered into stepwise selection procedures as linear terms or as dummy variables obtained after grouping. The assumption of linearity may be incorrect. Categorization introduces problems of defining cutpoint(s) (Altman et al., 1994), overparametrization and loss of efficiency (Morgan and Elashoff, 1986; Lagakos, 1988). In any case, a cutpoint model is an unrealistic way to describe a smooth relationship between a predictor and an outcome variable.

An alternative approach is to keep the variable continuous and allow some form of nonlinearity. Hitherto, quadratic or cubic polynomials have been used, but the range of curve shapes afforded by conventional low-order polynomials is limited. Box and Tidwell (1962) propose a method of determining a power transform of a predictor. A more general family of parametric models, proposed by Royston and Altman (1994), is based on fractional polynomial (FP) functions and can be traced back to Box and Tidwell's (1962) approach. Here, one, two or more terms of the form x^p are fitted, the exponents p being chosen from a small, preselected set of integer and noninteger values. FP functions encompass conventional polynomials as a special case.

To illustrate the main issues considered in our book, we start with two characteristic examples. In the first example, we consider a simple regression model for a continuous outcome variable with a single, continuous covariate. In the second example we illustrate several approaches to modelling the simultaneous effect of seven potential prognostic factors on a survival-time outcome with censoring. As with most real-life research questions, the prognostic factors are measured on different scales and are correlated. Finding satisfactory multivariable models which include continuous predictors is still a great challenge and is the main emphasis of our book.

1.1.3 Example 1: Continuous Response

We start with a simple illustration of how researchers try to cope with the problems of modelling a continuous covariate in real-life studies. Luke et al. (1997) described the relationship between percentage body fat content (`pbfm`) and body-mass index (`bmi`) in samples of black people from three countries (Nigeria, Jamaica and the USA). See Appendix A.2.1 for further details. The authors aimed to find out how well `bmi` predicted `pbfm`.

'Standard' Analysis (Polynomial Regression)

Percentage of body fat and body-mass index were highly correlated, with Spearman correlation coefficient $r_S(\texttt{bmi}, \texttt{pbfm}) = 0.92$. The authors stated that the relationship between `pbfm` and `bmi` was 'quadratic in all groups except Nigerian men, in whom it was linear'. No indication was given as to how the quadratic (or linear) model was arrived at. The left panel of Figure 1.1 shows the raw data and the authors' fitted quadratic curve for the subsample of 326 females from the USA. Although the fit looks reasonable, we see some minor lack of fit at the lower and upper extremes. More important, the quadratic curve turns downwards for $\texttt{bmi} > 50\,\text{kg}\,\text{m}^{-2}$, which makes little scientific sense. We would not expect the fattest women to have a lower body fat percentage than those slightly less obese. Quadratic functions always have a turning point, but it may or may not occur within the range of the observed data. We return to the critical issue of the scientific plausibility of an estimated function in Section 6.5.4.

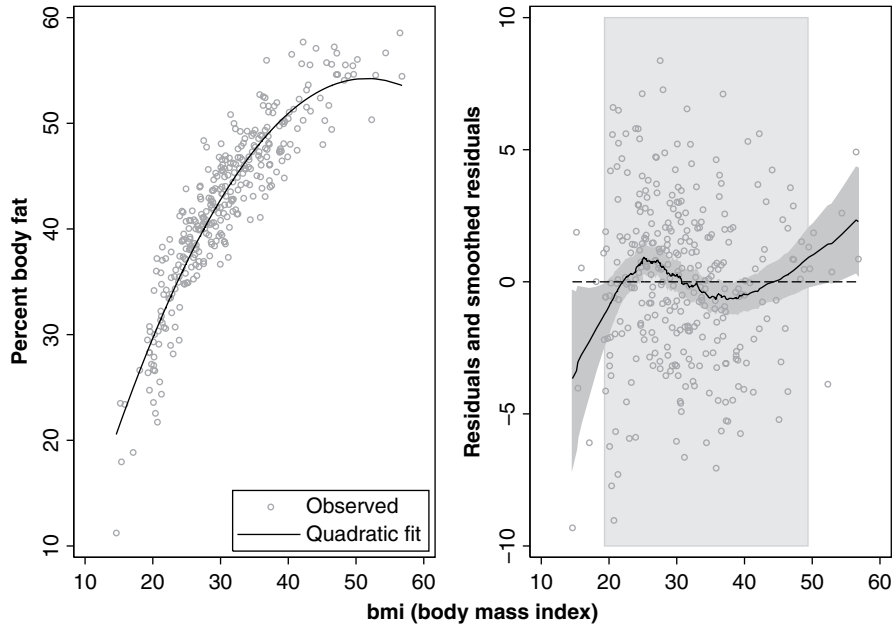

Figure 1.1 Research body-fat data. Left panel: raw values of `pbfm` and the fitted quadratic curve. Right panel: raw and smoothed residuals with 95% confidence interval (CI; see section for further explanation). Lack of fit of the quadratic is seen in the right panel.

Further evidence of lack of fit is seen in the right panel of Figure 1.1, which shows smoothed residuals from the quadratic curve together with a 95% pointwise CI. We used a locally linear (running-line) smoother here (Sasieni and Royston, 1998). A pattern somewhat characteristic of a cubic polynomial is apparent in the smoothed mean residual; indeed, adding a cubic term in `bmi` to the model improves the fit significantly ($P = 0.0001$). A quartic term is not significant at $P < 0.05$, so conventionally one would stop elaborating the model there.

Fractional Polynomial Functions

As an alternative, we also selected a curve from the family of degree-one FPs, also known as FP1 functions. FP1 functions are an extension of an *ad hoc* approach often taken by applied statisticians in the past and examined in some detail by John Tukey (Tukey, 1957; Mosteller and Tukey, 1977). A power p or logarithmic transformation is applied to a predictor x, giving a model whose systematic part is $\beta_0 + \beta_1 x^p$ or $\beta_0 + \beta_1 \ln x$. The power p for FP1 functions is restricted to the predefined set $S = \{-2, -1, -0.5, 0, 0.5, 1, 2, 3\}$ proposed by Royston and Altman (1994) for practical use. FP1 functions are easily extended to higher order FPs (FP2, FP3, ...) in which combinations of power transformations of x are used, resulting in a family of flexible functions. The full definition is given in Section 4.3. Selection of the best-fitting model is discussed in Sections 4.8 and 4.10.

The best-fitting FP1 function for the research body-fat data, among the eight transformations in set S, turns out to have power $p = -1$, for which the formula is $\beta_0 + \beta_1 x^{-1}$. The best-fitting FP2 function has powers $(-2, -1)$, for which the formula is $\beta_0 + \beta_1 x^{-2} + \beta_2 x^{-1}$. For comparison with traditional approaches, the fit of five models (linear, quadratic, cubic, FP1, FP2) is summarized in Table 1.1.

Table 1.1 Goodness-of-fit statistics for five models for the research body-fat data.

Model	d.f.[a]	Deviance D	R^2
Linear	1	1774.87	0.782
Quadratic	2	1647.18	0.853
Cubic	3	1630.70	0.860
FP1(−1)	2	1629.02	0.861
FP2(−2, −1)	4	1627.31	0.862

[a] Degrees of freedom of the (fractional) polynomial terms.

