Chapter 10
Binary Logistic Regression

10.1 Model

Binary responses are commonly studied in many fields. Examples include the presence or absence of a particular disease, death during surgery, or a consumer purchasing a product. Often one wishes to study how a set of predictor variables X is related to a dichotomous response variable Y. The predictors may describe such quantities as treatment assignment, dosage, risk factors, and calendar time. 1

For convenience we define the response to be $Y = 0$ or 1, with $Y = 1$ denoting the occurrence of the event of interest. Often a dichotomous outcome can be studied by calculating certain proportions, for example, the proportion of deaths among females and the proportion among males. However, in many situations, there are multiple descriptors, or one or more of the descriptors are continuous. Without a statistical model, studying patterns such as the relationship between age and occurrence of a disease, for example, would require the creation of arbitrary age groups to allow estimation of disease prevalence as a function of age.

Letting X denote the vector of predictors $\{X_1, X_2, \ldots, X_k\}$, a first attempt at modeling the response might use the ordinary linear regression model

$$E\{Y|X\} = X\beta, \tag{10.1}$$

since the expectation of a binary variable Y is Prob$\{Y = 1\}$. However, such a model by definition cannot fit the data over the whole range of the predictors since a purely linear model E$\{Y|X\}$ = Prob$\{Y = 1|X\} = X\beta$ can allow Prob$\{Y = 1\}$ to exceed 1 or fall below 0. The statistical model that is generally preferred for the analysis of binary responses is instead the binary logistic regression model, stated in terms of the probability that $Y = 1$ given X, the values of the predictors:

F.E. Harrell, Jr., *Regression Modeling Strategies*, Springer Series in Statistics, DOI 10.1007/978-3-319-19425-7_10

$$\text{Prob}\{Y = 1|X\} = [1 + \exp(-X\beta)]^{-1}. \tag{10.2}$$

As before, $X\beta$ stands for $\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k$. The binary logistic regression model was developed primarily by Cox[129] and Walker and Duncan.[647] The regression parameters β are estimated by the method of maximum likelihood (see below). 2

The function

$$P = [1 + \exp(-x)]^{-1} \tag{10.3}$$

is called the logistic function. This function is plotted in Figure 10.1 for x varying from -4 to $+4$. This function has an unlimited range for x while P is restricted to range from 0 to 1.

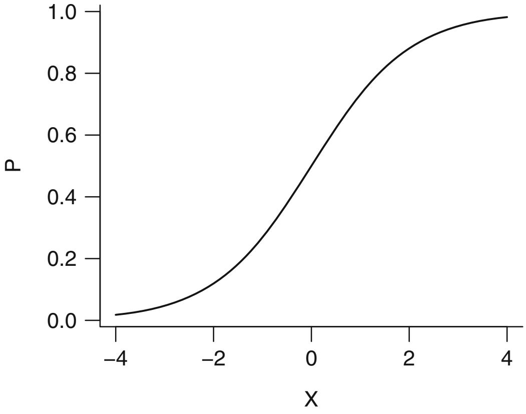

Fig. 10.1 Logistic function

For future derivations it is useful to express x in terms of P. Solving the equation above for x by using

$$1 - P = \exp(-x)/[1 + \exp(-x)] \tag{10.4}$$

yields the inverse of the logistic function:

$$x = \log[P/(1 - P)] = \log[\text{odds that } Y = 1 \text{ occurs}] = \text{logit}\{Y = 1\}. \tag{10.5}$$

Other methods that have been used to analyze binary response data include the probit model, which writes P in terms of the cumulative normal distribution, and discriminant analysis. Probit regression, although assuming a similar shape to the logistic function for the regression relationship between $X\beta$ and $\text{Prob}\{Y = 1\}$, involves more cumbersome calculations, and there is no natural interpretation of its regression parameters. In the past, discriminant analysis has been the predominant method since it is the simplest computationally. However, it makes more assumptions than logistic regression. The model used in discriminant analysis is stated in terms of the 3 distribution of X given the outcome group Y, even though one is seldom interested in the distribution of the predictors per se. The discriminant model has to be inverted using Bayes' rule to derive the quantity of primary interest, Prob$\{Y = 1\}$. By contrast, the logistic model is a *direct probability model* since it is stated in terms of Prob$\{Y = 1|X\}$. Since the distribution of a binary random variable Y is completely defined by the true probability that $Y = 1$ and since the model makes no assumption about the distribution of the predictors, the logistic model makes no distributional assumptions whatsoever.

10.1.1 Model Assumptions and Interpretation of Parameters

Since the logistic model is a direct probability model, its only assumptions relate to the form of the regression equation. Regression assumptions are verifiable, unlike the assumption of multivariate normality made by discriminant analysis. The logistic model assumptions are most easily understood by transforming Prob$\{Y = 1\}$ to make a model that is linear in $X\beta$:

$$\begin{aligned}\text{logit}\{Y = 1|X\} &= \text{logit}(P) = \log[P/(1 - P)]\\ &= X\beta, \qquad (10.6)\end{aligned}$$

where $P = \text{Prob}\{Y = 1|X\}$. Thus the model is a linear regression model in the log odds that $Y = 1$ since logit(P) is a weighted sum of the Xs. If all effects are additive (i.e., no interactions are present), the model assumes that for every predictor X_j,

$$\begin{aligned}\text{logit}\{Y = 1|X\} &= \beta_0 + \beta_1 X_1 + \ldots + \beta_j X_j + \ldots + \beta_k X_k\\ &= \beta_j X_j + C, \qquad (10.7)\end{aligned}$$

where if all other factors are held constant, C is a constant given by

$$C = \beta_0 + \beta_1 X_1 + \ldots + \beta_{j-1} X_{j-1} + \beta_{j+1} X_{j+1} + \ldots + \beta_k X_k. \qquad (10.8)$$

The parameter β_j is then the change in the log odds per unit change in X_j if X_j represents a single factor that is linear and does not interact with other factors and if all other factors are held constant. Instead of writing this relationship in terms of log odds, it could just as easily be written in terms of the odds that $Y = 1$:

$$\text{odds}\{Y = 1|X\} = \exp(X\beta), \qquad (10.9)$$

and if all factors other than X_j are held constant,

$$\text{odds}\{Y=1|X\}=\exp(\beta_j X_j + C)=\exp(\beta_j X_j)\exp(C). \tag{10.10}$$

The regression parameters can also be written in terms of *odds ratios*. The odds that $Y=1$ when X_j is increased by d, divided by the odds at X_j is

$$\begin{aligned}\frac{\text{odds}\{Y=1|X_1,X_2,\ldots,X_j+d,\ldots,X_k\}}{\text{odds}\{Y=1|X_1,X_2,\ldots,X_j,\ldots,X_k\}}&\\ =\frac{\exp[\beta_j(X_j+d)]\exp(C)}{[\exp(\beta_j X_j)\exp(C)]}&\\ =\exp[\beta_j X_j+\beta_j d-\beta_j X_j]=\exp(\beta_j d).&\end{aligned} \tag{10.11}$$

