Chapter 2
General Aspects of Fitting Regression Models

2.1 Notation for Multivariable Regression Models

The ordinary multiple linear regression model is frequently used and has parameters that are easily interpreted. In this chapter we study a general class of regression models, those stated in terms of a weighted sum of a set of independent or predictor variables. It is shown that after linearizing the model with respect to the predictor variables, the parameters in such regression models are also readily interpreted. Also, all the designs used in ordinary linear regression can be used in this general setting. These designs include analysis of variance (ANOVA) setups, interaction effects, and nonlinear effects. Besides describing and interpreting general regression models, this chapter also describes, in general terms, how the three types of assumptions of regression models can be examined.

First we introduce notation for regression models. Let Y denote the response (dependent) variable, and let $X = X_1, X_2, \ldots, X_p$ denote a list or vector of predictor variables (also called covariables or independent, descriptor, or concomitant variables). These predictor variables are assumed to be constants for a given individual or subject from the population of interest. Let $\beta = \beta_0, \beta_1, \ldots, \beta_p$ denote the list of regression coefficients (parameters). β_0 is an optional intercept parameter, and $\beta_1, \ldots, \beta_p$ are weights or regression coefficients corresponding to $X_1, \ldots, X_p$. We use matrix or vector notation to describe a weighted sum of the Xs:

$$X\beta = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p, \tag{2.1}$$

where there is an implied $X_0 = 1$.

A regression model is stated in terms of a connection between the predictors X and the response Y. Let $C(Y|X)$ denote a property of the distribution of Y given X (as a function of X). For example, $C(Y|X)$ could be $E(Y|X)$,

F.E. Harrell, Jr., *Regression Modeling Strategies*, Springer Series in Statistics, DOI 10.1007/978-3-319-19425-7_2

the expected value or average of Y given X, or $C(Y|X)$ could be the probability that $Y = 1$ given X (where $Y = 0$ or 1).

2.2 Model Formulations

We define a regression function as a function that describes interesting properties of Y that may vary across individuals in the population. X describes the list of factors determining these properties. Stated mathematically, a general regression model is given by

$$C(Y|X) = g(X). \tag{2.2}$$

We restrict our attention to models that, after a certain transformation, are linear in the unknown parameters, that is, models that involve X only through a weighted sum of all the Xs. The *general linear regression model* is given by

$$C(Y|X) = g(X\beta). \tag{2.3}$$

For example, the ordinary linear regression model is

$$C(Y|X) = E(Y|X) = X\beta, \tag{2.4}$$

and given X, Y has a normal distribution with mean $X\beta$ and constant variance σ^2. The binary logistic regression model[129,647] is

$$C(Y|X) = \text{Prob}\{Y = 1|X\} = (1 + \exp(-X\beta))^{-1}, \tag{2.5}$$

where Y can take on the values 0 and 1. In general the model, when stated in terms of the property $C(Y|X)$, may not be linear in $X\beta$; that is $C(Y|X) = g(X\beta)$, where $g(u)$ is nonlinear in u. For example, a regression model could be $E(Y|X) = (X\beta)^{.5}$. The model may be made linear in the unknown parameters by a transformation in the property $C(Y|X)$:

$$h(C(Y|X)) = X\beta, \tag{2.6}$$

where $h(u) = g^{-1}(u)$, the inverse function of g. As an example consider the binary logistic regression model given by

$$C(Y|X) = \text{Prob}\{Y = 1|X\} = (1 + \exp(-X\beta))^{-1}. \tag{2.7}$$

If $h(u) = \text{logit}(u) = \log(u/(1-u))$, the transformed model becomes

$$h(\text{Prob}(Y = 1|X)) = \log(\exp(X\beta)) = X\beta. \tag{2.8}$$

The transformation $h(C(Y|X))$ is sometimes called a *link function*. Let $h(C(Y|X))$ be denoted by $C'(Y|X)$. The general linear regression model then becomes

$$C'(Y|X) = X\beta. \tag{2.9}$$

In other words, the model states that some property C' of Y, given X, is a weighted sum of the Xs ($X\beta$). In the ordinary linear regression model, $C'(Y|X) = E(Y|X)$. In the logistic regression case, $C'(Y|X)$ is the logit of the probability that $Y = 1$, $\log \text{Prob}\{Y = 1\}/[1 - \text{Prob}\{Y = 1\}]$. This is the log of the odds that $Y = 1$ versus $Y = 0$.

It is important to note that the general linear regression model has two major components: $C'(Y|X)$ and $X\beta$. The first part has to do with a property or transformation of Y. The second, $X\beta$, is the *linear regression* or *linear predictor* part. The method of least squares can sometimes be used to fit the model if $C'(Y|X) = E(Y|X)$. Other cases must be handled using other methods such as maximum likelihood estimation or nonlinear least squares.

2.3 Interpreting Model Parameters

In the original model, $C(Y|X)$ specifies the way in which X affects a property of Y. Except in the ordinary linear regression model, it is difficult to interpret the individual parameters if the model is stated in terms of $C(Y|X)$. In the model $C'(Y|X) = X\beta = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p$, the regression parameter β_j is interpreted as the change in the property C' of Y per unit change in the descriptor variable X_j, all other descriptors remaining constant[a]:

$$\beta_j = C'(Y|X_1, X_2, \ldots, X_j + 1, \ldots, X_p) - C'(Y|X_1, X_2, \ldots, X_j, \ldots, X_p). \tag{2.10}$$

In the ordinary linear regression model, for example, β_j is the change in expected value of Y per unit change in X_j. In the logistic regression model β_j is the change in log odds that $Y = 1$ per unit change in X_j. When a non-interacting X_j is a dichotomous variable or a continuous one that is linearly related to C', X_j is represented by a single term in the model and its contribution is described fully by β_j.

In all that follows, we drop the $'$ from C' and assume that $C(Y|X)$ is the property of Y that is linearly related to the weighted sum of the Xs.

[a] Note that it is not necessary to "hold constant" all other variables to be able to interpret the effect of one predictor. It is sufficient to hold constant the weighted sum of all the variables other than X_j. And in many cases it is not physically possible to hold other variables constant while varying one, e.g., when a model contains X and X^2 (David Hoaglin, personal communication).

2.3.1 Nominal Predictors

Suppose that we wish to model the effect of two or more treatments and be able to test for differences between the treatments in some property of Y. A nominal or polytomous factor such as treatment group having k levels, in which there is no definite ordering of categories, is fully described by a series of $k-1$ binary indicator variables (sometimes called *dummy variables*). Suppose that there are four treatments, J, K, L, and M, and the treatment factor is denoted by T. The model can be written as

$$\begin{aligned} C(Y|T = J) &= \beta_0 \\ C(Y|T = K) &= \beta_0 + \beta_1 \\ C(Y|T = L) &= \beta_0 + \beta_2 \\ C(Y|T = M) &= \beta_0 + \beta_3. \end{aligned} \tag{2.11}$$

The four treatments are thus completely specified by three regression parameters and one intercept that we are using to denote treatment J, the reference treatment. This model can be written in the previous notation as

$$C(Y|T) = X\beta = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3, \tag{2.12}$$

where

$$\begin{aligned} X_1 &= 1 \text{ if } T = K, 0 \text{ otherwise} \\ X_2 &= 1 \text{ if } T = L, 0 \text{ otherwise} \\ X_3 &= 1 \text{ if } T = M, 0 \text{ otherwise.} \end{aligned} \tag{2.13}$$

For treatment J ($T = J$), all three Xs are zero and $C(Y|T = J) = \beta_0$. The test for any differences in the property $C(Y)$ between treatments is $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$.

This model is an *analysis of variance* or *k-sample*-type model. If there are other descriptor covariables in the model, it becomes an *analysis of covariance*-type model.

2.3.2 Interactions

Suppose that a model has descriptor variables X_1 and X_2 and that the effect of the two Xs cannot be separated; that is the effect of X_1 on Y depends on the level of X_2 and vice versa. One simple way to describe this *interaction* is to add the constructed variable $X_3 = X_1 X_2$ to the model:

$$C(Y|X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2. \tag{2.14}$$

It is now difficult to interpret β_1 and β_2 in isolation. However, we may quantify the effect of a one-unit increase in X_1 if X_2 is held constant as

Table 2.1 Parameters in a simple model with interaction

Parameter	Meaning
β_0	$C(Y\|age = 0, sex = m)$
β_1	$C(Y\|age = x + 1, sex = m) - C(Y\|age = x, sex = m)$
β_2	$C(Y\|age = 0, sex = f) - C(Y\|age = 0, sex = m)$
β_3	$C(Y\|age = x + 1, sex = f) - C(Y\|age = x, sex = f) -$ $[C(Y\|age = x + 1, sex = m) - C(Y\|age = x, sex = m)]$

$$\begin{aligned} C(Y|X_1 + 1, X_2) - C(Y|X_1, X_2) & \\ &= \beta_0 + \beta_1(X_1 + 1) + \beta_2 X_2 \\ &+ \beta_3(X_1 + 1)X_2 \\ &- [\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2] \\ &= \beta_1 + \beta_3 X_2. \end{aligned} \tag{2.15}$$

Likewise, the effect of a one-unit increase in X_2 on C if X_1 is held constant is $\beta_2 + \beta_3 X_1$. Interactions can be much more complex than can be modeled with a product of two terms. If X_1 is binary, the interaction may take the form of a difference in shape (and/or distribution) of X_2 versus $C(Y)$ depending on whether $X_1 = 0$ or $X_1 = 1$ (e.g., logarithm vs. square root). When both variables are continuous, the possibilities are much greater (this case is discussed later). Interactions among more than two variables can be exceedingly complex.