In terms of the proportion of variation explained R^2, there is little difference between the quadratic, cubic, FP1 and FP2 models, whereas the linear model is inferior. Informally, in terms of the deviance $D = -2l$ (where l is the maximized log likelihood), the cubic, FP1 and FP2 models are superior to the quadratic but differ little among themselves. It turns out that the FP2 model does not fit significantly better than the FP1 ($P = 0.4$; see Section 4.10.3). Figure 1.2 shows that smoothed residuals for the cubic and FP1 models appear roughly random. Note that FP1 models are by definition monotonic (i.e. have no turning point – see Section 4.4); so, provided the fit is adequate, they are a good choice in this type of example. Figure 1.3 shows the fitted curves from the models in Table 1.1.

The main lessons from this example are the benefits of a systematic approach to model selection (here, selection of the function) and of the need to assess the results critically, both

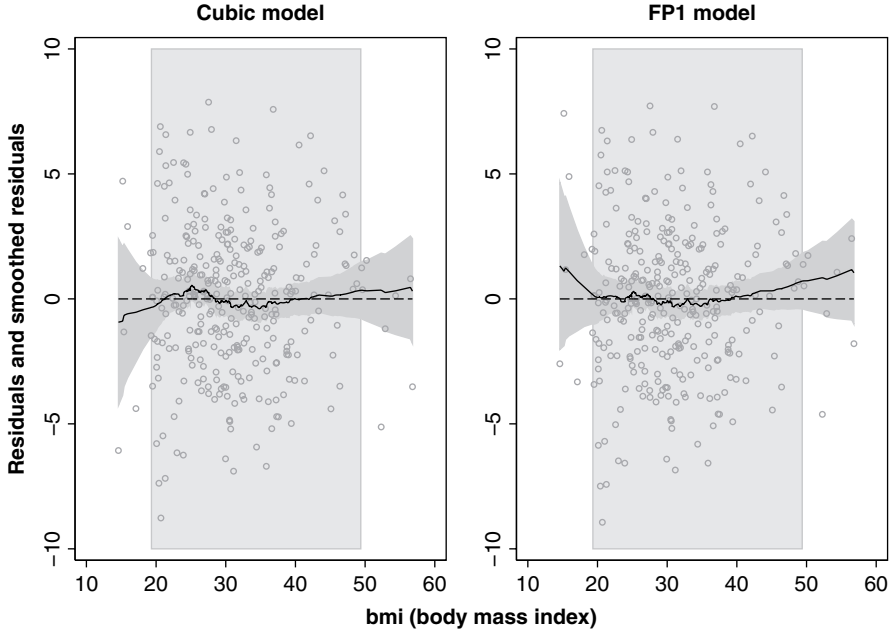

Figure 1.2 Research body-fat data. Left and right panels show residuals and smoothed residuals with 95% CIs from the cubic and FP1 models respectively.

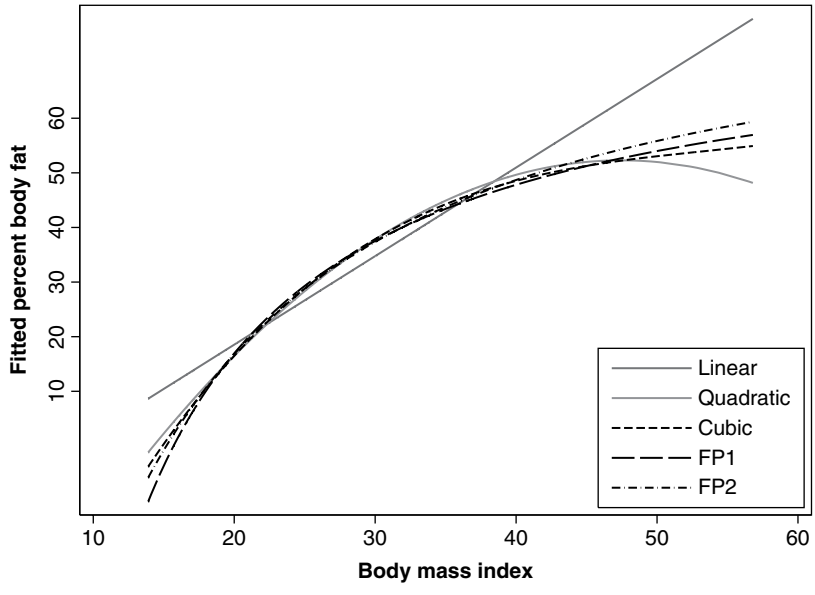

Figure 1.3 Research body-fat data. Fitted lines for five models. For raw data, see Figure 1.1.

from statistical and subject-matter perspectives. A plot such as Figure 1.1 clearly shows that a quadratic fits reasonably well, and equally clearly that it is not an ideal model – the curve is implausible. In purely statistical terms, an examination of smoothed residuals shows the deficiencies of the quadratic model and gives a hint on how to remedy them. The cubic, FP1 and FP2 models are better and fit the data about equally well, but the parsimony principle leads us to prefer the FP1 model. In our experience, simple models are generally more robust and generalize better to new data than complex ones.

1.1.4 Example 2: Multivariable Model for Survival Data

In a study of primary node positive breast cancer patients, seven standard prognostic factors (age, menopausal status, tumour size, tumour grade, number of positive lymph nodes, progesterone receptor status, and oestrogen receptor status) were considered in the development of a prognostic model for recurrence-free survival (RFS) time. For further details, see Appendix A.2.2. Figure 1.4 shows a Kaplan–Meier survival plot. Median RFS time was 4.9 years (95% CI, 4.2 to 5.5 years). Cox proportional hazards modelling (see Section 1.3.3) was used to investigate the simultaneous effect of the factors on the relative hazard of recurrence.

For continuous covariates, we used two standard methods. First, we assumed a linear relationship between the factor and the log relative hazard of an event (the 'linear approach'). Second, we categorized the factor into two or three groups according to predefined cutpoints (the 'step approach'). FPs were used as an additional method (the 'FP approach').

Univariate Models for `age`

To illustrate differences between the methods, we first consider the prognostic effect of age (`age`) in univariate analysis. Whether age is really a prognostic factor was controversial at the time of the original analysis of the GBSG study in the mid 1990s (see discussions in

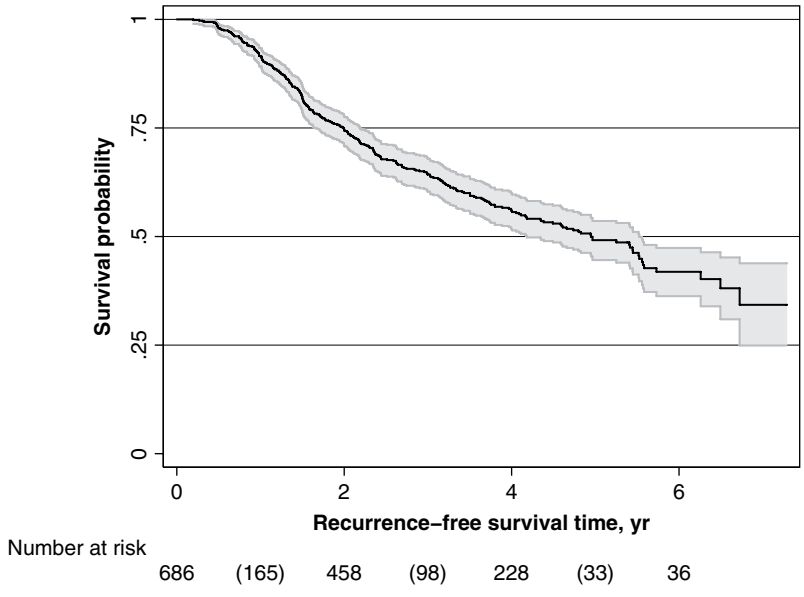

Figure 1.4 German Breast Cancer Study Group (GBSG) breast cancer data. Kaplan–Meier plot of RFS probabilities with 95% pointwise confidence band. Numbers at risk and (in parentheses) failing in each two-yearly interval are tabulated beneath the plot.