Thus the effect of increasing X_j by d is to increase the odds that $Y=1$ by a factor of $\exp(\beta_j d)$, or to increase the log odds that $Y=1$ by an increment of $\beta_j d$. In general, the ratio of the odds of response for an individual with predictor variable values X^* compared with an individual with predictors X is

$$\begin{aligned}X^*:X\text{ odds ratio}&=\exp(X^*\beta)/\exp(X\beta)\\ &=\exp[(X^*-X)\beta].\end{aligned} \tag{10.12}$$

Now consider some special cases of the logistic multiple regression model. If there is only one predictor X and that predictor is binary, the model can be written

$$\begin{aligned}\text{logit}\{Y=1|X=0\}&=\beta_0\\ \text{logit}\{Y=1|X=1\}&=\beta_0+\beta_1.\end{aligned} \tag{10.13}$$

Here β_0 is the log odds of $Y=1$ when $X=0$. By subtracting the two equations above, it can be seen that β_1 is the difference in the log odds when $X=1$ as compared with $X=0$, which is equivalent to the log of the ratio of the odds when $X=1$ compared with the odds when $X=0$. The quantity $\exp(\beta_1)$ is the odds ratio for $X=1$ compared with $X=0$. Letting $P^0=\text{Prob}\{Y=1|X=0\}$ and $P^1=\text{Prob}\{Y=1|X=1\}$, the regression parameters are interpreted by

$$\begin{aligned}\beta_0&=\text{logit}(P^0)=\log[P^0/(1-P^0)]\\ \beta_1&=\text{logit}(P^1)-\text{logit}(P^0)\\ &=\log[P^1/(1-P^1)]-\log[P^0/(1-P^0)]\\ &=\log\{[P^1/(1-P^1)]/[P^0/(1-P^0)]\}.\end{aligned} \tag{10.14}$$

Since there are only two quantities to model and two free parameters, there is no way that this two-sample model can't fit; the model in this case is essentially fitting two cell proportions. Similarly, if there are $g-1$ dummy indicator Xs representing g groups, the ANOVA-type logistic model must always fit.

If there is one continuous predictor X, the model is

$$\mathrm{logit}\{Y=1|X\} = \beta_0 + \beta_1 X, \tag{10.15}$$

and without further modification (e.g., taking log transformation of the predictor), the model assumes a straight line in the log odds, or that an increase in X by one unit increases the odds by a factor of $\exp(\beta_1)$.

Now consider the simplest analysis of covariance model in which there are two treatments (indicated by $X_1 = 0$ or 1) and one continuous covariable (X_2). The simplest logistic model for this setup is

$$\mathrm{logit}\{Y=1|X\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2, \tag{10.16}$$

which can be written also as

$$\begin{aligned}\mathrm{logit}\{Y=1|X_1=0,X_2\} &= \beta_0 + \beta_2 X_2\\ \mathrm{logit}\{Y=1|X_1=1,X_2\} &= \beta_0 + \beta_1 + \beta_2 X_2.\end{aligned} \tag{10.17}$$

The $X_1 = 1 : X_1 = 0$ odds ratio is $\exp(\beta_1)$, independent of X_2. The odds ratio for a one-unit increase in X_2 is $\exp(\beta_2)$, independent of X_1.

This model, with no term for a possible interaction between treatment and covariable, assumes that for each treatment the relationship between X_2 and log odds is linear, and that the lines have equal slope; that is, they are parallel. Assuming linearity in X_2, the only way that this model can fail is for the two slopes to differ. Thus, the only assumptions that need verification are linearity and lack of interaction between X_1 and X_2.

To adapt the model to allow or test for interaction, we write

$$\mathrm{logit}\{Y=1|X\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3, \tag{10.18}$$

where the derived variable X_3 is defined to be $X_1 X_2$. The test for lack of interaction (equal slopes) is $H_0 : \beta_3 = 0$. The model can be amplified as

$$\begin{aligned}\mathrm{logit}\{Y=1|X_1=0,X_2\} &= \beta_0 + \beta_2 X_2\\ \mathrm{logit}\{Y=1|X_1=1,X_2\} &= \beta_0 + \beta_1 + \beta_2 X_2 + \beta_3 X_2\\ &= \beta_0' + \beta_2' X_2,\end{aligned} \tag{10.19}$$

Table 10.1 Effect of an odds ratio of two on various risks

Without Risk Factor		With Risk Factor	
Probability	Odds	Odds	Probability
.2	.25	.5	.33
.5	1	2	.67
.8	4	8	.89
.9	9	18	.95
.98	49	98	.99

where $\beta_0' = \beta_0 + \beta_1$ and $\beta_2' = \beta_2 + \beta_3$. The model with interaction is therefore equivalent to fitting two separate logistic models with X_2 as the only predictor, one model for each treatment group. Here the $X_1 = 1 : X_1 = 0$ odds ratio is $\exp(\beta_1 + \beta_3 X_2)$.

10.1.2 Odds Ratio, Risk Ratio, and Risk Difference

As discussed above, the logistic model quantifies the effect of a predictor in terms of an odds ratio or log odds ratio. An odds ratio is a natural description of an effect in a probability model since an odds ratio *can* be constant. For example, suppose that a given risk factor doubles the odds of disease. Table 10.1 shows the effect of the risk factor for various levels of initial risk.

Since odds have an unlimited range, any positive odds ratio will still yield a valid probability. If one attempted to describe an effect by a risk ratio, the effect can only occur over a limited range of risk (probability). For example, a risk ratio of 2 can only apply to risks below .5; above that point the risk ratio must diminish. (Risk ratios are similar to odds ratios if the risk is small.) Risk differences have the same difficulty; the risk difference cannot be constant and must depend on the initial risk. Odds ratios, on the other hand, can describe an effect over the entire range of risk. An odds ratio can, for example, describe the effect of a treatment independently of covariables affecting risk. Figure 10.2 depicts the relationship between risk of a subject without the risk factor and the increase in risk for a variety of relative increases (odds ratios). It demonstrates how absolute risk increase is a function of the baseline risk. Risk increase will also be a function of factors that interact with the risk factor, that is, factors that modify its relative effect. Once a model is developed for estimating $\text{Prob}\{Y = 1|X\}$, this model can easily be used to estimate the absolute risk increase as a function of baseline risk factors as well as interacting factors. Let X_1 be a binary risk factor and let $A = \{X_2, \ldots, X_p\}$ be the other factors (which for convenience we assume do not interact with X_1). Then the estimate of $\text{Prob}\{Y = 1|X_1 = 1, A\} - \text{Prob}\{Y = 1|X_1 = 0, A\}$ is

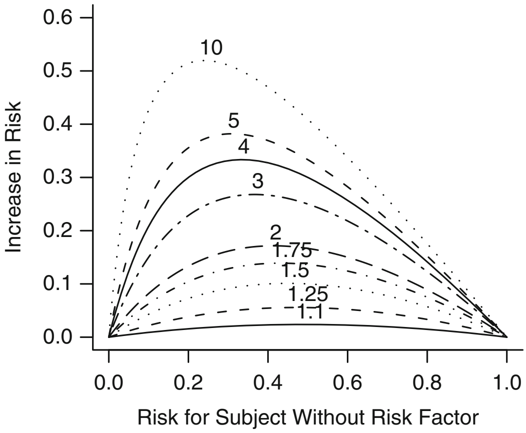

Fig. 10.2 Absolute benefit as a function of risk of the event in a control subject and the relative effect (odds ratio) of the risk factor. The odds ratios are given for each curve.