2.3.3 Example: Inference for a Simple Model

Suppose we postulated the model

$$C(Y|age, sex) = \beta_0 + \beta_1 age + \beta_2[sex = f] + \beta_3 age[sex = f],$$

where $[sex = f]$ is a 0–1 indicator variable for sex = female; the reference cell is sex = male corresponding to a zero value of the indicator variable. This is a model that assumes

1. age is linearly related to $C(Y)$ for males,
2. age is linearly related to $C(Y)$ for females, and
3. whatever distribution, variance, and independence assumptions are appropriate for the model being considered.

We are thus assuming that the interaction between age and sex is simple; that is it only alters the slope of the age effect. The parameters in the model have interpretations shown in Table 2.1. β_3 is the difference in slopes (female – male).

There are many useful hypotheses that can be tested for this model. First let's consider two hypotheses that are seldom appropriate although they are routinely tested.

1. $H_0 : \beta_1 = 0$: This tests whether age is associated with Y for males.
2. $H_0 : \beta_2 = 0$: This tests whether sex is associated with Y for zero-year olds.

Now consider more useful hypotheses. For each hypothesis we should write what is being tested, translate this to tests in terms of parameters, write the alternative hypothesis, and describe what the test has maximum power to detect. The latter component of a hypothesis test needs to be emphasized, as almost every statistical test is focused on one specific pattern to detect. For example, a test of association against an alternative hypothesis that a slope is nonzero will have maximum power when the true association is linear. If the true regression model is exponential in X, a linear regression test will have some power to detect "non-flatness" but it will not be as powerful as the test from a well-specified exponential regression effect. If the true effect is U-shaped, a test of association based on a linear model will have almost no power to detect association. If one tests for association against a quadratic (parabolic) alternative, the test will have some power to detect a logarithmic shape but it will have very little power to detect a cyclical trend having multiple "humps." In a quadratic regression model, a test of linearity against a quadratic alternative hypothesis will have reasonable power to detect a quadratic nonlinear effect but very limited power to detect a multiphase cyclical trend. Therefore in the tests in Table 2.2 keep in mind that power is maximal when linearity of the age relationship holds for both sexes. In fact it may be useful to write alternative hypotheses as, for example, "H_a : age is associated with $C(Y)$, powered to detect a *linear* relationship."

Note that if there is an interaction effect, we know that there is both an age and a sex effect. However, there can also be age or sex effects when the lines are parallel. That's why the tests of total association have 2 d.f.

2.4 Relaxing Linearity Assumption for Continuous Predictors

2.4.1 Avoiding Categorization

Relationships among variables are seldom linear, except in special cases such as when one variable is compared with itself measured at a different time. It is a common belief among practitioners who do not study bias and

efficiency in depth that the presence of non-linearity should be dealt with by chopping continuous variables into intervals. Nothing could be more disastrous.[13,14,17,45,82,185,187,215,294,300,379,446,465,521,533,559,597,646]

Table 2.2 Most Useful Tests for Linear $Age \times Sex$ Model

Null or Alternative Hypothesis	Mathematical Statement
Effect of age is independent of sex or Effect of sex is independent of age or Age and sex are additive Age effects are parallel	$H_0 : \beta_3 = 0$
Age interacts with sex Age modifies effect of sex Sex modifies effect of age Sex and age are non-additive (synergistic)	$H_a : \beta_3 \neq 0$
Age is not associated with Y Age is associated with Y Age is associated with Y for either Females or males	$H_0 : \beta_1 = \beta_3 = 0$ $H_a : \beta_1 \neq 0$ or $\beta_3 \neq 0$
Sex is not associated with Y Sex is associated with Y Sex is associated with Y for some Value of age	$H_0 : \beta_2 = \beta_3 = 0$ $H_a : \beta_2 \neq 0$ or $\beta_3 \neq 0$
Neither age nor sex is associated with Y Either age or sex is associated with Y	$H_0 : \beta_1 = \beta_2 = \beta_3 = 0$ $H_a : \beta_1 \neq 0$ or $\beta_2 \neq 0$ or $\beta_3 \neq 0$

Problems caused by dichotomization include the following.

1. Estimated values will have reduced precision, and associated tests will have reduced power.
2. Categorization assumes that the relationship between the predictor and the response is flat within intervals; this assumption is far less reasonable than a linearity assumption in most cases.
3. To make a continuous predictor be more accurately modeled when categorization is used, multiple intervals are required. The needed indicator variables will spend more degrees of freedom than will fitting a smooth relationship, hence power and precision will suffer. And because of sample size limitations in the very low and very high range of the variable, the outer intervals (e.g., outer quintiles) will be wide, resulting in significant heterogeneity of subjects within those intervals, and residual confounding.
4. Categorization assumes that there is a discontinuity in response as interval boundaries are crossed. Other than the effect of time (e.g., an instant stock price drop after bad news), there are very few examples in which such discontinuities have been shown to exist.
5. Categorization only seems to yield interpretable estimates such as odds ratios. For example, suppose one computes the odds ratio for stroke for persons with a systolic blood pressure > 160 mmHg compared with persons with a blood

pressure $\leq$ 160 mmHg. The interpretation of the resulting odds ratio will depend on the exact distribution of blood pressures in the sample (the proportion of subjects $>$ 170, $>$ 180, etc.). On the other hand, if blood pressure is modeled as a continuous variable (e.g., using a regression spline, quadratic, or linear effect) one can estimate the ratio of odds for *exact* settings of the predictor, e.g., the odds ratio for 200 mmHg compared with 120 mmHg.

6. Categorization does not condition on full information. When, for example, the risk of stroke is being assessed for a new subject with a known blood pressure (say 162 mmHg), the subject does not report to her physician "my blood pressure exceeds 160" but rather reports 162 mmHg. The risk for this subject will be much lower than that of a subject with a blood pressure of 200 mmHg.
7. If cutpoints are determined in a way that is not blinded to the response variable, calculation of P-values and confidence intervals requires special simulation techniques; ordinary inferential methods are completely invalid. For example, if cutpoints are chosen by trial and error in a way that utilizes the response, even informally, ordinary P-values will be too small and confidence intervals will not have the claimed coverage probabilities. The correct Monte-Carlo simulations must take into account both multiplicities and uncertainty in the choice of cutpoints. For example, if a cutpoint is chosen that minimizes the P-value and the resulting P-value is 0.05, the true type I error can easily be above 0.5[300].
8. Likewise, categorization that is not blinded to the response variable results in biased effect estimates[17,559].
9. "Optimal" cutpoints do not replicate over studies. Hollander et al.[300] state that "...the optimal cutpoint approach has disadvantages. One of these is that in almost every study where this method is applied, another cutpoint will emerge. This makes comparisons across studies extremely difficult or even impossible. Altman et al. point out this problem for studies of the prognostic relevance of the S-phase fraction in breast cancer published in the literature. They identified 19 different cutpoints used in the literature; some of them were solely used because they emerged as the 'optimal' cutpoint in a specific data set. In a meta-analysis on the relationship between cathepsin-D content and disease-free survival in node-negative breast cancer patients, 12 studies were in included with 12 different cutpoints ... Interestingly, neither cathepsin-D nor the S-phase fraction are recommended to be used as prognostic markers in breast cancer in the recent update of the American Society of Clinical Oncology." Giannoni et al.[215] demonstrated that many claimed "optimal cutpoints" are just the observed median values in the sample, which happens to optimize statistical power for detecting a separation in outcomes and have nothing to do with true outcome thresholds. Disagreements in cutpoints (which are bound to happen whenever one searches for things that do not exist) cause severe interpretation problems. One study may provide an odds ratio for comparing body mass index (BMI) $>$ 30 with BMI $\leq$ 30, another for comparing BMI $>$ 28 with BMI $\leq$ 28. Neither of these odds ratios has a good definition and the two estimates are not comparable.
10. Cutpoints are arbitrary and manipulatable; cutpoints can be found that can result in both positive and negative associations[646].
11. If a confounder is adjusted for by categorization, there will be residual confounding that can be explained away by inclusion of the continuous form of the predictor in the model in addition to the categories.

When cutpoints are chosen using Y, categorization represents one of those few times in statistics where both type I and type II errors are elevated.

A scientific quantity is a quantity which can be defined outside of the specifics of the current experiment. The kind of high:low estimates that result from categorizing a continuous variable are not scientific quantities; their interpretation depends on the entire sample distribution of continuous measurements within the chosen intervals.

To summarize problems with categorization it is useful to examine its effective assumptions. Suppose one assumes there is a single cutpoint c for predictor X. Assumptions implicit in seeking or using this cutpoint include (1) the relationship between X and the response Y is discontinuous at $X = c$ and only $X = c$; (2) c is correctly found as the cutpoint; (3) X vs. Y is flat to the left of c; (4) X vs. Y is flat to the right of c; (5) the "optimal" cutpoint does not depend on the values of other predictors. Failure to have these assumptions satisfied will result in great error in estimating c (because it doesn't exist), low predictive accuracy, serious lack of model fit, residual confounding, and overestimation of effects of remaining variables.