Sauerbrei et al. (1997; 1999)). In Table 1.2 we give deviance differences from the null model and P-values for the univariate effect of `age`, according to several types of model.

According to the linear approach, `age` has no apparent effect on RFS ($P > 0.4$) and the slope $\hat{\beta}$ on `age` is small (see Figure 1.5). As is usual with the Cox model, the effect of `age` is expressed as the log relative hazard with respect to an unspecified baseline hazard function. With a quadratic model, the P-value is 0.01. With the two predefined cutpoints used in the original analysis, the P-value is 0.15. Following discussions during the last decade and various data-driven searches for 'optimal' cutpoints, some researchers have argued in favor of 40 years as a good cutpoint for age. The analysis based on this cutpoint suggests an effect of `age` with

Table 1.2 GBSG breast cancer data. Deviance differences from the null model and P-values from univariate models assuming different functional forms.

Model	Deviance difference	d.f.	P-value
Linear	0.6	1	0.45
Quadratic	9.0	2	0.011
FP1(−1)	6.4	2	0.041
FP2(−2, −1)	17.6	4	0.002
Categorized (1)[a]	3.8	2	0.15
Categorized (2)[b]	5.3	1	0.021

[a] Predefined cutpoints 45 and 60 years.
[b] Selected cutpoint 40 years.

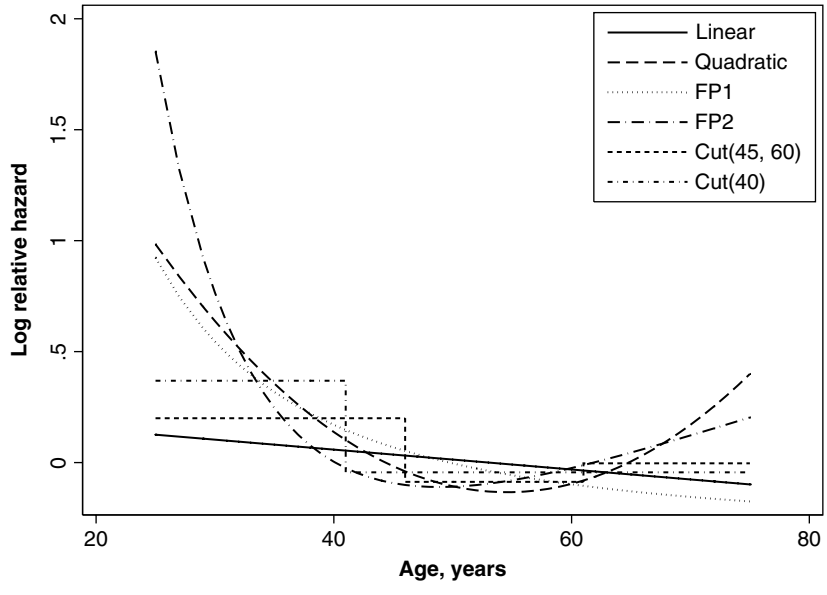

Figure 1.5 GBSG breast cancer data. Fitted functions from six Cox models for `age`.

a P-value of 0.02. Interpretation is obviously difficult. The model implies that patients aged 30 or 39 years have the same risk, whereas the risk decreases substantially (estimated relative hazard 0.66) for a patient aged 41 years. A patient of age 65 years has the same risk as a 41-year old, according to this model.

Use of the transformation $\texttt{age}^{-2}$ provides evidence that age really is a prognostic factor. This transformation gave the best fit within the class of eight FP1 transformations proposed by Royston and Altman (1994). Within the more flexible class of FP2 transformations, a further improvement in fit is obtained with the FP2 function $\hat{\beta}_1\texttt{age}^{-2} + \hat{\beta}_2\texttt{age}^{-0.5}$. The overall P-value for `age` in this model is 0.002. The functions from the six models shown in Figure 1.5 display major differences.

Multivariable Model-Building

So far, we have illustrated several problems when investigating the effect of one continuous variable. However, in the real world the norm is that several covariates must be considered simultaneously. Although univariate analyses of explanatory variables are a good starting point, ultimately a multivariable analysis is required.

Sauerbrei and Royston (1999) developed three multivariable models for the GBSG data by using backward elimination (BE) with a nominal P-value of 0.05. For continuous variables, they considered linear functions, step functions with predefined cutpoints, and functions from the FP class. We defined two dummy variables, `gradd1` and `gradd2`, from the ordered categorical factor `grade`. `gradd1`, `nodes` and `pgr` (progesterone receptor) were selected in all three models. The multivariable FP (MFP) procedure identified `age` as an additional prognostic factor. Its fitted curve is similar to the FP2 function from the univariate analysis (see Figure 1.5). The deviances for the linear, step and FP models are 3478.5, 3441.6 and 3427.9 respectively, showing that the FP2 model fits best. The lower the deviance is, the better the model fit is.

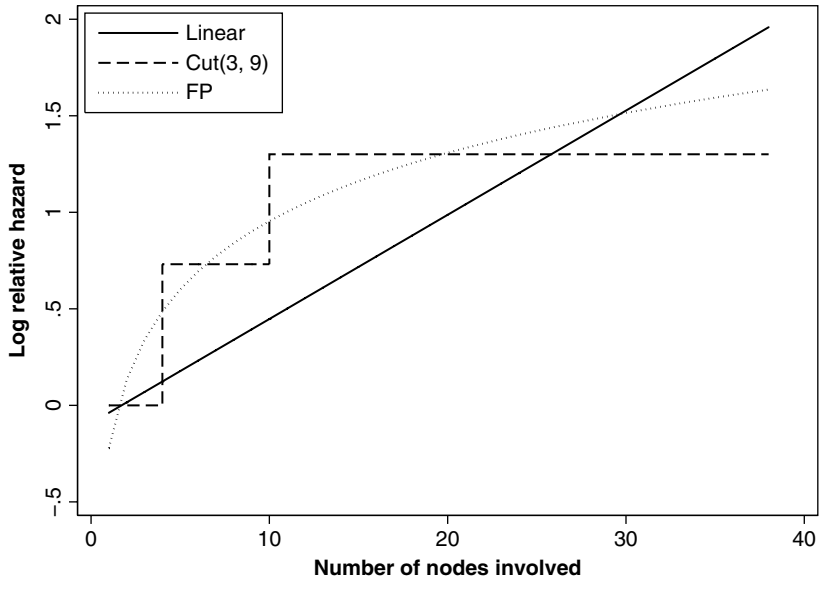

Figure 1.6 GBSG breast cancer data. Fitted functions from three models for `nodes`, estimated within multivariable Cox models. (Adapted from Sauerbrei et al. (1999) with permission from Macmillan Publishers Ltd. British Journal of Cancer, copyright 1999.)