Table 10.2 Example binary response data

Females	**Age**:	37	39	39	42	47	48	48	52	53	55	56	57	58	58	60	64	65	68	68	70
	Response:	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	1	1	1	1
Males	**Age**:	34	38	40	40	41	43	43	43	44	46	47	48	48	50	50	52	55	60	61	61
	Response:	1	1	0	0	0	1	1	1	0	0	1	1	1	0	1	1	1	1	1	1

$$\begin{aligned} &\frac{1}{1+\exp-[\hat{\beta}_0+\hat{\beta}_1+\hat{\beta}_2X_2+\ldots+\hat{\beta}_pX_p]} \\ &\qquad -\frac{1}{1+\exp-[\hat{\beta}_0+\hat{\beta}_2X_2+\ldots+\hat{\beta}_pX_p]} \\ &\qquad\qquad =\frac{1}{1+(\frac{1-\hat{R}}{\hat{R}})\exp(-\hat{\beta}_1)}-\hat{R}, \end{aligned} \tag{10.20}$$

where $\hat{R}$ is the estimate of the baseline risk, Prob$\{Y=1|X_1=0\}$. The risk difference estimate can be plotted against $\hat{R}$ or against levels of variables in A to display absolute risk increase against overall risk (Figure 10.2) or against specific subject characteristics. 4

10.1.3 Detailed Example

Consider the data in Table 10.2. A graph of the data, along with a fitted logistic model (described later), appears in Figure 10.3. The graph also displays proportions of responses obtained by stratifying the data by sex and age group ($< 45, 45 - 54, \geq 55$). The age points on the abscissa for these groups are the overall mean ages in the three age intervals (40.2, 49.1, and 61.1, respectively).

```
require(rms)
```

```
getHdata(sex.age.response)
d ← sex.age.response
dd ← datadist(d); options(datadist='dd')
f ← lrm(response ~ sex + age, data=d)
fasr ← f   # Save for later
w ← function(...)
  with(d, {
    m ← sex=='male'
    f ← sex=='female'
    lpoints(age[f], response[f], pch=1)
    lpoints(age[m], response[m], pch=2)
    af ← cut2(age, c(45,55), levels.mean=TRUE)
    prop ← tapply(response, list(af, sex), mean,
                   na.rm=TRUE)
    agem ← as.numeric(row.names(prop))
    lpoints(agem, prop[,'female'],
            pch=4, cex=1.3, col='green')
    lpoints(agem, prop[,'male'],
            pch=5, cex=1.3, col='green')
    x ← rep(62, 4); y ← seq(.25, .1, length=4)
    lpoints(x, y, pch=c(1, 2, 4, 5),
            col=rep(c('blue','green'),each=2))
    ltext(x+5, y,
          c('F Observed','M Observed',
            'F Proportion','M Proportion'), cex=.8)
  } )   # Figure 10.3

plot(Predict(f, age=seq(34, 70, length=200), sex, fun=plogis),
     ylab='Pr[response]', ylim=c(-.02, 1.02), addpanel=w)
ltx ← function(fit) latex(fit, inline=TRUE, columns=54,
                           file='', after='$.', digits=3,
        size='Ssize', before='$X\\hat{\\beta}=')
ltx(f)
```

$X\hat{\beta} = -9.84 + 3.49[\text{male}] + 0.158\,\text{age}.$

Descriptive statistics for assessing the association between sex and response, age group and response, and age group and response stratified by sex are found below. Corresponding fitted logistic models, with sex coded as 0 = female, 1 = male are also given. Models were fitted first with sex as the only predictor, then with age as the (continuous) predictor, then with sex and age simultaneously. First consider the relationship between sex and response, ignoring the effect of age.

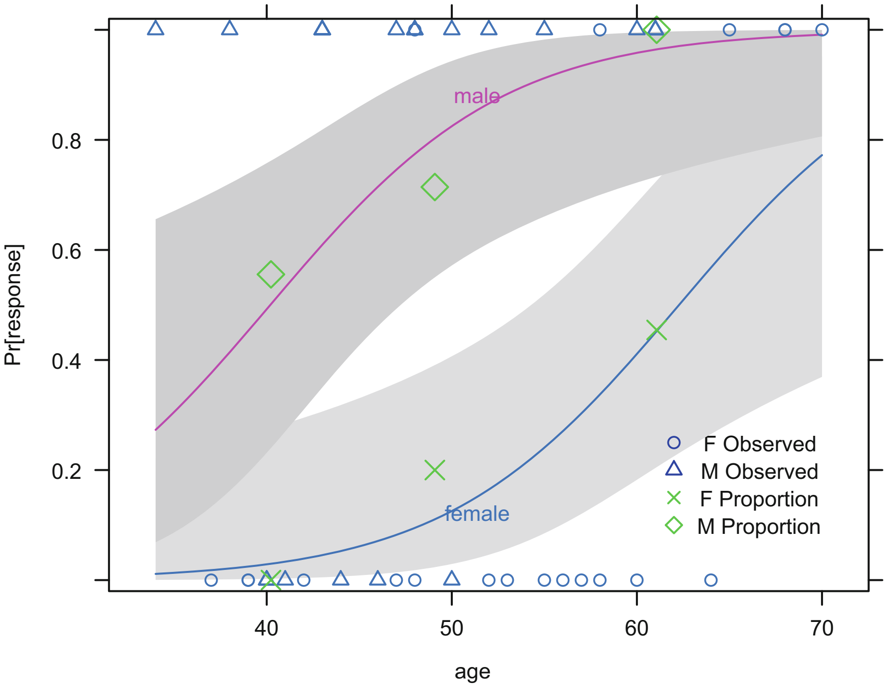

Fig. 10.3 Data, subgroup proportions, and fitted logistic model, with 0.95 pointwise confidence bands

```
sex         response
Frequency
Row Pct        0         1     Total      Odds/Log

F             14         6        20    6/14=.429
           70.00     30.00                  -.847

M              6        14        20    14/6=2.33
           30.00     70.00                   .847

Total         20        20        40

M:F odds ratio = (14/6)/(6/14) = 5.44, log=1.695
```

Statistics for sex × response

Statistic	d.f.	Value	*P*
χ^2	1	6.400	0.011
Likelihood Ratio χ^2	1	6.583	0.010

Parameter	Estimate	Std Err	Wald χ^2	P
β_0	-0.8473	0.4880	3.0152	
β_1	1.6946	0.6901	6.0305	0.0141

Note that the estimate of β_0, $\hat{\beta}_0$ is the log odds for females and that $\hat{\beta}_1$ is the log odds (M:F) ratio. $\hat{\beta}_0 + \hat{\beta}_1 = .847$, the log odds for males. The likelihood ratio test for H_0 : no effect of sex on probability of response is obtained as follows.

$$\begin{array}{ll} \text{Log likelihood } (\beta_1 = 0) & : -27.727 \\ \text{Log likelihood (max)} & : -24.435 \\ \text{LR } \chi^2(H_0 : \beta_1 = 0) & : -2(-27.727 - -24.435) = 6.584. \end{array}$$

(Note the agreement of the LR χ^2 with the contingency table likelihood ratio χ^2, and compare 6.584 with the Wald statistic 6.03.)