A better approach that maximizes power and that only assumes a smooth relationship is to use regression splines for predictors that are not known to predict linearly. Use of flexible parametric approaches such as this allows standard inference techniques (P-values, confidence limits) to be used, as will be described below. Before introducing splines, we consider the simplest approach to allowing for nonlinearity.

2.4.2 Simple Nonlinear Terms

If a continuous predictor is represented, say, as X_1 in the model, the model is assumed to be linear in X_1. Often, however, the property of Y of interest does not behave linearly in all the predictors. The simplest way to describe a nonlinear effect of X_1 is to include a term for $X_2 = X_1^2$ in the model:

$$C(Y|X_1) = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2. \tag{2.16}$$

If the model is truly linear in X_1, β_2 will be zero. This model formulation allows one to test H_0 : model is linear in X_1 against H_a : model is quadratic (parabolic) in X_1 by testing $H_0 : \beta_2 = 0$.

Nonlinear effects will frequently not be of a parabolic nature. If a transformation of the predictor is known to induce linearity, that transformation (e.g., $\log(X)$) may be substituted for the predictor. However, often the transformation is not known. Higher powers of X_1 may be included in the model to approximate many types of relationships, but polynomials have some undesirable properties (e.g., undesirable peaks and valleys, and the fit in one region of X can be greatly affected by data in other regions[433]) and will not adequately fit many functional forms.[156] For example, polynomials do not adequately fit logarithmic functions or "threshold" effects.

2.4.3 Splines for Estimating Shape of Regression Function and Determining Predictor Transformations

A draftsman's *spline* is a flexible strip of metal or rubber used to draw curves. Spline functions are piecewise polynomials used in curve fitting. That is, they are polynomials within intervals of X that are connected across different intervals of X. Splines have been used, principally in the physical sciences, to approximate a wide variety of functions. The simplest spline function is a linear spline function, a piecewise linear function. Suppose that the x axis is divided into intervals with endpoints at a, b, and c, called *knots*. The linear spline function is given by

$$f(X) = \beta_0 + \beta_1 X + \beta_2 (X-a)_+ + \beta_3 (X-b)_+ + \beta_4 (X-c)_+, \tag{2.17}$$

where

$$\begin{aligned} (u)_+ &= u, \; u > 0, \\ &\quad 0, \; u \leq 0. \end{aligned} \tag{2.18}$$

The number of knots can vary depending on the amount of available data for fitting the function. The linear spline function can be rewritten as

$$\begin{aligned} f(X) &= \beta_0 + \beta_1 X, && X \leq a \\ &= \beta_0 + \beta_1 X + \beta_2 (X-a) && a < X \leq b \\ &= \beta_0 + \beta_1 X + \beta_2 (X-a) + \beta_3 (X-b) && b < X \leq c \\ &= \beta_0 + \beta_1 X + \beta_2 (X-a) \\ &\quad + \beta_3 (X-b) + \beta_4 (X-c) && c < X. \end{aligned} \tag{2.19}$$

A linear spline is depicted in Figure 2.1.

The general linear regression model can be written assuming only piecewise linearity in X by incorporating constructed variables X_2, X_3, and X_4 :

$$C(Y|X) = f(X) = X\beta, \tag{2.20}$$

where $X\beta = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4$, and

$$\begin{aligned} X_1 = X \quad X_2 = (X-a)_+ \\ X_3 = (X-b)_+ \quad X_4 = (X-c)_+. \end{aligned} \tag{2.21}$$

By modeling a slope increment for X in an interval $(a, b]$ in terms of $(X-a)_+$, the function is constrained to join ("meet") at the knots. Overall linearity in X can be tested by testing $H_0 : \beta_2 = \beta_3 = \beta_4 = 0$.

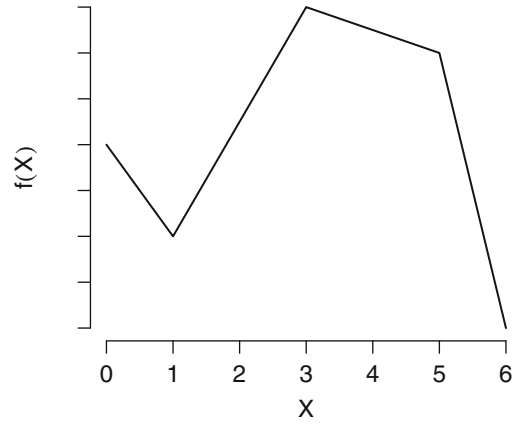

Fig. 2.1 A linear spline function with knots at $a = 1, b = 3, c = 5$.

2.4.4 Cubic Spline Functions

Although the linear spline is simple and can approximate many common relationships, it is not smooth and will not fit highly curved functions well. These problems can be overcome by using piecewise polynomials of order higher than linear. Cubic polynomials have been found to have nice properties with good ability to fit sharply curving shapes. Cubic splines can be made to be smooth at the join points (knots) by forcing the first and second derivatives of the function to agree at the knots. Such a smooth cubic spline function with three knots (a, b, c) is given by

$$\begin{aligned} f(X) &= \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 \\ &\quad + \beta_4 (X-a)_+^3 + \beta_5 (X-b)_+^3 + \beta_6 (X-c)_+^3 \\ &= X\beta \end{aligned} \tag{2.22}$$

with the following constructed variables:

$$\begin{aligned} X_1 = X \quad X_2 = X^2 \\ X_3 = X^3 \quad X_4 = (X-a)_+^3 \\ X_5 = (X-b)_+^3 \quad X_6 = (X-c)_+^3. \end{aligned} \tag{2.23}$$

If the cubic spline function has k knots, the function will require estimating $k + 3$ regression coefficients besides the intercept. See Section 2.4.6 for information on choosing the number and location of knots. 1

There are more numerically stable ways to form a design matrix for cubic spline functions that are based on B-splines instead of the truncated power basis[152,575] used here. However, B-splines are more complex and do not allow for extrapolation beyond the outer knots, and the truncated power basis seldom presents estimation problems (see Section 4.6) when modern methods such as the Q–R decomposition are used for matrix inversion. 2

2.4.5 Restricted Cubic Splines

Stone and Koo[595] have found that cubic spline functions do have a drawback in that they can be poorly behaved in the tails, that is before the first knot and after the last knot. They cite advantages of constraining the function to be linear in the tails. Their restricted cubic spline function (also called *natural splines*) has the additional advantage that only $k-1$ parameters must be estimated (besides the intercept) as opposed to $k+3$ parameters with the unrestricted cubic spline. The restricted spline function with k knots $t_1, \ldots, t_k$ is given by[156]

$$f(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_{k-1} X_{k-1}, \tag{2.24}$$

where $X_1 = X$ and for $j = 1, \ldots, k-2$,

$$\begin{aligned} X_{j+1} = {} & (X - t_j)_+^3 - (X - t_{k-1})_+^3 (t_k - t_j)/(t_k - t_{k-1}) \\ & + (X - t_k)_+^3 (t_{k-1} - t_j)/(t_k - t_{k-1}). \end{aligned} \tag{2.25}$$

It can be shown that X_j is linear in X for $X \geq t_k$. For numerical behavior and to put all basis functions for X on the same scale, `R Hmisc` and `rms` package functions by default divide the terms in Eq. 2.25 by

$$\tau = (t_k - t_1)^2. \tag{2.26}$$

Figure 2.2 displays the τ-scaled spline component variables X_j for $j = 2, 3, 4$ and $k = 5$ and one set of knots. The left graph magnifies the lower portion of the curves.

```
require(Hmisc)
```

```
x ← rcspline.eval(seq(0,1,.01),
                  knots=seq(.05,.95,length=5), inclx=T)
xm ← x
xm[xm > .0106] ← NA
matplot(x[,1], xm, type="l", ylim=c(0,.01),
        xlab=expression(X), ylab='', lty=1)
matplot(x[,1], x,  type="l",
        xlab=expression(X), ylab='', lty=1)
```

Figure 2.3 displays some typical shapes of restricted cubic spline functions with $k = 3, 4, 5$, and 6. These functions were generated using random β.

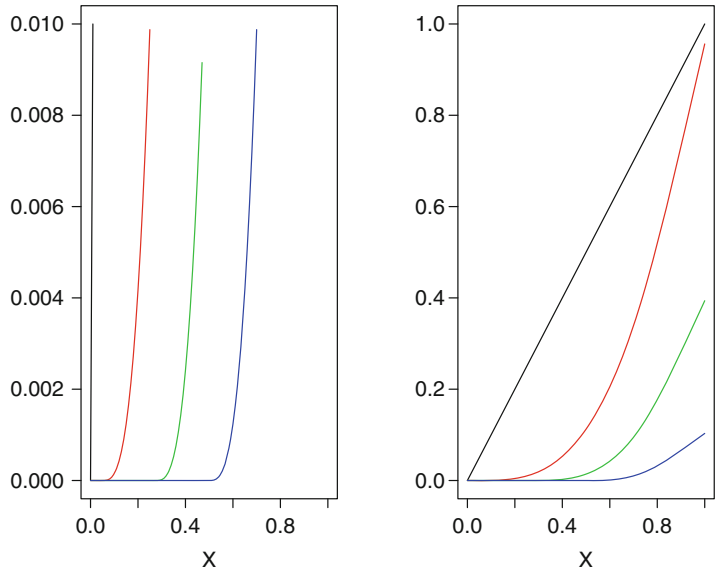

Fig. 2.2 Restricted cubic spline component variables for $k = 5$ and knots at $X = .05, .275, .5, .725$, and $.95$. Nonlinear basis functions are scaled by τ. The left panel is a y–magnification of the right panel. Fitted functions such as those in Figure 2.3 will be linear combinations of these basis functions as long as knots are at the same locations used here.

```
x ← seq(0, 1, length=300)
for(nk in 3:6) {
  set.seed(nk)
  knots ← seq(.05, .95, length=nk)
  xx ← rcspline.eval(x, knots=knots, inclx=T)
  for(i in 1 : (nk - 1))
    xx[,i] ← (xx[,i] - min(xx[,i])) /
              (max(xx[,i]) - min(xx[,i]))
  for(i in 1 : 20) {
    beta  ← 2*runif(nk-1) - 1
    xbeta ← xx %*% beta + 2 * runif(1) - 1
    xbeta ← (xbeta - min(xbeta)) /
             (max(xbeta) - min(xbeta))
    if(i == 1) {
      plot(x, xbeta, type="l", lty=1,
           xlab=expression(X), ylab='', bty="l")
      title(sub=paste(nk,"knots"), adj=0, cex=.75)
      for(j in 1 : nk)
        arrows(knots[j], .04, knots[j], -.03,
               angle=20, length=.07, lwd=1.5)
    }
    else lines(x, xbeta, col=i)
  }
}
```