Besides `age`, the models differ in the functional form for `nodes` and `pgr`, whereas `gradd1` had a similar effect. The number of positive nodes has been identified as the most important prognostic factor in many studies of early breast cancer. The hazard of an event increases with the number of nodes. Such medical knowledge may be incorporated in the modelling process by restricting the candidate functions to be monotonic. Here, we allowed only FP1 functions (guaranteed monotonic) for `nodes`, the resulting function being log(`nodes`) – see model II* of Sauerbrei and Royston (1999).

The three functions are shown in Figure 1.6. The step function is a rough approximation to the FP (log) function; with more cutpoints, the approximation would be closer, but 'noisier'. The linear function seriously underestimates the hazard for a small number of nodes and overestimates it for a large number.

The functions derived for the continuous variable `pgr` also differ (see Sauerbrei et al. (1999)). In this example, the three ways of handling continuous variables in a multivariable analysis give different results and raise some general issues that are an important topic of our book.

1.2 ISSUES IN MODELLING CONTINUOUS PREDICTORS

1.2.1 Effects of Assumptions

An assumption of linearity may prevent one from recognizing a strong effect of a variable (e.g. `age`), or to lead one to mismodel the effect (e.g. `nodes`). The popular cutpoint approach introduces several well-known difficulties. Data-dependent determination of the cutpoint, at the most extreme by using the 'optimal' cutpoint approach, results in P-values that are too small and in an overestimation of the effect (Royston and Altman, 1994; Royston et al., 2006).

In other words, severely biased results can be obtained. Predefined cutpoints may not reflect the true, but unknown, functional relationship, and the number of cutpoints required is unknown. Furthermore, cutpoint approaches do not use the full information from the data, and step functions may conflict with biological principles, which demand smooth(er) functions. For further discussion, see Section 3.4.

With the FP approach, much more information from the data is utilized. The question of whether a nonlinear transformation improves the fit of a model is assessed systematically, with further advantages of flexibility in handling continuous variables, predicated on established statistical principles, transparency and ease of use.

1.2.2 Global versus Local Influence Models

Several approaches to modelling continuous covariates other than linear and FP functions are available. In general, it is useful to distinguish between regression functions for a continuous variable with the property of either global or local influence. For a function of x with the global-influence property, the fit at a given value x_0 of x may be relatively unaffected by local perturbations of the response at x_0, but the fit at points distant to x_0 may be affected, perhaps considerably. This property may be regarded by proponents of local-influence models as a fatal flaw; see the discussion and the example given by Hastie and Tibshirani in Royston and Altman (1994). A rigorous definition of the global-influence property has not to our knowledge been framed, but such models are usually 'parametric' in nature. Examples include polynomials, nonlinear models such as exponential and logistic functions, and FPs. By contrast, functions with the local-influence property, including regression splines (de Boer 2001), smoothing splines (Green and Silverman 1994), and kernel-based scatter-plot smoothers such as lowess (Cleveland and Devlin, 1988), are typically 'nonparametric' in character. Perturbation of the response at x_0 usually greatly affects the fit at x_0 but hardly affects it at points distant to x_0. One key argument favouring functions with global influence is their potential for use in future applications and datasets. Without such an aim, functions with local influence might appear the more attractive (Hand and Vinciotti, 2003).

Although FP functions retain the global-influence property, they are much more flexible than polynomials. Indeed, low-dimensional FP curves may provide a satisfactory fit where high-order polynomials fail (Royston and Altman, 1994). FPs are intermediate between polynomials and nonlinear curves. They may be seen as a good compromise between ultra-flexible but potentially unstable local-influence models and the relatively inflexible conventional polynomials. The title of our book makes it clear that our main emphasis is on FPs, but we also compare and discuss results from global- and local-influence models in several examples.

1.2.3 Disadvantages of Fractional Polynomial Modelling

Modelling with FP functions and our MFP approach also has difficulties. Perhaps the most important aspects are insufficient power to detect a nonlinear function and the possible sensitivity to extreme values at either end of the distribution of a covariate. An example of the latter is a 'hook' in the best FP2 function for `nodes` (see model II of Sauerbrei and Royston (1999)). In the analysis in Section 1.1.4, we used only FP1 functions for `nodes`, whereas in the original paper we discussed advantages of working with the preliminary transformation $\texttt{enodes} = \exp(-0.12 \times \texttt{nodes})$, guaranteeing a monotonic function with an asymptote. This transformation was used in all our subsequent analyses of the breast cancer data and is also often used in our book.

The power issue has several aspects. Owing to insufficient sample size (in survival data, too few events), variables with a modest or weak effect may not be selected; or, by default, linear effects may be chosen instead of more realistic nonlinear functions. See Section 4.16 for some simulation results on the topic. Our approach to multivariable model-building may, therefore, add to the problem of low power inherent in all types of statistical modelling in small samples (see also Section 6.9.2). It is questionable whether variable selection makes sense at all in small samples. In medium-sized studies, loss of power is a justifiable cost, balanced by the benefits of combining variable selection with the selection of the functional form for continuous variables.

1.2.4 Controlling Model Complexity

We prefer a simple model unless the data indicate the need for greater complexity. A simple model transfers better to other settings and is more suited to practical use. The simplest dose–response relationship is the linear model, and that is our default option in most modelling situations. We would use an FP model if prior knowledge dictated such a model. In the absence of prior knowledge, we would use an FP model if there was sufficient evidence of nonlinearity within the data. The FP approach contrasts with local regression modelling (e.g. splines, kernel smoothers, etc.), which often starts and ends with a complex model.

1.3 TYPES OF REGRESSION MODEL CONSIDERED

For the types of model that we use, textbooks describing all aspects in great detail are available. Therefore, we introduce the models only very briefly, referring as necessary to other sources. A modern text addressing many detailed issues in model building is Harrell (2001); however, it appears to have been written for people already expert in regression modelling. Other recommendable textbooks on regression analysis, referenced in our book, are Cohen et al. (2003), DeMaris (2004), Vittinghof et al. (2005) and Weisberg (2005). The first two are based in behavioural and social sciences.

Although we work with multiple linear regression models, in our own research and in examples we are more concerned with important generalizations, including logistic and Cox models, and generalized linear models (GLMs), of which logistic regression is a special case. Such types of model are familiar tools nowadays, allowing the analyst to deal flexibly with many different types of response variable.

Methods of variable selection and related aspects have usually been developed and investigated for the multiple linear regression model. However, the methods, or at least their basic ideas, are commonly transferred to GLMs and to models for survival data. Additional difficulties may then arise; for example, obtaining satisfactory definitions of residuals or of equivalents of the proportion of explained variation R^2.

1.3.1 Normal-Errors Regression

For an individual with response y, the multiple linear regression model with normal errors $\varepsilon \sim N(0, \sigma^2)$ and covariate vector $\mathbf{x} = (x_1, \ldots, x_k)$ with k variables, may be written

$$y = E(y) + \varepsilon = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \varepsilon = \beta_0 + \mathbf{x}\beta + \varepsilon \tag{1.1}$$

The linear predictor or 'index', $\eta = \beta_0 + \mathbf{x}\beta$, is an important quantity in multivariable modelling (see also Section 1.3.5). Throughout our book, (1.1) is called the 'normal-errors model'. Although, as presented here, the model is linear in the covariates, models that are nonlinear in $\mathbf{x}$ come under the same heading.