Next, consider the relationship between age and response, ignoring sex.

```
age        response
Frequency
Row Pct       0         1     Total      Odds/Log

<45           8         5        13      5/8=.625
           61.5      38.4                    -.47

45-54         6         6        12         6/6=1
           50.0      50.0                       0

55+           6         9        15       9/6=1.5
           40.0      60.0                    .405

Total        20        20        40

55+ : <45 odds ratio = (9/6)/(5/8) = 2.4, log=.875
```

Parameter	Estimate	Std Err	Wald χ^2	P
β_0	−2.7338	1.8375	2.2134	0.1368
β_1	0.0540	0.0358	2.2763	0.1314

The estimate of β_1 is in rough agreement with that obtained from the frequency table. The 55+ : < 45 log odds ratio is .875, and since the respective mean ages in the 55+ and <45 age groups are 61.1 and 40.2, an estimate of the log odds ratio increase per year is $.875/(61.1 - 40.2) = .875/20.9 = .042$.

The likelihood ratio test for H_0 : no association between age and response is obtained as follows.

$$\begin{array}{ll} \text{Log likelihood } (\beta_1 = 0) & : -27.727 \\ \text{Log likelihood (max)} & : -26.511 \\ \text{LR } \chi^2(H_0 : \beta_1 = 0) & : -2(-27.727 - -26.511) = 2.432. \end{array}$$

(Compare 2.432 with the Wald statistic 2.28.)

Next we consider the simultaneous association of age and sex with response.

```
                    sex=F

age          response
Frequency
Row Pct         0         1    Total

<45             4         0        4
            100.0       0.0

45-54           4         1        5
             80.0      20.0

55+             6         5       11
             54.6      45.4

Total          14         6       20

                    sex=M

age          response
Frequency
Row Pct         0         1    Total

<45             4         5        9
             44.4      55.6

45-54           2         5        7
             28.6      71.4

55+             0         4        4
              0.0     100.0

Total           6        14       20
```

A logistic model for relating sex and age simultaneously to response is given below.

Parameter	Estimate	Std Err	Wald χ^2	P
β_0	-9.8429	3.6758	7.1706	0.0074
β_1 (sex)	3.4898	1.1992	8.4693	0.0036
β_2 (age)	0.1581	0.0616	6.5756	0.0103

Likelihood ratio tests are obtained from the information below.

Log likelihood $(\beta_1 = 0, \beta_2 = 0)$	$: -27.727$
Log likelihood (max)	$: -19.458$
Log likelihood $(\beta_1 = 0)$	$: -26.511$
Log likelihood $(\beta_2 = 0)$	$: -24.435$
LR χ^2 $(H_0 : \beta_1 = \beta_2 = 0)$	$: -2(-27.727 - -19.458) = 16.538$
LR χ^2 $(H_0 : \beta_1 = 0)$ sex|age	$: -2(-26.511 - -19.458) = 14.106$
LR χ^2 $(H_0 : \beta_2 = 0)$ age|sex	$: -2(-24.435 - -19.458) = 9.954.$

The 14.1 should be compared with the Wald statistic of 8.47, and 9.954 should be compared with 6.58. The fitted logistic model is plotted separately

for females and males in Figure 10.3. The fitted model is

$$\text{logit}\{\text{Response} = 1|\text{sex,age}\} = -9.84 + 3.49 \times \text{sex} + .158 \times \text{age}, \qquad (10.21)$$

where as before sex = 0 for females, 1 for males. For example, for a 40-year-old female, the predicted logit is $-9.84 + .158(40) = -3.52$. The predicted probability of a response is $1/[1 + \exp(3.52)] = .029$. For a 40-year-old male, the predicted logit is $-9.84 + 3.49 + .158(40) = -.03$, with a probability of .492.

10.1.4 Design Formulations

The logistic multiple regression model can incorporate the same designs as can ordinary linear regression. An analysis of variance (ANOVA) model for a treatment with k levels can be formulated with $k - 1$ dummy variables. This logistic model is equivalent to a $2 \times k$ contingency table. An analysis of covariance logistic model is simply an ANOVA model augmented with covariables used for adjustment.

One unique design that is interesting to consider in the context of logistic models is a simultaneous comparison of multiple factors between two groups. Suppose, for example, that in a randomized trial with two treatments one wished to test whether any of 10 baseline characteristics are mal-distributed between the two groups. If the 10 factors are continuous, one could perform a two-sample Wilcoxon–Mann–Whitney test or a t-test for each factor (if each is normally distributed). However, this procedure would result in multiple comparison problems and would also not be able to detect the combined effect of small differences across all the factors. A better procedure would be a multivariate test. The Hotelling T^2 test is designed for just this situation. It is a k-variable extension of the one-variable unpaired t-test. The T^2 test, like discriminant analysis, assumes multivariate normality of the k factors. This assumption is especially tenuous when some of the factors are polytomous. A better alternative is the global test of no regression from the logistic model. This test is valid because it can be shown that H_0 : mean X is the same for both groups (= H_0 : mean X does not depend on group = H_0 : mean $X|$ group = constant) is true if and only if H_0 : Prob$\{$group$|X\}$ = constant. Thus k factors can be tested simultaneously for differences between the two groups using the binary logistic model, which has far fewer assumptions than does the Hotelling T^2 test. The logistic global test of no regression (with k d.f.) would be expected to have greater power if there is non-normality. Since the logistic model makes no assumption regarding the distribution of the descriptor variables, it can easily test for simultaneous group differences involving a mixture of continuous, binary, and nominal variables. In observational studies, such models for treatment received or exposure (propensity score models) hold great promise for adjusting for confounding.[117, 380, 526, 530, 531] 5

O'Brien[479] has developed a general test for comparing group 1 with group 2 for a single measurement. His test detects location and scale differences by fitting a logistic model for Prob{Group 2} using X and X^2 as predictors.