Once $\beta_0, \ldots, \beta_{k-1}$ are estimated, the restricted cubic spline can be restated in the form

$$f(X) = \beta_0 + \beta_1 X + \beta_2 (X - t_1)_+^3 + \beta_3 (X - t_2)_+^3 + \ldots + \beta_{k+1} (X - t_k)_+^3 \quad (2.27)$$

by dividing $\beta_2, \ldots, \beta_{k-1}$ by τ (Eq. 2.26) and computing

$$\begin{aligned} \beta_k &= [\beta_2 (t_1 - t_k) + \beta_3 (t_2 - t_k) + \ldots + \beta_{k-1} (t_{k-2} - t_k)]/(t_k - t_{k-1}) \quad (2.28) \\ \beta_{k+1} &= [\beta_2 (t_1 - t_{k-1}) + \beta_3 (t_2 - t_{k-1}) + \ldots + \beta_{k-1} (t_{k-2} - t_{k-1})]/(t_{k-1} - t_k). \end{aligned}$$

A test of linearity in X can be obtained by testing

$$H_0 : \beta_2 = \beta_3 = \ldots = \beta_{k-1} = 0. \quad (2.29)$$

4 The truncated power basis for restricted cubic splines does allow for rational (i.e., linear) extrapolation beyond the outer knots. However, when the outer knots are in the tails of the data, extrapolation can still be dangerous.

When nonlinear terms in Equation 2.25 are normalized, for example, by dividing them by the square of the difference in the outer knots to make all terms have units of X, the ordinary truncated power basis has no numerical difficulties when modern matrix algebra software is used.

2.4.6 Choosing Number and Position of Knots

We have assumed that the locations of the knots are specified in advance; that is, the knot locations are not treated as free parameters to be estimated. If knots were free parameters, the fitted function would have more flexibility but at the cost of instability of estimates, statistical inference problems, and inability to use standard regression modeling software for estimating regression parameters.

How then does the analyst pre-assign knot locations? If the regression relationship were described by prior experience, pre-specification of knot locations would be easy. For example, if a function were known to change curvature at $X = a$, a knot could be placed at a. However, in most situations there is no way to pre-specify knots. Fortunately, Stone[593] has found that the location of knots in a restricted cubic spline model is not very crucial in most situations; the fit depends much more on the choice of k, the number of

5 knots. Placing knots at fixed quantiles (percentiles) of a predictor's marginal distribution is a good approach in most datasets. This ensures that enough points are available in each interval, and also guards against letting outliers overly influence knot placement. Recommended equally spaced quantiles are shown in Table 2.3.

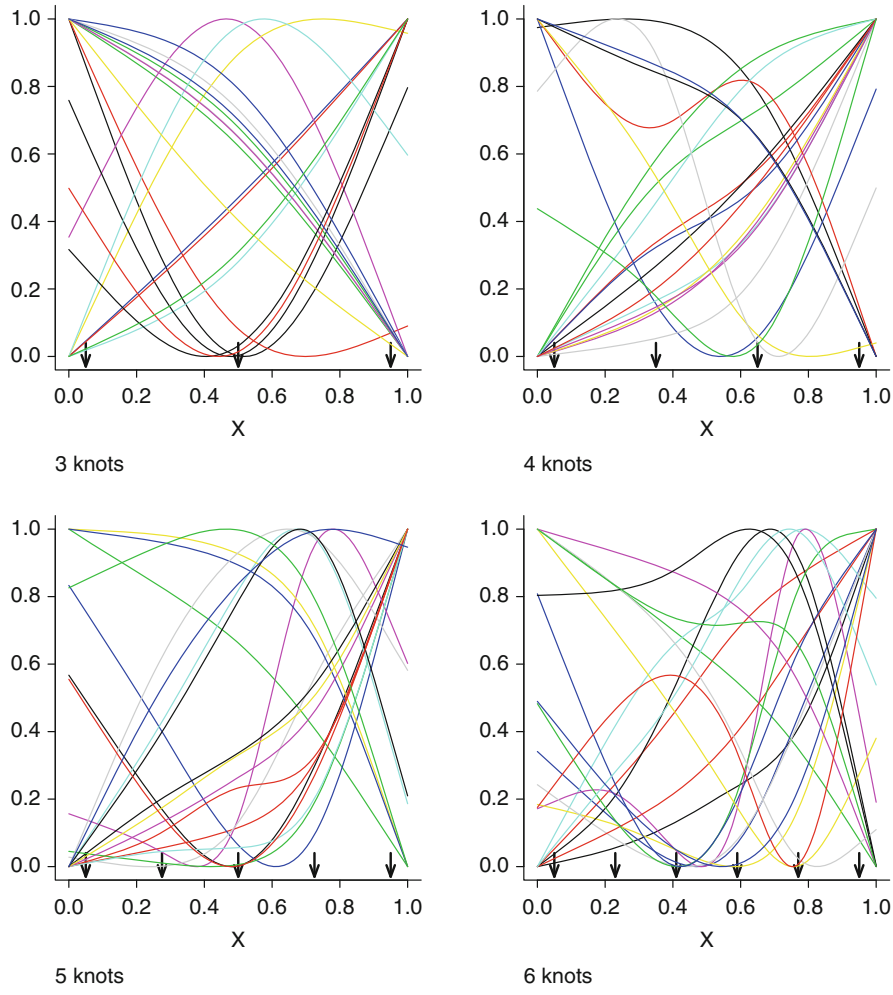

Fig. 2.3 Some typical restricted cubic spline functions for $k = 3, 4, 5, 6$. The y–axis is $X\beta$. Arrows indicate knots. These curves were derived by randomly choosing values of β subject to standard deviations of fitted functions being normalized.

Table 2.3 Default quantiles for knots

k	Quantiles						
3			.10	.5	.90		
4			.05	.35	.65	.95	
5		.05	.275	.5	.725	.95	
6	.05	.23	.41	.59	.77	.95	
7	.025	.1833	.3417	.5	.6583	.8167	.975

The principal reason for using less extreme default quantiles for $k = 3$ and more extreme ones for $k = 7$ is that one usually uses $k = 3$ for small sample sizes and $k = 7$ for large samples. When the sample size is less than 100, the outer quantiles should be replaced by the fifth smallest and fifth largest data points, respectively.[595] What about the choice of k? The flexibility of possible fits must be tempered by the sample size available to estimate the unknown parameters. Stone[593] has found that more than 5 knots are seldom required in a restricted cubic spline model. The principal decision then is between $k = 3, 4$, or 5. For many datasets, $k = 4$ offers an adequate fit of the model and is a good compromise between flexibility and loss of precision caused by overfitting a small sample. When the sample size is large (e.g., $n \geq 100$ with a continuous uncensored response variable), $k = 5$ is a good choice. Small samples (< 30, say) may require the use of $k = 3$. Akaike's information criterion (AIC, Section 9.8.1) can be used for a data-based choice of k. The value of k maximizing the model likelihood ratio $\chi^2 - 2k$ would be the best "for the money" using AIC.

The analyst may wish to devote more knots to variables that are thought to be more important, and risk lack of fit for less important variables. In this way the total number of estimated parameters can be controlled (Section 4.1).

2.4.7 Nonparametric Regression

One of the most important results of an analysis is the estimation of the tendency (trend) of how X relates to Y. This trend is useful in its own right and it may be sufficient for obtaining predicted values in some situations, but trend estimates can also be used to guide formal regression modeling (by suggesting predictor variable transformations) and to check model assumptions.

Nonparametric smoothers are excellent tools for determining the shape of the relationship between a predictor and the response. The standard nonparametric smoothers work when one is interested in assessing one continuous predictor at a time and when the property of the response that *should* be linearly related to the predictor is a standard measure of central tendency. For example, when $C(Y)$ is $E(Y)$ or $\Pr[Y = 1]$, standard smoothers are useful, but when $C(Y)$ is a measure of variability or a rate (instantaneous risk), or when Y is only incompletely measured for some subjects (e.g., Y is censored for some subjects), simple smoothers will not work.