Suppose we have a set of n observations

$$(y_1, \mathbf{x}_1), \ldots, (y_n, \mathbf{x}_n) \equiv (y_1, x_{11}, x_{12}, \ldots, x_{1k}), \ldots, (y_n, x_{n1}, x_{n2}, \ldots, x_{nk})$$

conforming to Equation (1.1). Here (but not in general), for simplicity in presenting the equation for $\widehat{\beta}$, each covariate is assumed to have been centered around its observed mean. Thus, for the jth covariate ($j = 1, \ldots, k$) we have $\sum_{i=1}^{n} x_{ij} = 0$. The principle of ordinary least squares (OLS) estimation leads to the estimated regression coefficients

$$\widehat{\beta} = \left(\widehat{\beta}_1, \ldots, \widehat{\beta}_k\right)^{\mathrm{T}} = \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$$

where the (i, j)th element of the matrix $\mathbf{X}$ is x_{ij} and $\mathbf{y} = (y_1, \ldots, y_n)^{\mathrm{T}}$. The fitted or 'predicted' values are

$$\widehat{y}_i = \widehat{\beta}_0 + \mathbf{x}_i\widehat{\beta}$$

where

$$\widehat{\beta}_0 = \overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

Details of the theory of multiple linear regression may be found in a standard textbook such as Draper and Smith (1998) or Weisberg (2005).

Transformation of the Response Variable

A major topic in our book is the use of transformed predictors to accommodate nonlinear regression relationships. In contrast, we do not consider transformation of the response to improve fit, e.g. using the well-known method of Box and Cox (1964). In all the examples involving normal-errors models (see Tables A.1 and A.2), we assume that an appropriate transformation of the response has already been derived if required. Thus, the symbol y denotes the response variable in a given model, possibly after suitable transformation.

Residuals

OLS residuals are defined as

$$r_i = y_i - \widehat{y}_i$$

Raw residuals are typically too 'noisy' to be helpful; therefore, in our book we are generally concerned with smoothed residuals (see also Section 1.4). Mismodelling in Equation (1.1) may appear as a systematic pattern in the local mean residual as a function of some covariate x or the index η, and may be revealed by a smoothed scatter plot of the residuals on x (e.g. Cleveland and Devlin, 1988, Sasieni and Royston, 1998). The mean residual at a given value of x is interpretable as an estimate of the bias for Equation (1.1) in the mean value of y at x, a quantity of direct interest. Usually, 95% pointwise CIs are shown on the plot, to help one judge whether observed patterns are likely to be 'real' or due to chance.

1.3.2 Logistic Regression

Logistic regression, a special case of a GLM (see Section 1.3.4), is concerned with modelling binary responses. The multiple linear logistic regression model with covariates $x_1, \ldots, x_k$ asserts that the probability π of occurrence of a binary event y of interest, e.g. death or 'caseness' in a case-control study, may be represented by

$$\text{logit}\,\pi \;=\; \log\frac{\pi}{1-\pi} \;=\; \beta_0 + \sum_{j=1}^{k}\beta_j x_j \tag{1.2}$$

$\pi/(1-\pi)$ is known as the odds of an event. Suppose y takes the values 1 for an event and 0 for a nonevent. If (1.2) is correct, then y has a Bernoulli distribution with probability parameter (and expected value) π.

In a model with just a single binary covariate x, taking the values 0 and 1, then logit $\pi = \beta_0$ when $x = 0$ and $\beta_0 + \beta_1$ when $x = 1$. Let $\pi_{(1)} = \text{logit}(\pi|x=1)$ and $\pi_{(0)} = \text{logit}(\pi|x=0)$. It follows that

$$\begin{aligned}\text{logit}\,\pi_{(1)} - \text{logit}\,\pi_{(0)} \;&=\; \log\left(\frac{\pi_{(1)}}{1-\pi_{(1)}}\right) - \log\left(\frac{\pi_{(0)}}{1-\pi_{(0)}}\right) \;=\; (\beta_0+\beta_1) - \beta_0 \;=\; \beta_1 \\ &=\; \log\left[\left(\frac{\pi_{(1)}}{1-\pi_{(1)}}\right) \Big/ \left(\frac{\pi_{(0)}}{1-\pi_{(0)}}\right)\right]\end{aligned}$$

This shows the well-known result that the log odds ratio of an event when $x = 1$ compared with that when $x = 0$ equals the regression slope β_1 (or that the odds ratio equals $\exp\beta_1$).

Having now a sample of n observations $(y_1, \mathbf{x}_1), \ldots, (y_n, \mathbf{x}_n)$, estimates of β_0 and β are found by maximum likelihood. Let $\widehat{\eta}_i = \widehat{\beta}_0 + \mathbf{x}\widehat{\beta}_i$ be the index (linear predictor) from Equation (1.2). The probability that the ith observation is an event, i.e. $\Pr(y_i = 1|\mathbf{x}_i)$, is estimated by

$$\widehat{\pi}_i \;=\; \frac{\exp(\widehat{\eta}_i)}{1+\exp(\widehat{\eta}_i)}$$

Further details of the theory and practice of logistic regression may be found in Hosmer and Lemeshow (2000) or Collett (2003a).

Residuals

Several types of residual are available in logistic regression (e.g. Pearson and deviance). We work with the simplest and most accessible, the raw residuals $r_i = y_i - \widehat{\pi}_i$. As with normal-errors regression, mismodeling in (1.2) may be revealed by a scatter plot smooth of the residuals on x with 95% pointwise CIs, conceptually similar to Copas's (1983a) suggestion of 'plotting p against x'. In logistic regression, the smoothed residual at x is an estimate of the bias for (1.2) in the probability of an event given x.

1.3.3 Cox Regression

The 'Cox model' (Cox, 1972), also known as the proportional hazards model, is designed for modelling censored survival data. In its simplest form, the Cox model with covariates $x_1, \ldots, x_k$ describes the hazard of an event of interest at a time $t > 0$ after a starting point or time origin $t = 0$:

$$\lambda(t; \mathbf{x}) = \lambda_0(t) \exp\left(\sum_1^k \beta_j x_j\right)$$

For a single binary covariate x with regression coefficient β_1, the Cox model implies that

$$\frac{\lambda(t; 1)}{\lambda(t; 0)} = \frac{\lambda_0(t) \exp(\beta_1 \times 1)}{\lambda_0(t) \exp(\beta_1 \times 0)} = \exp \beta_1$$

The quantity $\lambda(t; 1)/\lambda(t; 0)$ is known as the *hazard ratio* (HR) for $x = 1$ compared with $x = 0$. More generally, the HR is the hazard at $\mathbf{x}$ divided by that at $\mathbf{x} = 0$. The HR plays a central role in survival analysis, since it is a convenient summary of the relationship between two entire survival curves. The crucial assumption of proportional hazards (PHs) is equivalent to saying that the HR is independent of t. If a non-PH is detected, then the Cox model may be extended in various ways to accommodate it (Therneau and Grambsch, 2000). A strategy for assessing potential non-PH and modelling it is described in Section 11.1.