For a randomized study where adjustment for confounding is seldom necessary, adjusting for covariables using a binary logistic model results in *increases* in standard errors of regression coefficients.[527] This is the opposite of what happens in linear regression where there is an unknown variance parameter that is estimated using the residual squared error. Fortunately, adjusting for covariables using logistic regression, by accounting for subject heterogeneity, will result in larger regression coefficients even for a randomized treatment variable. The increase in estimated regression coefficients more than offsets the increase in standard error[284, 285, 527, 588].

10.2 Estimation

10.2.1 Maximum Likelihood Estimates

The parameters in the logistic regression model are estimated using the maximum likelihood (ML) method. The method is based on the same principles as the one-sample proportion example described in Section 9.1. The difference is that the general logistic model is not a single sample or a two-sample problem. The probability of response for the ith subject depends on a particular set of predictors X_i, and in fact the list of predictors may not be the same for any two subjects. Denoting the response and probability of response of the ith subject by Y_i and P_i, respectively, the model states that

$$P_i = \text{Prob}\{Y_i = 1|X_i\} = [1 + \exp(-X_i\beta)]^{-1}. \tag{10.22}$$

The likelihood of an observed response Y_i given predictors X_i and the unknown parameters β is

$$P_i^{Y_i}[1 - P_i]^{1-Y_i}. \tag{10.23}$$

The joint likelihood of all responses $Y_1, Y_2, \ldots, Y_n$ is the product of these likelihoods for $i = 1, \ldots, n$. The likelihood and log likelihood functions are rewritten by using the definition of P_i above to allow them to be recognized as a function of the unknown parameters β. Except in simple special cases (such as the k-sample problem in which all Xs are dummy variables), the ML estimates (MLE) of β cannot be written explicitly. The Newton–Raphson method described in Section 9.4 is usually used to solve iteratively for the list of values β that maximize the log likelihood. The MLEs are denoted by

$\hat{\beta}$. The inverse of the estimated observed information matrix is taken as the estimate of the variance–covariance matrix of $\hat{\beta}$.

Under $H_0 : \beta_1 = \beta_2 = \ldots = \beta_k = 0$, the intercept parameter β_0 can be estimated explicitly and the log likelihood under this global null hypothesis can be computed explicitly. Under the global null hypothesis, $P_i = P = [1 + \exp(-\beta_0)]^{-1}$ and the MLE of P is $\hat{P} = s/n$ where s is the number of responses and n is the sample size. The MLE of β_0 is $\hat{\beta}_0 = \text{logit}(\hat{P})$. The log likelihood under this null hypothesis is 6

$$\begin{aligned} & s \ \log(\hat{P}) + (n-s)\log(1-\hat{P}) \\ = & \ \ s \ \log(s/n) + (n-s)\log[(n-s)/n] \\ = & \ s \ \log s + (n-s)\log(n-s) - n\log(n). \end{aligned} \qquad (10.24)$$

10.2.2 Estimation of Odds Ratios and Probabilities

Once β is estimated, one can estimate any log odds, odds, or odds ratios. The MLE of the $X_j + 1 : X_j$ log odds ratio is $\hat{\beta}_j$, and the estimate of the $X_j + d : X_j$ log odds ratio is $\hat{\beta}_j d$, all other predictors remaining constant (assuming the absence of interactions and nonlinearities involving X_j). For large enough samples, the MLEs are normally distributed with variances that are consistently estimated from the estimated variance–covariance matrix. Letting z denote the $1-\alpha/2$ critical value of the standard normal distribution, a two-sided $1 - \alpha$ confidence interval for the log odds ratio for a one-unit increase in X_j is $[\hat{\beta}_j - zs, \hat{\beta}_j + zs]$, where s is the estimated standard error of $\hat{\beta}_j$. (Note that for $\alpha = .05$, i.e., for a 95% confidence interval, $z = 1.96$.)

A theorem in statistics states that the MLE of a function of a parameter is that same function of the MLE of the parameter. Thus the MLE of the $X_j + 1 : X_j$ odds ratio is $\exp(\hat{\beta}_j)$. Also, if a $1 - \alpha$ confidence interval of a parameter β is $[c, d]$ and $f(u)$ is a one-to-one function, a $1 - \alpha$ confidence interval of $f(\beta)$ is $[f(c), f(d)]$. Thus a $1-\alpha$ confidence interval for the $X_j + 1 : X_j$ odds ratio is $\exp[\hat{\beta}_j \pm zs]$. Note that while the confidence interval for β_j is symmetric about $\hat{\beta}_j$, the confidence interval for $\exp(\beta_j)$ is not. By the same theorem just used, the MLE of $P_i = \text{Prob}\{Y_i = 1|X_i\}$ is

$$\hat{P}_i = [1 + \exp(-X_i\hat{\beta})]^{-1}. \qquad (10.25)$$

A confidence interval for P_i could be derived by computing the standard error of $\hat{P}_i$, yielding a symmetric confidence interval. However, such an interval would have the disadvantage that its endpoints could fall below zero or exceed one. A better approach uses the fact that for large samples $X\hat{\beta}$ is approximately normally distributed. An estimate of the variance of $X\hat{\beta}$ in matrix notation is XVX' where V is the estimated variance–covariance matrix of $\hat{\beta}$ (see Equation 9.51). This variance is the sum of all variances and covariances of $\hat{\beta}$ weighted by squares and products of the predictors. The estimated standard error of $X\hat{\beta}$, s, is the square root of this variance estimate. A $1-\alpha$ confidence interval for P_i is then 7

$$\{1+\exp[-(X_i\hat{\beta} \pm zs)]\}^{-1}. \tag{10.26}$$

10.2.3 Minimum Sample Size Requirement

Suppose there were no covariates, so that the only parameter in the model is the intercept. What is the sample size required to allow the estimate of the intercept to be precise enough so that the predicted probability is within 0.1 of the true probability with 0.95 confidence, when the true intercept is in the neighborhood of zero? The answer is n=96. What if there were one covariate, and it was binary with a prevalence of $\frac{1}{2}$? One would need 96 subjects with $X = 0$ and 96 with $X = 1$ to have an upper bound on the margin of error for estimating Prob$\{Y = 1|X = x\}$ not exceed 0.1 for either value of x[a].

Now consider a very simple single continuous predictor case in which X has a normal distribution with mean zero and standard deviation σ, with the true Prob$\{Y = 1|X = x\} = [1+\exp(-x)]^{-1}$. The expected number of events is $\frac{n}{2}$[b]. The following simulation answers the question "What should n be so that the expected maximum absolute error (over $x \in [-1.5, 1.5]$) in $\hat{P}$ is less than ϵ?"

```
sigmas   ← c(.5, .75, 1, 1.25, 1.5, 1.75, 2, 2.5, 3, 4)
ns       ← seq(25, 300, by=25)
nsim     ← 1000
xs       ← seq(-1.5, 1.5, length=200)
pactual  ← plogis(xs)

dn ← list(sigma=format(sigmas), n=format(ns))
maxerr ← N1 ← array(NA, c(length(sigmas), length(ns)), dn)
require(rms)

i ← 0
for(s in sigmas) {
  i ← i + 1
  j ← 0
  for(n in ns) {
```

[a] The general formula for the sample size required to achieve a margin of error of δ in estimating a true probability of θ at the 0.95 confidence level is $n = (\frac{1.96}{\delta})^2 \times \theta(1-\theta)$. Set $\theta = \frac{1}{2}$ (intercept=0) for the worst case.