The oldest and simplest nonparametric smoother is the moving average. Suppose that the data consist of the points $X = 1, 2, 3, 5$, and 8, with the corresponding Y values $2.1, 3.8, 5.7, 11.1$, and 17.2. To smooth the relationship we could estimate $E(Y|X = 2)$ by $(2.1 + 3.8 + 5.7)/3$ and $E(Y|X = (2 + 3 + 5)/3)$ by $(3.8 + 5.7 + 11.1)/3$. Note that overlap is fine; that is one point may be contained in two sets that are averaged. You can immediately see that the

simple moving average has a problem in estimating $E(Y)$ at the outer values of X. The estimates are quite sensitive to the choice of the number of points (or interval width) to use in "binning" the data.

A moving least squares linear regression smoother is far superior to a moving flat line smoother (moving average). Cleveland's[111] moving linear regression smoother *loess* has become the most popular smoother. To obtain the smoothed value of Y at $X = x$, we take all the data having X values within a suitable interval about x. Then a linear regression is fitted to all of these points, and the predicted value from this regression at $X = x$ is taken as the estimate of $E(Y|X = x)$. Actually, `loess` uses weighted least squares estimates, which is why it is called a *locally weighted least squares* method. The weights are chosen so that points near $X = x$ are given the most weight[b] in the calculation of the slope and intercept. Surprisingly, a good default choice for the interval about x is an interval containing 2/3 of the data points! The weighting function is devised so that points near the extremes of this interval receive almost no weight in the calculation of the slope and intercept.

Because `loess` uses a moving straight line rather than a moving flat one, it provides much better behavior at the extremes of the Xs. For example, one can fit a straight line to the first three data points and then obtain the predicted value at the lowest X, which takes into account that this X is not the middle of the three Xs.

`loess` obtains smoothed values for $E(Y)$ at each observed value of X. Estimates for other Xs are obtained by linear interpolation.

The `loess` algorithm has another component. After making an initial estimate of the trend line, `loess` can look for outliers off this trend. It can then delete or down-weight those apparent outliers to obtain a more robust trend estimate. Now, different points will appear to be outliers with respect to this second trend estimate. The new set of outliers is taken into account and another trend line is derived. By default, the process stops after these three iterations. `loess` works exceptionally well for binary Y as long as the iterations that look for outliers are not done, that is only one iteration is performed.

For a single X, Friedman's "super smoother"[207] is another efficient and flexible nonparametric trend estimator. For both `loess` and the super smoother the amount of smoothing can be controlled by the analyst. Hastie and Tibshirani[275] provided an excellent description of smoothing methods and developed a generalized additive model for multiple Xs, in which each continuous predictor is fitted with a nonparametric smoother (see Chapter 16). Interactions are not allowed. Cleveland et al.[96] have extended two-dimensional smoothers to multiple dimensions without assuming additivity. Their *local regression model* is feasible for up to four or so predictors. Local regression models are extremely flexible, allowing parts of the model to be

6

[b] This weight is not to be confused with the regression coefficient; rather the weights are $w_1, w_2, \ldots, w_n$ and the fitting criterion is $\sum_i^n w_i(Y_i - \hat{Y}_i)^2$.

parametrically specified, and allowing the analyst to choose the amount of smoothing or the effective number of degrees of freedom of the fit.

Smoothing splines are related to nonparametric smoothers. Here a knot is placed at every data point, but a penalized likelihood is maximized to derive the smoothed estimates. Gray[237,238] developed a general method that is halfway between smoothing splines and regression splines. He pre-specified, say, 10 fixed knots, but uses a penalized likelihood for estimation. This allows
7 the analyst to control the effective number of degrees of freedom used.

Besides using smoothers to estimate regression relationships, smoothers are valuable for examining trends in residual plots. See Sections 14.6 and 21.2 for examples.

2.4.8 Advantages of Regression Splines over Other Methods

There are several advantages of regression splines:[271]

1. Parametric splines are piecewise polynomials and can be fitted using any existing regression program after the constructed predictors are computed. Spline regression is equally suitable to multiple linear regression, survival models, and logistic models for discrete outcomes.
2. Regression coefficients for the spline function are estimated using standard techniques (maximum likelihood or least squares), and statistical inferences can readily be drawn. Formal tests of no overall association, linearity, and additivity can readily be constructed. Confidence limits for the estimated regression function are derived by standard theory.
3. The fitted spline function directly estimates the transformation that a predictor should receive to yield linearity in $C(Y|X)$. The fitted spline transformation sometimes suggests a simple transformation (e.g., square root) of a predictor that can be used if one is not concerned about the proper number of degrees of freedom for testing association of the predictor with the response.
4. The spline function can be used to represent the predictor in the final model. Nonparametric methods do not yield a prediction equation.
5. Splines can be extended to non-additive models (see below). Multidimensional nonparametric estimators often require burdensome computations.

2.5 Recursive Partitioning: Tree-Based Models

Breiman et al.[69] have developed an essentially model-free approach called *classification and regression trees* (CART), a form of recursive partitioning.

For some implementations of CART, we say "essentially" model-free since a model-based statistic is sometimes chosen as a splitting criterion. The essence of recursive partitioning is as follows.

1. Find the predictor so that the best possible binary split on that predictor has a larger value of some statistical criterion than any other split on any other predictor. For ordinal and continuous predictors, the split is of the form $X < c$ versus $X \geq c$. For polytomous predictors, the split involves finding the best separation of categories, without preserving order.
2. Within each previously formed subset, find the best predictor and best split that maximizes the criterion in the subset of observations passing the previous split.
3. Proceed in like fashion until fewer than k observations remain to be split, where k is typically 20 to 100.
4. Obtain predicted values using a statistic that summarizes each terminal node (e.g., mean or proportion).
5. Prune the tree backward so that a tree with the same number of nodes developed on 0.9 of the data validates best on the remaining 0.1 of the data (average over the 10 cross-validations). Alternatively, shrink the node estimates toward the mean, using a progressively stronger shrinkage factor, until the best cross-validation results.

Tree models have the advantage of not requiring any functional form for the predictors and of not assuming additivity of predictors (i.e., recursive partitioning can identify complex interactions). Trees can deal with missing data flexibly. They have the disadvantages of not utilizing continuous variables effectively and of overfitting in three directions: searching for best predictors, for best splits, and searching multiple times. The penalty for the extreme amount of data searching required by recursive partitioning surfaces when the tree does not cross-validate optimally until it is pruned all the way back to two or three splits. Thus reliable trees are often not very discriminating.

Tree models are especially useful in messy situations or settings in which overfitting is not so problematic, such as confounder adjustment using propensity scores[117] or in missing value imputation. A major advantage of tree modeling is savings of analyst time, but this is offset by the underfitting needed to make trees validate.

2.6 Multiple Degree of Freedom Tests of Association

When a factor is a linear or binary term in the regression model, the test of association for that factor with the response involves testing only a single regression parameter. Nominal factors and predictors that are represented as a quadratic or spline function require multiple regression parameters to be

tested simultaneously in order to assess association with the response. For a nominal factor having k levels, the overall ANOVA-type test with $k-1$ d.f. tests whether there are any differences in responses between the k categories. It is recommended that this test be done before attempting to interpret individual parameter estimates. If the overall test is not significant, it can be dangerous to rely on individual pairwise comparisons because the type I error will be increased. Likewise, for a continuous predictor for which linearity is not assumed, all terms involving the predictor should be tested simultaneously to check whether the factor is associated with the outcome. This test should precede the test for linearity and should usually precede the attempt to eliminate nonlinear terms. For example, in the model

$$C(Y|X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_2^2, \tag{2.30}$$

one should test $H_0 : \beta_2 = \beta_3 = 0$ with 2 d.f. to assess association between X_2 and outcome. In the five-knot restricted cubic spline model

$$C(Y|X) = \beta_0 + \beta_1 X + \beta_2 X' + \beta_3 X'' + \beta_4 X''', \tag{2.31}$$

the hypothesis $H_0 : \beta_1 = \ldots = \beta_4 = 0$ should be tested with 4 d.f. to assess whether there is any association between X and Y. If this 4 d.f. test is insignificant, it is dangerous to interpret the shape of the fitted spline function because the hypothesis that the overall function is flat has not been rejected.

A dilemma arises when an overall test of association, say one having 4 d.f., is insignificant, the 3 d.f. test for linearity is insignificant, but the 1 d.f. test for linear association, after deleting nonlinear terms, becomes significant. Had the test for linearity been borderline significant, it would not have been warranted to drop these terms in order to test for a linear association. But with the evidence for nonlinearity not very great, one could attempt to test for association with 1 d.f. This however is not fully justified, because the 1 d.f. test statistic does not have a χ^2 distribution with 1 d.f. since pretesting was done. The original 4 d.f. test statistic does have a χ^2 distribution with 4 d.f. because it was for a pre-specified test.

For quadratic regression, Grambsch and O'Brien[234] showed that the 2 d.f. test of association is nearly optimal when pretesting is done, even when the true relationship is linear. They considered an ordinary regression model $E(Y|X) = \beta_0 + \beta_1 X + \beta_2 X^2$ and studied tests of association between X and Y. The strategy they studied was as follows. First, fit the quadratic model and obtain the partial test of $H_0 : \beta_2 = 0$, that is the test of linearity. If this partial F-test is significant at the $\alpha = 0.05$ level, report as the final test of association between X and Y the 2 d.f. F-test of $H_0 : \beta_1 = \beta_2 = 0$. If the test of linearity is insignificant, the model is refitted without the quadratic term and the test of association is then a 1 d.f. test, $H_0 : \beta_1 = 0 | \beta_2 = 0$. Grambsch and O'Brien demonstrated that the type I error from this two-stage test is greater than the stated α, and in fact a fairly accurate P-value can be obtained if it is computed from an F distribution with 2 numerator

d.f. even when testing at the second stage. This is because in the original 2 d.f. test of association, the 1 d.f. corresponding to the nonlinear effect is deleted if the nonlinear effect is very small; that is one is retaining the most significant part of the 2 d.f. F statistic.