A sample of n observations for survival analysis by a multivariable Cox model takes the form $(t_1, \mathbf{x}_1, \delta_1), \ldots, (t_n, \mathbf{x}_n, \delta_n)$, where δ_i is the 'censoring indicator'. δ_i takes the value 1 when t_i is an observed failure time and 0 when t_i is right-censored (i.e. when the precise time-to-event is unobserved, but is known to be $\geq t_i$). All times t_i must be positive; values of zero make no contribution to the estimation process. The parameter vector β is estimated by maximum partial likelihood. Many theoretical and practical details of survival modelling may be found in Hosmer and Lemeshow (1999) and Collett (2003b).

Residuals

We use martingale residuals and scaled Schoenfeld residuals. For details, see Hosmer and Lemeshow (1999, pp. 163, 198) or Therneau and Grambsch (2000, pp. 80, 85). Unscaled martingale residuals give a local estimate of the difference between the observed and predicted number of events. The pattern of the (smoothed) martingale residuals provides information on the functional form of a continuous covariate x in a model (Therneau et al., 1990). For a proposed function of x, systematic patterns seen in a plot of the smoothed martingale residuals against x indicate lack of fit, and may suggest how the chosen function of x may be improved. Note that, for comparability with the function estimated from a Cox model, the martingale residuals should be scaled by dividing by the ratio of the number of events to the number of individuals (Therneau et al., 1990). Such scaling does not affect the *pattern* of martingale residuals, only their *magnitude*, and is not applied in the relevant examples in our book.

Scaled Schoenfeld residuals are based on score residuals and are useful in a visual assessment of the PH assumption. Under PH, the mean of these residuals is zero, and is independent of time. A systematic pattern in the smoothed residuals when plotted against time suggests a time-varying effect of the covariate. The Grambsch–Therneau test (Grambsch and Therneau 1994) may be applied to test the PH assumption formally, for specific covariates or globally over all the covariates in the model.

1.3.4 Generalized Linear Models

A GLM (McCullagh and Nelder, 1989) comprises a random component, a systematic component and a link function which connects the two components. The response y is assumed to have a probability density function from the exponential family, namely

$$\exp\left[\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)\right]$$

where θ is known as the natural parameter, ϕ is a dispersion (scale) parameter, and $a(.)$, $b(.)$ and $c(.,.)$ are known functions. This density function defines the random component. The model also prescribes that the expectation μ of y is related to covariates $x_1, \ldots, x_k$ by $g(\mu) = \eta$ where $\eta = \beta_0 + \sum_1^k \beta_j x_j$. The index or linear predictor η is the systematic component and $g(.)$ is known as the link function.

The mean μ is related to θ by $\mu = \mathrm{d}b/\mathrm{d}\theta$. A convenient link for a given member of the exponential family is the canonical link, in which $g(\mu)$ is chosen so that $\eta = \theta$. For the Bernoulli distribution, which underlies logistic regression with a binary outcome, we have $E(y) = \Pr(y = 1)$. The usual choice in data analysis is the canonical link, which for logistic regression is the logit (see Equation (1.2)); sometimes the probit or the complementary log–log link is used. Standard practice is to define the model in terms of μ and η, so that θ plays no further part.

Given choices for the random and systematic components and the link function, and a sample $\mathbf{y}$ of n observations, estimation of the model parameters is done iteratively by maximum likelihood.

Residuals

Again we use the simplest and most interpretable residuals for GLMs, the raw residuals $r_i = y_i - \widehat{\mu}_i$. The r_i are conceptually identical to the raw residuals $y_i - \widehat{\pi}_i$ in logistic regression. They are in fact scaled Pearson residuals (see McCullagh and Nelder (1989, equation (2.11), p. 37)). Variations in $\mathrm{var}(r_i)$ as a function of a covariate or the index η_i are not of concern, since they are accommodated by the varying width of pointwise CIs given by a scatter-plot smoother applied to the raw residuals.

1.3.5 Linear and Additive Predictors

A different type of generalization of Equation (1.1) is from models with a linear predictor or index $\beta_0 + \sum_1^k \beta_j x_j$ to those with an *additive* predictor (also called an index) of the form $\beta_0 + \sum_1^k f_j(x_j)$, where $f_j(x_j)$ is a more complicated function of x_j than $\beta_j x_j$. These models apply only when x_j is a continuous covariate, such as age or blood pressure. Examples of types of $f_j(x_j)$ include polynomials, FPs, regression splines, smoothing splines, wavelets, Fourier series, and so on. Hastie and Tibshirani (1990) devote a whole book to an approach to regression modelling based on various types of additive function. Convenient classes of functions to use are cubic smoothing splines (Green and Silverman, 1994) and regression splines (de Boer 2001).

Note that $f_j(x_j)$ can be broken down into simpler parts, sometimes leading again to a linear predictor. For example, the quadratic regression model for a single predictor x may be written in additive format as $\beta_0 + f(x)$, where $f(x) = \beta_1 x + \beta_2 x^2$, or in linear format as $\beta_0 + \beta_1 x + \beta_2 x^2$. Be clear that the 'linearity' in the linear format relates to the *two* variables x and x^2. The quadratic model is additive in x, nonlinear in x, and linear in (x, x^2).

1.4 ROLE OF RESIDUALS

1.4.1 Uses of Residuals

Residuals have many roles in statistics. Most of them are connected with some form of model criticism (e.g. see Belsley et al. (1980)). In our book, we use residuals almost exclusively as a graphical tool to study the (lack of) fit of a function of a continuous predictor.

Preferred types of residual for the normal-errors, logistic and Cox models were discussed in Section 1.3. There are arguments favouring different residuals for different purposes. For example, residuals which are identically distributed under the assumption that the model is correct are particularly suitable for detecting outliers in the response. Residuals which are easily interpretable when smoothed are advantageous for detecting meaningful lack of fit of a function. For further discussion of residuals, please refer to Belsley et al. (1980).

Examples of smoothed residuals from models with a single predictor have already been presented (e.g. Figures 1.1 and 1.2). Such plots are useful for picking up anomalies in the fit. We use a univariate running-line smoother (Fan and Gijbels, 1996), implemented for Stata in the command `running` (Sasieni et al., 2005). We use the default amount of smoothing provided by `running`. Running-line smoothers of residuals provide a detailed picture of the relationship. As a result, they can give quite 'noisy' results, but the message from the data, of a lack of fit or otherwise, is usually sufficiently clear.

1.4.2 Graphical Analysis of Residuals

Generically, our favoured graphical analysis of residuals for a continuous predictor x, exemplified in Figures 1.1 and 1.2, has the following elements combined into a single plot:

1. A smooth of the residuals as a function of x, plotted as a solid line.
2. A pointwise 95% CI for the smooth, plotted as a shaded area.
3. A lightly shaded box within the plot region, bounded vertically by the 2.5th and 97.5th centiles of the observed distribution of x and horizontally by a convenient range of values. The box shows where most (95%) of the observations of x lie. The aim is to down-weight the visual impact of extreme values of x on the estimated function of x. The data are usually sparse near extreme values of x, and 'end effects' (unstable estimates of the function) are most likely to occur there.
4. A horizontal line representing $y = 0$, the expected value of the residuals if the model is correct.
5. (Optionally) a scatter plot of the raw residuals against x. This component may be omitted if the variation among the residuals visually overwhelms the smooth and its CI. To enhance legibility, martingale residuals from time-to-event models that are less than -1 may be truncated at -1.

Note that the pointwise 95% CIs should be interpreted cautiously, since they do not represent a global confidence region. For example, the value 0 may be excluded for some small range of x values, even when there is no serious lack of fit.