[b] The R code can easily be modified for other event frequencies, or the minimum of the number of events and non-events for a dataset at hand can be compared with $\frac{n}{2}$ in this simulation. An average maximum absolute error of 0.05 corresponds roughly to a half-width of the 0.95 confidence interval of 0.1.

```
    j ← j + 1
    n1 ← maxe ← 0
    for(k in 1:nsim) {
      x ← rnorm(n, 0, s)
      P ← plogis(x)
      y ← ifelse(runif(n) ≤ P, 1, 0)
      n1 ← n1 + sum(y)
      beta ← lrm.fit(x, y)$coefficients
      phat ← plogis(beta[1] + beta[2] * xs)
      maxe ← maxe + max(abs(phat - pactual))
    }
    n1 ← n1/nsim
    maxe ← maxe/nsim
    maxerr[i,j] ← maxe
    N1[i,j] ← n1
  }
}
xrange ← range(xs)
simerr ← llist(N1, maxerr, sigmas, ns, nsim, xrange)

maxe ← reShape(maxerr)
# Figure 10.4
xYplot(maxerr ∼ n, groups=sigma, data=maxe,
       ylab=expression(paste('Average Maximum  ',
           abs(hat(P) - P))),
       type='l', lty=rep(1:2, 5), label.curve=FALSE,
       abline=list(h=c(.15, .1, .05), col=gray(.85)))
Key(.8, .68, other=list(cex=.7,
               title=expression(∼∼∼∼∼∼∼∼∼∼∼sigma)))
```

10.3 Test Statistics

The likelihood ratio, score, and Wald statistics discussed earlier can be used to test any hypothesis in the logistic model. The likelihood ratio test is generally preferred. When true parameters are near the null values all three statistics usually agree. The Wald test has a significant drawback when the true parameter value is very far from the null value. In such case the standard error estimate becomes too large. As $\hat{\beta}_j$ increases from 0, the Wald test statistic for $H_0 : \beta_j = 0$ becomes larger, but after a certain point it becomes smaller. The statistic will eventually drop to zero if $\hat{\beta}_j$ becomes infinite.[278] Infinite estimates can occur in the logistic model especially when there is a binary predictor whose mean is near 0 or 1. Wald statistics are especially problematic in this case. For example, if 10 out of 20 males had a disease and 5 out of 5 females had the disease, the female : male odds ratio is infinite and so is the logistic regression coefficient for sex. If such a situation occurs, the likelihood ratio or score statistic should be used instead of the Wald statistic.

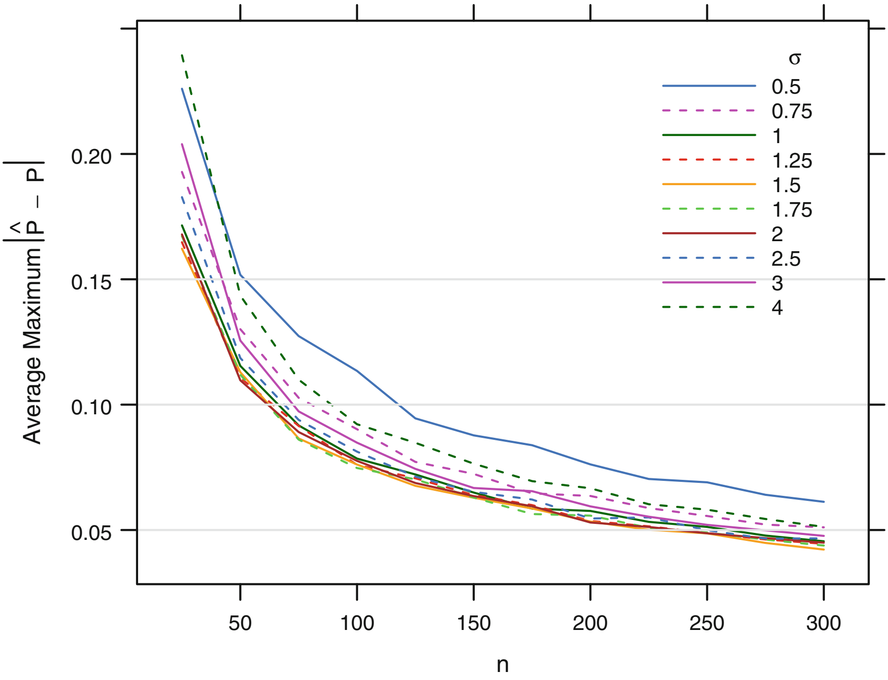

Fig. 10.4 Simulated expected maximum error in estimating probabilities for $x \in [-1.5, 1.5]$ with a single normally distributed X with mean zero

For k-sample (ANOVA-type) logistic models, logistic model statistics are equivalent to contingency table χ^2 statistics. As exemplified in the logistic model relating sex to response described previously, the global likelihood ratio statistic for all dummy variables in a k-sample model is identical to the contingency table (k-sample binomial) likelihood ratio χ^2 statistic. The score statistic for this same situation turns out to be identical to the $k-1$ degrees of freedom Pearson χ^2 for a $k \times 2$ table.

As mentioned in Section 2.6, it can be dangerous to interpret individual parameters, make pairwise treatment comparisons, or test linearity if the overall test of association for a factor represented by multiple parameters is insignificant.

10.4 Residuals

Several types of residuals can be computed for binary logistic model fits. Many of these residuals are used to examine the influence of individual observations on the fit. The *partial residual* can be used for directly assessing how each 8 predictor should be transformed. For the ith observation, the partial residual for the jth element of X is defined by

$$r_{ij} = \hat{\beta}_j X_{ij} + \frac{Y_i - \hat{P}_i}{\hat{P}_i(1 - \hat{P}_i)}, \tag{10.27}$$

where X_{ij} is the value of the jth variable in the ith observation, Y_i is the corresponding value of the response, and $\hat{P}_i$ is the predicted probability that $Y_i = 1$. A smooth plot (using, e.g., loess) of X_{ij} against r_{ij} will provide an estimate of how X_j should be transformed, adjusting for the other Xs (using their current transformations). Typically one tentatively models X_j linearly and checks the smoothed plot for linearity. A U-shaped relationship in this plot, for example, indicates that a squared term or spline function needs to be added for X_j. This approach does assume additivity of predictors. 9

10.5 Assessment of Model Fit

As the logistic regression model makes no distributional assumptions, only the assumptions of linearity and additivity need to be verified (in addition to the usual assumptions about independence of observations and inclusion of important covariables). In ordinary linear regression there is no global test for lack of model fit unless there are replicate observations at various settings of X. This is because ordinary regression entails estimation of a separate variance parameter σ^2. In logistic regression there are global tests for goodness of fit. Unfortunately, some of the most frequently used ones are inappropriate. For example, it is common to see a deviance test of goodness of fit based on the "residual" log likelihood, with P-values obtained from a χ^2 distribution with $n - p$ d.f. This P-value is inappropriate since the deviance does not have an asymptotic χ^2 distribution, due to the facts that the number of parameters estimated is increasing at the same rate as n and the expected cell frequencies are far below five (by definition).