If we use a 2 d.f. F critical value to assess the X effect even when X^2 is not in the model, it is clear that the two-stage approach can only lose power and hence it has no advantage whatsoever. That is because the sum of squares due to regression from the quadratic model is greater than the sum of squares computed from the linear model.

2.7 Assessment of Model Fit

2.7.1 Regression Assumptions

In this section, the regression part of the model is isolated, and methods are described for validating the regression assumptions or modifying the model to meet the assumptions. The general linear regression model is

$$C(Y|X) = X\beta = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k. \tag{2.32}$$

The assumptions of linearity and additivity need to be verified. We begin with a special case of the general model,

$$C(Y|X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2, \tag{2.33}$$

where X_1 is binary and X_2 is continuous. One needs to verify that the property of the response $C(Y)$ is related to X_1 and X_2 according to Figure 2.4.

There are several methods for checking the fit of this model. The first method below is based on critiquing the simple model, and the other methods directly "estimate" the model.

1. Fit the simple linear additive model and critically examine residual plots for evidence of systematic patterns. For least squares fits one can compute estimated residuals $e = Y - X\hat{\beta}$ and box plots of e stratified by X_1 and scatterplots of e versus X_1 and $\hat{Y}$ with trend curves. If one is assuming constant conditional variance of Y, the spread of the residual distribution against each of the variables can be checked at the same time. If the normality assumption is needed (i.e., if significance tests or confidence limits are used), the distribution of e can be compared with a normal distribution with mean zero. **Advantage**: Simplicity. **Disadvantages**: Standard residuals can only be computed for continuous uncensored response variables. The judgment of non-randomness is largely subjective, it is difficult to detect interaction, and if interaction is present it is difficult to check any of the other assumptions. Unless trend lines are added to plots, pat-

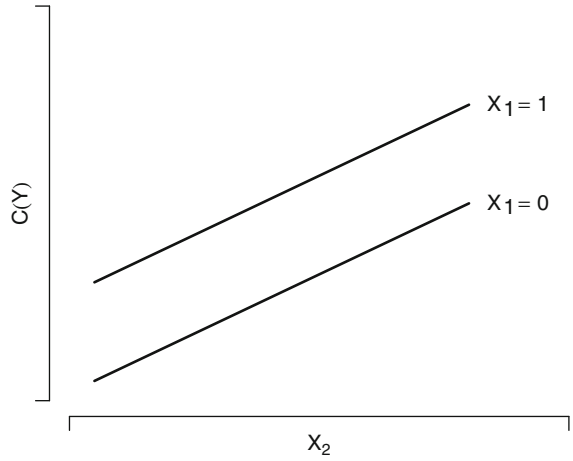

Fig. 2.4 Regression assumptions for one binary and one continuous predictor

terns may be difficult to discern if the sample size is very large. Detecting patterns in residuals does not always inform the analyst of what corrective action to take, although partial residual plots can be used to estimate the needed transformations if interaction is absent.

2. Make a scatterplot of Y versus X_2 using different symbols according to values of X_1. **Advantages**: Simplicity, and one can sometimes see all regression patterns including interaction. **Disadvantages**: Scatterplots cannot be drawn for binary, categorical, or censored Y. Patterns are difficult to see if relationships are weak or if the sample size is very large.
3. Stratify the sample by X_1 and quantile groups (e.g., deciles) of X_2. Within each $X_1 \times X_2$ stratum an estimate of $C(Y|X_1, X_2)$ is computed. If X_1 is continuous, the same method can be used after grouping X_1 into quantile groups. **Advantages**: Simplicity, ability to see interaction patterns, can handle censored Y if care is taken. **Disadvantages**: Subgrouping requires relatively large sample sizes and does not use continuous factors effectively as it does not attempt any interpolation. The ordering of quantile groups is not utilized by the procedure. Subgroup estimates have low precision (see p. 488 for an example). Each stratum must contain enough information to allow trends to be apparent above noise in the data. The method of grouping chosen (e.g., deciles vs. quintiles vs. rounding) can alter the shape of the plot.
4. Fit a nonparametric smoother separately for levels of X_1 (Section 2.4.7) relating X_2 to Y. **Advantages**: All regression aspects of the model can be summarized efficiently with minimal assumptions. **Disadvantages**: Does not easily apply to censored Y, and does not easily handle multiple predictors.

5. Fit a flexible parametric model that allows for most of the departures from the linear additive model that you wish to entertain. **Advantages**: One framework is used for examining the model assumptions, fitting the model, and drawing formal inference. Degrees of freedom are well defined and all aspects of statistical inference "work as advertised." **Disadvantages**: Complexity, and it is generally difficult to allow for interactions when assessing patterns of effects.

The first four methods each have the disadvantage that if confidence limits or formal inferences are desired it is difficult to know how many degrees of freedom were effectively used so that, for example, confidence limits will have the stated coverage probability. For method five, the restricted cubic spline function is an excellent tool for estimating the true relationship between X_2 and $C(Y)$ for continuous variables without assuming linearity. By fitting a model containing X_2 expanded into $k-1$ terms, where k is the number of knots, one can obtain an estimate of the function of X_2 that could be used linearly in the model:

$$\begin{aligned}\hat{C}(Y|X) &= \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \hat{\beta}_3 X_2' + \hat{\beta}_4 X_2'' \\ &= \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{f}(X_2), \end{aligned} \tag{2.34}$$

where

$$\hat{f}(X_2) = \hat{\beta}_2 X_2 + \hat{\beta}_3 X_2' + \hat{\beta}_4 X_2'', \tag{2.35}$$

and X_2' and X_2'' are constructed spline variables (when $k = 4$) as described previously. We call $\hat{f}(X_2)$ the spline-estimated transformation of X_2. Plotting the estimated spline function $\hat{f}(X_2)$ versus X_2 will generally shed light on how the effect of X_2 should be modeled. If the sample is sufficiently large, the spline function can be fitted separately for $X_1 = 0$ and $X_1 = 1$, allowing detection of even unusual interaction patterns. A formal test of linearity in X_2 is obtained by testing $H_0 : \beta_3 = \beta_4 = 0$, using a computationally efficient score test, for example (Section 9.2.3).

If the model is nonlinear in X_2, either a transformation suggested by the spline function plot (e.g., $\log(X_2)$) or the spline function itself (by placing X_2, X_2', and X_2'' simultaneously in any model fitted) may be used to describe X_2 in the model. If a tentative transformation of X_2 is specified, say $g(X_2)$, the adequacy of this transformation can be tested by expanding $g(X_2)$ in a spline function and testing for linearity. If one is concerned only with prediction and not with statistical inference, one can attempt to find a simplifying transformation for a predictor by plotting $g(X_2)$ against $\hat{f}(X_2)$ (the estimated spline transformation) for a variety of g, seeking a linearizing transformation of X_2. When there are nominal or binary predictors in the model in addition to the continuous predictors, it should be noted that there are no shape assumptions to verify for the binary/nominal predictors. One need only test for interactions between these predictors and the others.

If the model contains more than one continuous predictor, all may be expanded with spline functions in order to test linearity or to describe nonlinear relationships. If one did desire to assess simultaneously, for example, the linearity of predictors X_2 and X_3 in the presence of a linear or binary predictor X_1, the model could be specified as

$$\begin{aligned} C(Y|X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_2' + \beta_4 X_2'' \\ + \beta_5 X_3 + \beta_6 X_3' + \beta_7 X_3'', \end{aligned} \tag{2.36}$$

where X_2', X_2'', X_3', and X_3'' represent components of four knot restricted cubic spline functions.

The test of linearity for X_2 (with 2 d.f.) is $H_0 : \beta_3 = \beta_4 = 0$. The overall test of linearity for X_2 and X_3 is $H_0 : \beta_3 = \beta_4 = \beta_6 = \beta_7 = 0$, with 4 d.f. But as described further in Section 4.1, even though there are many reasons for allowing relationships to be nonlinear, there are reasons for not testing the nonlinear components for significance, as this might tempt the analyst to simplify the model thus distorting inference.[234] Testing for linearity is usually best done to justify to non-statisticians the need for complexity to explain or predict outcomes.

2.7.2 Modeling and Testing Complex Interactions

For testing interaction between X_1 and X_2 (after a needed transformation may have been applied), often a product term (e.g., X_1X_2) can be added to the model and its coefficient tested. A more general simultaneous test of linearity and lack of interaction for a two-variable model in which one variable is binary (or is assumed linear) is obtained by fitting the model

$$\begin{aligned} C(Y|X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_2' + \beta_4 X_2'' \\ + \beta_5 X_1 X_2 + \beta_6 X_1 X_2' + \beta_7 X_1 X_2'' \end{aligned} \tag{2.37}$$

and testing $H_0 : \beta_3 = \ldots = \beta_7 = 0$. This formulation allows the shape of the X_2 effect to be completely different for each level of X_1. There is virtually no departure from linearity and additivity that cannot be detected from this expanded model formulation if the number of knots is adequate and X_1 is binary. For binary logistic models, this method is equivalent to fitting two separate spline regressions in X_2.