In a multivariable model, a plot of smoothed residuals may be drawn for every relevant predictor, perhaps including those not selected in the model. When feasible, a composite plot showing all such smooths in a single graphic is helpful. An example is Figure 6.5 (see Section 6.5.2).

1.5 ROLE OF SUBJECT-MATTER KNOWLEDGE IN MODEL DEVELOPMENT

The consensus is that subject-matter knowledge should generally guide model building. A study should be carefully planned, guided by the research questions and taking into account the methods envisaged for analysing the data. In randomized trials, the main research question is precisely specified – for example, is the new treatment better than the current standard with a hazard ratio of < 0.75? A detailed analysis plan is written before the analysis starts.

The situation is more complex and difficult with observational studies. At first glance, the research question may still be simple, e.g. whether there is an association between an exposure and the probability of developing a disease, adjusting for known confounders. However, this simple question may pose at least two serious challenges to the analyst. First, the relationship between a continuous exposure and the disease probability may have many possible functional forms. Does subject-matter knowledge support a particular functional form? That is unlikely when a 'new' exposure for a disease is under investigation. Assuming linearity is certainly a good starting point, but it must be checked and should be abandoned if the data 'say' otherwise. Second, what are the 'known' confounders in a given setting, and in what form should they be included in the model? If they are indeed 'known', why does it often happen that different sets of confounders are used in different studies? A small number of confounders may be prespecified, but why are many more variables typically collected? The potential problems of the statistical analysis increase if it is necessary to determine which of a large number of variables are associated with the outcome in a multivariable context (a typical question in studies of prognosis). Because subject-matter knowledge is typically limited or at best fragile, data-dependent model building is necessary (Harrell, 2001).

Unless stated otherwise, we assume in our book that subject-matter knowledge is so limited that it does not affect model building. However, when such knowledge does exist, analyses using fractional polynomials can easily be adapted to include it. For example, in Sauerbrei and Royston (1999) we noted that an FP function seemed biologically implausible and suggested a way to ensure that the estimated function was monotonic with an asymptote. Variables should sometimes be included in a model without being statistically significant, or should be excluded despite statistical significance. The former situation is more common with 'known' confounders, whereas the latter can happen if an 'expensive-to-measure' variable adds little to the model fit or to the explained variation. These two situations can easily be handled by 'forcing' variable(s) into or out of the model.

Sometimes, subject-matter knowledge may require a restricted class of nonlinear functions for certain variables (e.g. monotonic functions). When this is the case, the power to detect variables with a weak influence is increased by choosing the FP1 class as the most complex allowed functional form. See some simulation results in Section 4.16 and a fuller discussion in Section 6.9.2.

Most analyses of observational studies rely on a blend of subject-matter knowledge and data-dependent decisions. Initial decisions include grouping of categorical variables or, for continuous variables, considering how to handle extreme values or outliers (see Section 2.3). The concerns are specific to a dataset, but where possible the decisions should be based on subject-matter knowledge. The aims in general are to obtain an 'optimal' fit to the data, interpretable covariate effects, consistency with subject-matter knowledge where available, general usability by others, and transportability to other settings (external validation). In summary, multivariable model-building has elements of art in the attempt to provide satisfactory answers to more or less vague questions through the analysis of the data at hand under much uncertainty.

1.6 SCOPE OF MODEL BUILDING IN OUR BOOK

The techniques discussed in our book are intended and appear to work best for model building under certain conditions, as summarized in Table 1.3.

Table 1.3 Issues in building regression models, when the aim is to identify influential variables and to determine the functional form for continuous variables.

Issue	Assumption in our book (unless stated otherwise)	Reason for the assumption
Subject matter knowledge	No knowledge	Subject-matter knowledge should always be incorporated in the model-building process or should even guide an analysis. However, often it is limited or nonexistent, and data-dependent model building is required
Number of variables	About 5 to 30	With a smaller number of variables, selection may not be required. With many more variables (e.g. high-dimensional data), the approaches may no longer be feasible or will require (substantial) modification
Correlation structure	Correlations are not 'very' strong (e.g. correlation coefficient below 0.7)	Stronger correlations often appear in fields such as econometrics, less commonly in medicine. For large correlations, nonstatistical criteria may be used to select a variable. Alternatively, a 'representative', e.g. a linear combination of the correlated variables, may be chosen
Sample size	At least 10 observations per variable	With a (much) smaller sample size, selection bias and model instability become major issues. An otherwise satisfactory approach to variable and/or function selection may fail, or may require extension (e.g. shrinkage to correct for selection bias)
Completeness of data	No missing data	Particularly with multivariable data, missing covariate data introduces many additional problems. Not considered here
Variable selection procedure	Only sequential and all-subsets selection strategies are considered	Stepwise and all-subsets procedures are the main types used in practice. BE and an appropriate choice of significance level gives results similar to all-subsets selection
Functional form of continuous covariates	Full information from the covariate is used	Categorizing continuous variables should be avoided. A linear function is often justifiable, but sometimes may not fit the data. Check with FPs or splines whether non-linear functions markedly improve the fit
Interaction between covariates	No interactions	Investigation of interactions complicates multivariable model-building. Investigation of interactions should take subject-matter knowledge into account

(Adapted from Sauerbrei et al. (2007a, Table 1) and Sauerbrei et al. (1999) with permission from John Wiley & Sons Ltd.)

The restriction to no interactions is applied rigorously in Chapters 1–6, but is lifted in Chapter 7 and subsequently. Relaxation of the assumptions is possible in some cases, e.g. the number of variables may be about 40, or 9 observations per variable are acceptable. The limitations make it easier to answer the main research questions, but do not seriously reduce the scope of what may be done and the recommendations for practice (see Section 12.2). Also, in some of our examples we violate some of the assumptions.

1.7 MODELLING PREFERENCES

1.7.1 General Issues

The application of complex statistical methods for the development of regression models has greatly increased in recent years. Advances in statistical methodology now allow us to create and estimate more realistic models than ever before, and the necessary computer programs are often available. By contrast, however, the properties of many model-building procedures are still unknown, and the few comparisons that do exist tend to be based on (small) simulation studies. This unfortunate situation is a key reason why tastes in model building vary so much between statisticians.

The aims of an investigation play an important role in every statistical analysis. At least in the health sciences, there seems to be a consensus that subject-matter knowledge must be incorporated in such analyses. With some minor modifications, this can usually be done with all the procedures discussed in our book. However, practical experience shows that, in most analyses of observational studies, data-driven model building still plays an important role. Some variables are inevitably chosen mainly by statistical principles – essentially, P-values for including or excluding variables, or information criteria. The definition of a 'best' strategy to produce a model that has good predictive properties in new data is difficult.

1.7.2 Criteria for a Good Model

It is important to distinguish between two main aims when creating a model. The first is prediction, with little consideration of the model structure; the second is explanation, where we try to identify influential predictors and gain insight into the relationship between the predictors and the outcome. Much published research focuses on prediction, in which model fit and mean-square prediction error are the main criteria for model adequacy. With our background in clinical epidemiology, the second aim is more appropriate. Studies are done to investigate whether particular variables are prognostically or diagnostically important, or are associated with an increased risk of some outcome. For continuous predictors, the shape of the function is often of interest, e.g. whether there is an increasing trend or a plateau at high values of x. Because disease causation is invariably multifactorial, such assessments must be done in a multivariable context. In reality, many variables may be considered as potential predictors, but only a few have a relevant effect. The task is to identify them. Often, generalizability and practical usefulness must also be kept in mind when developing a model. Consider, for example, a prognostic model comprising many variables. All constituent variables would have to be measured in an identical or at least in a similar way, even when their effects are very small. Such a model is impractical, therefore 'not clinically useful' and likely to be 'quickly forgotten' (Wyatt and Altman, 1995). In reality, a model satisfying the second aim, although

not providing an optimal predictor in the sense of minimizing mean-square error or similar criteria, typically has only slightly inferior performance. A model that fits the current dataset well may be too data driven to reflect the underlying relationships adequately.