Hosmer and Lemeshow[304] have developed a commonly used test for goodness of fit for binary logistic models based on grouping into deciles of predicted probability and performing an ordinary χ^2 test for the mean predicted probability against the observed fraction of events (using 8 d.f. to account for evaluating fit on the model development sample). The Hosmer–Lemeshow test is dependent on the choice of how predictions are grouped[303] and it is not clear that the choice of the number of groups should be independent of n. Hosmer et al.[303] have compared a number of global goodness of fit tests for binary logistic regression. They concluded that the simple unweighted sum of squares test of Copas[124] as modified by le Cessie and van Houwelingen[387] is as good as any. They used a normal Z-test for the sum of squared errors ($n \times B$, where B is the Brier index in Equation 10.35). This test takes into account the fact that one cannot obtain a χ^2 distribution for the sum of squares. It also takes into account the estimation of β. It is not yet clear for which types of lack of fit this test has reasonable power. Returning to the external validation case where uncertainty of β does not need to be accounted for, Stallard[584] has further documented the lack of power of the original Hosmer-Lemeshow test and found more power with a logarithmic scoring rule (deviance test) and a χ^2 test that, unlike the simple unweighted sum of squares test, weights each squared error by dividing it by $\hat{P}_i(1-\hat{P}_i)$. A scaled χ^2 distribution seemed to provide the best approximation to the null distribution of the test statistics.

More power for detecting lack of fit is expected to be obtained from testing specific alternatives to the model. In the model

$$\text{logit}\{Y = 1|X\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2, \tag{10.28}$$

where X_1 is binary and X_2 is continuous, one needs to verify that the log odds is related to X_1 and X_2 according to Figure 10.5.

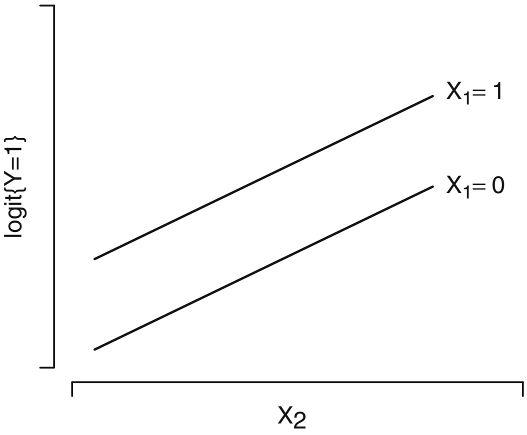

Fig. 10.5 Logistic regression assumptions for one binary and one continuous predictor

The simplest method for validating that the data are consistent with the no-interaction linear model involves stratifying the sample by X_1 and quantile groups (e.g., deciles) of X_2.[265] Within each stratum the proportion of responses $\hat{P}$ is computed and the log odds calculated from $\log[\hat{P}/(1 - \hat{P})]$. The number of quantile groups should be such that there are at least 20 (and perhaps many more) subjects in each $X_1 \times X_2$ group. Otherwise, probabilities cannot be estimated precisely enough to allow trends to be seen above "noise" in the data. Since at least 3 X_2 groups must be formed to allow assessment of linearity, the total sample size must be at least $2 \times 3 \times 20 = 120$ for this method to work at all.

Figure 10.6 demonstrates this method for a large sample size of 3504 subjects stratified by sex and deciles of age. Linearity is apparent for males while there is evidence for slight interaction between age and sex since the age trend for females appears curved.

```
getHdata(acath)
acath$sex ← factor(acath$sex, 0:1, c('male','female'))
dd ← datadist(acath); options(datadist='dd')
f ← lrm(sigdz ∼ rcs(age, 4) * sex, data=acath)
```

```
w ← function(...)
  with(acath, {
    plsmo(age, sigdz, group=sex, fun=qlogis, lty='dotted',
          add=TRUE, grid=TRUE)
    af ← cut2(age, g=10, levels.mean=TRUE)
    prop ← qlogis(tapply(sigdz, list(af, sex), mean,
                        na.rm=TRUE))
    agem ← as.numeric(row.names(prop))
    lpoints(agem, prop[,'female'], pch=4, col='green')
    lpoints(agem, prop[,'male'],   pch=2, col='green')
  } )   # Figure 10.6
plot(Predict(f, age, sex), ylim=c(-2,4), addpanel=w,
     label.curve=list(offset=unit(0.5, 'cm')))
```

The subgrouping method requires relatively large sample sizes and does not use continuous factors effectively. The ordering of values is not used at all between intervals, and the estimate of the relationship for a continuous variable has little resolution. Also, the method of grouping chosen (e.g., deciles vs. quintiles vs. rounding) can alter the shape of the plot.

In this dataset with only two variables, it is efficient to use a nonparametric smoother for age, separately for males and females. Nonparametric smoothers, such as `loess`[111] used here, work well for binary response variables (see Section 2.4.7); the logit transformation is made on the smoothed probability estimates. The smoothed estimates are shown in Figure 10.6. 10

When there are several predictors, the restricted cubic spline function is better for estimating the true relationship between X_2 and logit$\{Y = 1\}$ 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 transformation of X_2 as discussed in Section 2.4:

$$\begin{aligned} \text{logit}\{Y = 1|X\} &= \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \hat{\beta}_3 X_2' + \hat{\beta}_4 X_2'' \\ &= \hat{\beta}_0 + \hat{\beta}_1 X_1 + f(X_2), \end{aligned} \tag{10.29}$$

where X_2' and X_2'' are constructed spline variables (when $k = 4$). Plotting the estimated spline function $f(X_2)$ versus X_2 will estimate 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 $H_0 : \beta_3 = \beta_4 = 0$.

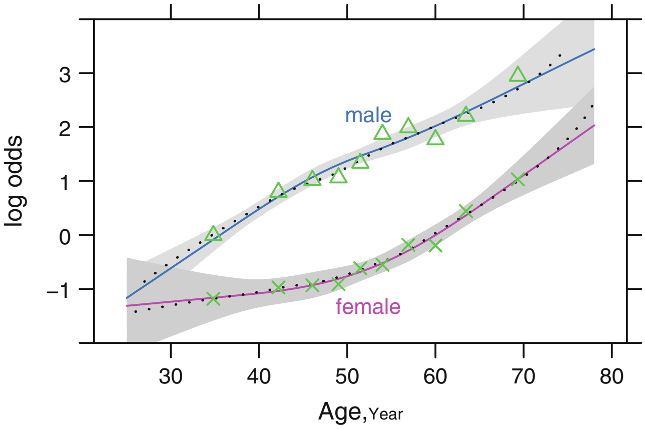

Fig. 10.6 Logit proportions of significant coronary artery disease by sex and deciles of age for n=3504 patients, with spline fits (smooth curves). Spline fits are for $k = 4$ knots at age= 36, 48, 56, and 68 years, and interaction between age and sex is allowed. Shaded bands are pointwise 0.95 confidence limits for predicted log odds. Smooth nonparametric estimates are shown as dotted curves. Data courtesy of the Duke Cardiovascular Disease Databank.