10

Interactions can be complex when all variables are continuous. An approximate approach is to reduce the variables to two transformed variables, in which case interaction may sometimes be approximated by a single product of the two new variables. A disadvantage of this approach is that the estimates of the transformations for the two variables will be different depending

on whether interaction terms are adjusted for when estimating "main effects." A good compromise method involves fitting interactions of the form $X_1 f(X_2)$ and $X_2 g(X_1)$:

$$\begin{aligned} C(Y|X) = \beta_0 &+ \beta_1 X_1 + \beta_2 X_1' + \beta_3 X_1'' \\ &+ \beta_4 X_2 + \beta_5 X_2' + \beta_6 X_2'' \\ &+ \beta_7 X_1 X_2 + \beta_8 X_1 X_2' + \beta_9 X_1 X_2'' \\ &+ \beta_{10} X_2 X_1' + \beta_{11} X_2 X_1'' \end{aligned} \tag{2.38}$$

(for $k = 4$ knots for both variables). The test of additivity is $H_0 : \beta_7 = \beta_8 = \ldots = \beta_{11} = 0$ with 5 d.f. A test of lack of fit for the simple product interaction with X_2 is $H_0 : \beta_8 = \beta_9 = 0$, and a test of lack of fit for the simple product interaction with X_1 is $H_0 : \beta_{10} = \beta_{11} = 0$.

A general way to model and test interactions, although one requiring a larger number of parameters to be estimated, is based on modeling the $X_1 \times X_2 \times Y$ relationship with a smooth three-dimensional surface. A cubic spline surface can be constructed by covering the $X_1 - X_2$ plane with a grid and fitting a patch-wise cubic polynomial in two variables. The grid is $(u_i, v_j), i = 1, \ldots, k, j = 1, \ldots, k$, where knots for X_1 are $(u_1, \ldots, u_k)$ and knots for X_2 are $(v_1, \ldots, v_k)$. The number of parameters can be reduced by constraining the surface to be of the form $aX_1 + bX_2 + cX_1X_2$ in the lower left and upper right corners of the plane. The resulting restricted cubic spline surface is described by a multiple regression model containing spline expansions in X_1 and X_2 and all cross-products of the restricted cubic spline components (e.g., X_1X_2'). If the same number of knots k is used for both predictors, the number of interaction terms is $(k-1)^2$. Examples of various ways of modeling interaction are given in Chapter 10. Spline functions made up of cross-products of all terms of individual spline functions are called *tensor splines*.[50,274]

11

The presence of more than two predictors increases the complexity of tests for interactions because of the number of two-way interactions and because of the possibility of interaction effects of order higher than two. For example, in a model containing age, sex, and diabetes, the important interaction could be that older male diabetics have an exaggerated risk. However, higher-order interactions are often ignored unless specified a priori based on knowledge of the subject matter. Indeed, the number of two-way interactions alone is often too large to allow testing them all with reasonable power while controlling multiple comparison problems. Often, the only two-way interactions we can afford to test are those that were thought to be important before examining the data. A good approach is to test for all such pre-specified interaction effects with a single global (pooled) test. Then, unless interactions involving only one of the predictors are of special interest, one can either drop all interactions or retain all of them.

For some problems a reasonable approach is, for each predictor separately, to test simultaneously the joint importance of all interactions involving that predictor. For p predictors this results in p tests each with $p-1$ degrees of freedom. The multiple comparison problem would then be reduced from $p(p-1)/2$ tests (if all two-way interactions were tested individually) to p tests.

In the fields of biostatistics and epidemiology, some types of interactions that have consistently been found to be important in predicting outcomes and thus may be pre-specified are the following.

1. Interactions between treatment and the severity of disease being treated. Patients with little disease can receive little benefit.
2. Interactions involving age and risk factors. Older subjects are generally less affected by risk factors. They had to have been robust to survive to their current age with risk factors present.
3. Interactions involving age and type of disease. Some diseases are incurable and have the same prognosis regardless of age. Others are treatable or have less effect on younger patients.
4. Interactions between a measurement and the state of a subject during a measurement. Respiration rate measured during sleep may have greater predictive value and thus have a steeper slope versus outcome than respiration rate measured during activity.
5. Interaction between menopausal status and treatment or risk factors.
6. Interactions between race and disease.
7. Interactions between calendar time and treatment. Some treatments have learning curves causing secular trends in the associations.
8. Interactions between month of the year and other predictors, due to seasonal effects.
9. Interaction between the quality and quantity of a symptom, for example, daily frequency of chest pain × severity of a typical pain episode.
10. Interactions between study center and treatment.

12

2.7.3 Fitting Ordinal Predictors

For the case of an ordinal predictor, spline functions are not useful unless there are so many categories that in essence the variable is continuous. When the number of categories k is small (three to five, say), the variable is usually modeled as a polytomous factor using indicator variables or equivalently as one linear term and $k-2$ indicators. The latter coding facilitates testing for linearity. For more categories, it may be reasonable to stratify the data by levels of the variable and to compute summary statistics (e.g., logit proportions for a logistic model) or to examine regression coefficients associated with indicator variables over categories. Then one can attempt to summarize the pattern with a linear or some other simple trend. Later hypothesis tests

must take into account this data-driven scoring (by using > 1 d.f., for example), but the scoring can save degrees of freedom when testing for interaction with other factors. In one dataset, the number of comorbid diseases was used to summarize the risk of a set of diseases that was too large to model. By plotting the logit of the proportion of deaths versus the number of diseases, it was clear that the square of the number of diseases would properly score the variables.

Sometimes it is useful to code an ordinal predictor with $k-1$ indicator variables of the form $[X \geq v_j]$, where $j = 2, \ldots, k$ and $[h]$ is 1 if h is true, 0 otherwise.[648] Although a test of linearity does not arise immediately from this coding, the regression coefficients are interpreted as amounts of change from the previous category. A test of whether the last m categories can be combined with the category $k-m$ does follow easily from this coding.

2.7.4 Distributional Assumptions

The general linear regression model is stated as $C(Y|X) = X\beta$ to highlight its regression assumptions. For logistic regression models for binary or nominal responses, there is no distributional assumption if simple random sampling is used and subjects' responses are independent. That is, the binary logistic model and all of its assumptions are contained in the expression $\text{logit}\{Y = 1|X\} = X\beta$. For ordinary multiple regression with constant variance σ^2, we usually assume that $Y - X\beta$ is normally distributed with mean 0 and variance σ^2. This assumption can be checked by estimating β with $\hat{\beta}$ and plotting the overall distribution of the residuals $Y - X\hat{\beta}$, the residuals against $\hat{Y}$, and the residuals against each X. For the latter two, the residuals should be normally distributed within each neighborhood of $\hat{Y}$ or X. A weaker requirement is that the overall distribution of residuals is normal; this will be satisfied if all of the stratified residual distributions are normal. Note a hidden assumption in both models, namely, that there are no omitted predictors. Other models, such as the Weibull survival model or the Cox[132] proportional hazards model, also have distributional assumptions that are not fully specified by $C(Y|X) = X\beta$. However, regression and distributional assumptions of some of these models are encapsulated by

$$C(Y|X) = C(Y = y|X) = d(y) + X\beta \tag{2.39}$$

for some choice of C. Here $C(Y = y|X)$ is a property of the response Y evaluated at $Y = y$, given the predictor values X, and $d(y)$ is a component of the distribution of Y. For the Cox proportional hazards model, $C(Y = y|X)$ can be written as the log of the hazard of the event at time y, or equivalently as the log of the $-\log$ of the survival probability at time y, and $d(y)$ can be thought of as a log hazard function for a "standard" subject.

If we evaluated the property $C(Y = y|X)$ at predictor values X^1 and X^2, the difference in properties is

$$\begin{aligned} C(Y = y|X^1) - C(Y = y|X^2) &= d(y) + X^1\beta \\ &\quad - [d(y) + X^2\beta] \\ &= (X^1 - X^2)\beta, \end{aligned} \tag{2.40}$$

which is independent of y. One way to verify part of the distributional assumption is to estimate $C(Y = y|X^1)$ and $C(Y = y|X^2)$ for set values of X^1 and X^2 using a method that does not make the assumption, and to plot $C(Y = y|X^1) - C(Y = y|X^2)$ versus y. This function should be flat if the distributional assumption holds. The assumption can be tested formally if $d(y)$ can be generalized to be a function of X as well as y. A test of whether $d(y|X)$ depends on X is a test of one part of the distributional assumption. For example, writing $d(y|X) = d(y) + X\Gamma \log(y)$ where

$$X\Gamma = \Gamma_1 X_1 + \Gamma_2 X_2 + \ldots + \Gamma_k X_k \tag{2.41}$$

and testing $H_0 : \Gamma_1 = \ldots = \Gamma_k = 0$ is one way to test whether $d(y|X)$ depends on X. For semiparametric models such as the Cox proportional hazards model, the only distributional assumption is the one stated above, namely, that the difference in properties between two subjects depends only on the difference in the predictors between the two subjects. Other, parametric, models assume in addition that the property $C(Y = y|X)$ has a specific shape as a function of y, that is that $d(y)$ has a specific functional form. For example, the Weibull survival model has a specific assumption regarding the shape of the hazard or survival distribution as a function of y.

Assessments of distributional assumptions are best understood by applying these methods to individual models as is demonstrated in later chapters.