The distinction between prediction and interest in the effects of individual variables was stressed by Copas (1983b). He noted that the loss functions are different, and stated that a good predictor 'may include variables which are not significant, exclude others which are, and may involve coefficients which are systematically biased'. Such a predictor would clearly fail to satisfy the explanatory aim of many studies. Apart from these general considerations, no clear guidance on how to develop a multivariable model fulfilling such an aim appears to be available.

1.7.3 Personal Preferences

Our general philosophy is based on experience in real applications and simulation studies, and on investigations of model stability by bootstrap resampling (Efron, 1979; Sauerbrei and Schumacher, 1992; Sauerbrei, 1999; Royston and Sauerbrei, 2003; Ambler and Royston, 2001). It is influenced by the potential use of our models in future applications, an important aim of most of the models we have developed so far. These considerations have led us to prefer simple models unless the data indicate the need for greater complexity. In the context of time-series forecasting, Chatfield (2002) states that the cost of achieving an excellent fit to the current data may be a poor fit to future data, and 'this emphasizes the importance of checking any model with new data . . . and explains my preference for simple models'. Hand (2006) expresses similar views regarding classification methods.

A distinctive feature of FP modelling is the availability of a rigorous procedure for selecting variables and functions. The principles for selecting an FP function are easily explained. Combination with BE results in a procedure applicable without detailed expert knowledge.

Arguments favouring BE over other stepwise methods have been given by Mantel (1970). Sauerbrei (1999) argued that BE(0.157) (i.e. BE using a nominal significance level of 0.157) may be used as a substitute for all-subsets procedures with C_p or Akaike's information criterion (AIC). His conclusion was based on asymptotic and simulation results on the significance level for all-subsets procedures, on simulation results for the stepwise methods, and on empirical comparisons in particular datasets (Teräsvirta and Mellin, 1986; Sauerbrei, 1992; Sauerbrei, 1993). Models selected with BE(0.157) and AIC usually have at most minor differences (Blettner and Sauerbrei, 1993; Sauerbrei, 1993). For further discussion, see Section 2.6. We consider BE to be a good candidate for a sensible variable selection strategy. We also believe that the class of FP functions is a good candidate for finding nonlinear relationships with continuous covariates and at the same time for generating interpretable and transferable (generally useful) models (Royston and Sauerbrei, 2005; Sauerbrei et al., 2007a).

The MFP procedure combines these two components (selection of variables and functions). It is computationally not too demanding, statistically comprehensible, and may be applied to most types of regression model. Furthermore, it addresses the two main tasks in multivariable model-building: elimination of 'unimportant' variables and selection of a 'reasonable' dose–response function for continuous variables. We are well aware that every model can only be a crude approximation to the complex relationships existing in reality. We do not aim to fit the data in any sense 'optimally'. A model that includes at least the strong predictors and whose unknown functional form seems to be 'roughly' modelled in a plausible way is, from

our point of view, acceptable. Since the MFP modelling process is comprehensible without detailed expert knowledge, the resulting models are interpretable and transferable. We consider MFP modelling to be an important pragmatic approach to determine multivariable models for continuous variables. In a similar vein, Hand (2006) observes that 'more complicated models often require tuning . . . and, in general, experts are more able to obtain good results than are inexperienced users. On the other hand, simple models can often be applied successfully by inexperienced users'.

In the MFP approach, the nominal significance levels are the tuning parameters, which largely determine the nature of the resulting model with respect to both the number of variables chosen and the complexity of any selected functions. Depending on the aim of a study, significance levels, which may be different for selection of variables and of complexity of functions, may be chosen. For example, when determining adjustment factors in an epidemiological study, a nominal P-value of 0.2 may be sensible, whereas in a study developing a multivariable diagnostic index a P-value of 0.01 may be more appropriate.

MFP has been progressively extended to perform wider tasks, e.g. modelling interactions between a categorical and continuous covariate (Royston and Sauerbrei 2004a), and determining time-varying functions of regression coefficients in the Cox model (Sauerbrei et al., 2007c). We are well aware that data-dependent modelling ignores the uncertainties of the model-building process and leads to potentially biased estimates of parameters and underestimation of their standard errors.

1.8 GENERAL NOTATION

Here, we provide a concise explanation of general notation used in our book. We have kept mathematical exposition and notation to an absolute minimum throughout.

In general, x denotes a predictor (covariate, independent variable, explanatory variable, risk factor, etc.) and y an outcome variable (response, dependent variable, etc.). We use lowercase letters (e.g. x, β) to denote scalar quantities. Uppercase letters are used sparingly, sometimes denoting models (e.g. M_1). Lowercase bold letters (e.g. $\mathbf{x}$, $\boldsymbol{\beta}$, $\mathbf{p}$) are used for (row) vectors. The expression $\eta = \beta_0 + \mathbf{x}\boldsymbol{\beta}$ denotes the 'index' of a model (see Section 1.3.5), where $\mathbf{x}$ and $\boldsymbol{\beta}$, each vectors with k elements, are explanatory variables and regression parameters of the model respectively, and β_0 is the intercept. Strictly speaking, $\mathbf{x}\boldsymbol{\beta}$ should be written as $\mathbf{x}\boldsymbol{\beta}^{\mathrm{T}}$, where the superscript T denotes vector or matrix transpose, but no ambiguity results from omitting the transpose.

Expectation is denoted by $E()$, variance by var(), standard deviation by SD or SD() and standard error by SE or SE(), and ninety-five percent confidence interval by 95% CI. The distribution of y is sometimes indexed by $\mu = E(y)$, or by an equivalent parameter $g(\mu)$, where g is a monotonic function known as the link function (see Section 1.3.4).

The quantity D denotes the 'deviance' or minus twice the (maximized) log likelihood of a model. A Gaussian or normal distribution with mean μ and variance σ^2 is denoted by $N(\mu, \sigma^2)$. The chi-squared distribution with d degrees of freedom (d.f.) is denoted by χ^2_d, and $F_{m,n}$ refers to the F-distribution with m and n degrees of freedom. A chi-square test statistic obtained from a likelihood ratio test is denoted by χ^2.

When describing examples and case studies, variable names such as `age` and `pgr` are written in typewriter font to distinguish them from the rest of the text. Sometimes, for brevity, variable names are given in algebraic form (e.g. x_1, $x_5 - x_8$).

Notation relating to model selection algorithms with nominal significance levels α or (α_1, α_2), e.g. BE(α), FSP(α), MFP(α_1, α_2), is introduced when required in Chapters 2, 4 and 6. Notation specific to FPs is introduced in Section 4.3.

When no confusion can arise, the same notation may be used for different quantities. For example, x^* denotes scaled x in Section 4.11 and a negative exponential transformation in Section 5.6.2.