2.8 Further Reading

1. References [152, 575, 578] have more information about cubic splines.
2. See Smith[578] for a good overview of spline functions.
3. More material about natural splines may be found in de Boor[152]. McNeil et al.[451] discuss the overall smoothness of natural splines in terms of the integral of the square of the second derivative of the regression function, over the range of the data. Govindarajulu et al.[230] compared restricted cubic splines, penalized splines, and fractional polynomial[532] fits and found that the first two methods agreed with each other more than with estimated fractional polynomials.
4. A tutorial on restricted cubic splines is in [271].
5. Durrleman and Simon[168] provide examples in which knots are allowed to be estimated as free parameters, jointly with the regression coefficients. They found that even though the "optimal" knots were often far from a priori knot locations, the model fits were virtually identical.

6 Contrast Hastie and Tibshirani's generalized nonparametric additive models[275] with Stone and Koo's[595] additive model in which each continuous predictor is represented with a restricted cubic spline function.

7 Gray[237,238] provided some comparisons with ordinary regression splines, but he compared penalized regression splines with non-restricted splines with only two knots. Two knots were chosen so as to limit the degrees of freedom needed by the regression spline method to a reasonable number. Gray argued that regression splines are sensitive to knot locations, and he is correct when only two knots are allowed and no linear tail restrictions are imposed. Two knots also prevent the (ordinary maximum likelihood) fit from utilizing some local behavior of the regression relationship. For penalized likelihood estimation using B-splines, Gray[238] provided extensive simulation studies of type I and II error for testing association in which the true regression function, number of knots, and amount of likelihood penalization were varied. He studied both normal regression and Cox regression.

8 Breiman et al.'s original CART method[69] used the Gini criterion for splitting. Later work has used log-likelihoods.[109] Segal,[562] LeBlanc and Crowley,[389] and Ciampi et al.[107,108] and Keleş and Segal[342]have extended recursive partitioning to censored survival data using the log-rank statistic as the criterion. Zhang[682] extended tree models to handle multivariate binary responses. Schmoor et al.[556] used a more general splitting criterion that is useful in therapeutic trials, namely, a Cox test for main and interacting effects. Davis and Anderson[149] used an exponential survival model as the basis for tree construction. Ahn and Loh[7] developed a Cox proportional hazards model adaptation of recursive partitioning along with bootstrap and cross-validation-based methods to protect against "over-splitting." The Cox-based regression tree methods of Ciampi et al.[107] have a unique feature that allows for construction of "treatment interaction trees" with hierarchical adjustment for baseline variables. Zhang et al.[683] provided a new method for handling missing predictor values that is simpler than using surrogate splits. See [34,140,270,629] for examples using recursive partitioning for binary responses in which the prediction trees did not validate well.

9 [443,629] discuss other problems with tree models.

10 For ordinary linear models, the regression estimates are the same as obtained with separate fits, but standard errors are different (since a pooled standard error is used for the combined fit). For Cox[132] regression, separate fits can be slightly different since each subset would use a separate ranking of Y.

11 Gray's penalized fixed-knot regression splines can be useful for estimating joint effects of two continuous variables while allowing the analyst to control the effective number of degrees of freedom in the fit [237,238, Section 3.2]. When Y is a non-censored variable, the local regression model of Cleveland et al.,[96] a multidimensional scatterplot smoother mentioned in Section 2.4.7, provides a good graphical assessment of the joint effects of several predictors so that the forms of interactions can be chosen. See Wang et al.[653] and Gustafson[248] for several other flexible approaches to analyzing interactions among continuous variables.

12 Study site by treatment interaction is often the interaction that is worried about the most in multi-center randomized clinical trials, because regulatory agencies are concerned with consistency of treatment effects over study centers. However, this type of interaction is usually the weakest and is difficult to assess when there are many centers due to the number of interaction parameters to estimate. Schemper[545] discusses various types of interactions and a general nonparametric test for interaction.

2.9 Problems

For problems 1 to 3, state each model statistically, identifying each predictor with one or more component variables. Identify and interpret each regression parameter except for coefficients of nonlinear terms in spline functions. State each hypothesis below as a formal statistical hypothesis involving the proper parameters, and give the (numerator) degrees of freedom of the test. State alternative hypotheses carefully with respect to unions or intersections of conditions and list the type of alternatives to the null hypothesis that the test is designed to detect.[c]

1. A property of Y such as the mean is linear in age and blood pressure and there may be an interaction between the two predictors. Test H_0 : there is no interaction between age and blood pressure. Also test H_0 : blood pressure is not associated with Y (in any fashion). State the effect of blood pressure as a function of age, and the effect of age as a function of blood pressure.
2. Consider a linear additive model involving three treatments (control, drug Z, and drug Q) and one continuous adjustment variable, age. Test H_0 : treatment group is not associated with response, adjusted for age. Also test H_0 : response for drug Z has the same property as the response for drug Q, adjusted for age.
3. Consider models each with two predictors, temperature and white blood count (WBC), for which temperature is always assumed to be linearly related to the appropriate property of the response, and WBC may or may not be linear (depending on the particular model you formulate for each question). Test:
   a. H_0 : WBC is not associated with the response versus H_a : WBC is linearly associated with the property of the response.
   b. H_0 : WBC is not associated with Y versus H_a : WBC is quadratically associated with Y. Also write down the formal test of linearity against this quadratic alternative.
   c. H_0 : WBC is not associated with Y versus H_a : WBC related to the property of the response through a smooth spline function; for example, for WBC the model requires the variables WBC, WBC$'$, and WBC$''$ where WBC$'$ and WBC$''$ represent nonlinear components (if there are four knots in a restricted cubic spline function). Also write down the formal test of linearity against this spline function alternative.
   d. Test for a lack of fit (combined nonlinearity or non-additivity) in an overall model that takes the form of an interaction between temperature and WBC, allowing WBC to be modeled with a smooth spline function.
4. For a fitted model $Y = a + bX + cX^2$ derive the estimate of the effect on Y of changing X from x_1 to x_2.

[c] In other words, under what assumptions does the test have maximum power?

5. In "The Class of 1988: A Statistical Portrait," the College Board reported mean SAT scores for each state. Use an ordinary least squares multiple regression model to study the mean verbal SAT score as a function of the percentage of students taking the test in each state. Provide plots of fitted functions and defend your choice of the "best" fit. Make sure the shape of the chosen fit agrees with what you know about the variables. Add the raw data points to plots.

   a. Fit a linear spline function with a knot at $X = 50\%$. Plot the data and the fitted function and do a formal test for linearity and a test for association between X and Y. Give a detailed interpretation of the estimated coefficients in the linear spline model, and use the partial t-test to test linearity in this model.
   b. Fit a restricted cubic spline function with knots at $X = 6$, 12, 58, and 68% (not percentile).[d] Plot the fitted function and do a formal test of association between X and Y. Do two tests of linearity that test the same hypothesis:
      i. by using a *contrast* to simultaneously test the correct set of coefficients against zero (done by the `anova` function in `rms`);[e]
      ii. by comparing the R^2 from the complex model with that from a simple linear model using a partial F-test.

      Explain why the tests of linearity have the d.f. they have.
   c. Using subject matter knowledge, pick a final model (from among the previous models or using another one) that makes sense.

   The data are found in Table 2.4 and may be created in `R` using the `sat.r` code on the RMS course web site.
6. Derive the formulas for the restricted cubic spline component variables without cubing or squaring any terms.
7. Prove that each component variable is linear in X when $X \geq t_k$, the last knot, using general principles and not algebra or calculus. Derive an expression for the restricted spline regression function when $X \geq t_k$.
8. Consider a two–stage procedure in which one tests for linearity of the effect of a predictor X on a property of the response $C(Y|X)$ against a quadratic alternative. If the two–tailed test of linearity is significant at the α level, a two d.f. test of association between X and Y is done. If the test for linearity is not significant, the square term is dropped and a linear model is fitted. The test of association between X and Y is then (apparently) a one d.f. test.

   a. Write a formal expression for the test statistic for association.

[d] Note: To pre-specify knots for restricted cubic spline functions, use something like `rcs(predictor, c(t1,t2,t3,t4))`, where the knot locations are `t1, t2, t3, t4`.

[e] Note that `anova` in `rms` computes all needed test statistics from a single model fit object.

b. Write an expression for the nominal P–value for testing association using this strategy.
c. Write an expression for the actual P–value or alternatively for the type–I error if using a fixed critical value for the test of association.
d. For the same two–stage strategy consider an estimate of the effect on $C(Y|X)$ of increasing X from a to b. Write a brief symbolic algorithm for deriving a true two–sided $1-\alpha$ confidence interval for the $b:a$ effect (difference in $C(Y)$) using the bootstrap.

Table 2.4 SAT data from the College Board, 1988

% Taking SAT (X)	Mean Verbal Score (Y)	% Taking SAT (X)	Mean Verbal Score (Y)
4	482	24	440
5	498	29	460
5	513	37	448
6	498	43	441
6	511	44	424
7	479	45	417
9	480	49	422
9	483	50	441
10	475	52	408
10	476	55	412
10	487	57	400
10	494	58	401
12	474	59	430
12	478	60	433
13	457	62	433
13	485	63	404
14	451	63	424
14	471	63	430
14	473	64	431
16	467	64	437
17	470	68	446
18	464	69	424
20	471	72	420
22	455	73	432
23	452	81	436