For testing interaction between X_1 and X_2, 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

$$\begin{aligned}\text{logit}\{Y = 1|X\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_2' + \beta_4 X_2'' \\ + \beta_5 X_1 X_2 + \beta_6 X_1 X_2' + \beta_7 X_1 X_2''\end{aligned} \quad (10.30)$$

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. The most computationally efficient test for lack of fit is the score test (e.g., X_1 and X_2 are forced into a tentative model and the remaining variables are candidates). Figure 10.6 also depicts a fitted spline logistic model with $k = 4$, allowing for general interaction between age and sex as parameterized above. The fitted function, after expanding the restricted cubic spline function for simplicity (see Equation 2.27), is given above. Note the good agreement between the empirical estimates of log odds and the spline fits and nonparametric estimates in this large dataset.

An analysis of log likelihood for this model and various sub-models is found in Table 10.3. The χ^2 for global tests is corrected for the intercept and the degrees of freedom does not include the intercept.

Table 10.3 LR χ^2 tests for coronary artery disease risk

Model / Hypothesis	Likelihood Ratio χ^2	d.f.	P	Formula
a: sex, age (linear, no interaction)	766.0	2		
b: sex, age, age × sex	768.2	3		
c: sex, spline in age	769.4	4		
d: sex, spline in age, interaction	782.5	7		
H_0 : no age × sex interaction given linearity	2.2	1	.14	$(b - a)$
H_0 : age linear | no interaction	3.4	2	.18	$(c - a)$
H_0 : age linear, no interaction	16.6	5	.005	$(d - a)$
H_0 : age linear, product form interaction	14.4	4	.006	$(d - b)$
H_0 : no interaction, allowing for nonlinearity in age	13.1	3	.004	$(d - c)$

Table 10.4 AIC on χ^2 scale by number of knots

k	Model χ^2	AIC
0	99.23	97.23
3	112.69	108.69
4	121.30	115.30
5	123.51	115.51
6	124.41	114.51

This analysis confirms the first impression from the graph, namely, that age × sex interaction is present but it is not of the form of a simple product between age and sex (change in slope). In the context of a linear age effect, there is no significant product interaction effect ($P = .14$). Without allowing for interaction, there is no significant nonlinear effect of age ($P = .18$). However, the general test of lack of fit with 5 d.f. indicates a significant departure from the linear additive model ($P = .005$).

In Figure 10.7, data from 2332 patients who underwent cardiac catheterization at Duke University Medical Center and were found to have significant ($\geq 75\%$) diameter narrowing of at least one major coronary artery were analyzed (the dataset is available from the Web site). The relationship between the time from the onset of symptoms of coronary artery disease (e.g., angina, myocardial infarction) to the probability that the patient has severe (three-vessel disease or left main disease—`tvdlm`) coronary disease was of interest. There were 1129 patients with `tvdlm`. A logistic model was used with the duration of symptoms appearing as a restricted cubic spline function with $k = 3, 4, 5$, and 6 equally spaced knots in terms of quantiles between .05 and .95. The best fit for the number of parameters was chosen using Akaike's information criterion (AIC), computed in Table 10.4 as the model likelihood ratio χ^2 minus twice the number of parameters in the model aside from the intercept. The linear model is denoted $k = 0$.

```
dz ← subset(acath, sigdz==1)
dd ← datadist(dz)
```

```
f ← lrm(tvdlm ∼ rcs(cad.dur, 5), data=dz)
w ← function(...)
  with(dz, {
    plsmo(cad.dur, tvdlm, fun=qlogis, add=TRUE,
          grid=TRUE, lty='dotted')
    x ← cut2(cad.dur, g=15, levels.mean=TRUE)
    prop ← qlogis(tapply(tvdlm, x, mean, na.rm=TRUE))
    xm ← as.numeric(names(prop))
    lpoints(xm, prop, pch=2, col='green')
  } )    # Figure 10.7
plot(Predict(f, cad.dur), addpanel=w)
```

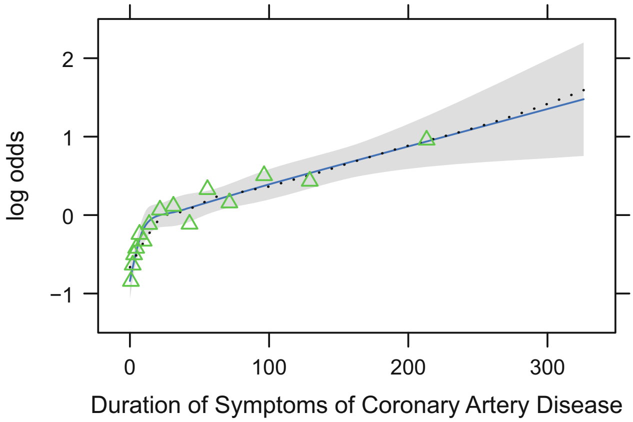

Fig. 10.7 Estimated relationship between duration of symptoms and the log odds of severe coronary artery disease for $k = 5$. Knots are marked with arrows. Solid line is spline fit; dotted line is a nonparametric loess estimate.

Figure 10.7 displays the spline fit for $k = 5$. The triangles represent subgroup estimates obtained by dividing the sample into groups of 150 patients. For example, the leftmost triangle represents the logit of the proportion of `tvdlm` in the 150 patients with the shortest duration of symptoms, versus the mean duration in that group. A Wald test of linearity, with 3 d.f., showed highly significant nonlinearity (χ^2= 23.92 with 3 d.f.). The plot of the spline transformation suggests a log transformation, and when log (duration of symptoms in months + 1) was fitted in a logistic model, the log likelihood of the model (119.33 with 1 d.f.) was virtually as good as the spline model (123.51 with 4 d.f.); the corresponding Akaike information criteria (on the χ^2 scale) are 117.33 and 115.51. To check for adequacy in the log transformation, a five-knot restricted cubic spline function was fitted to $\log_{10}(\text{months} + 1)$, as displayed in Figure 10.8. There is some evidence for lack of fit on the right, but the Wald χ^2 for testing linearity yields $P = .27$.

```
f ← lrm(tvdlm ~ log10(cad.dur + 1), data=dz)
w ← function(...)
  with(dz, {
    x ← cut2(cad.dur, m=150, levels.mean=TRUE)
    prop ← tapply(tvdlm, x, mean, na.rm=TRUE)
    xm ← as.numeric(names(prop))
    lpoints(xm, prop, pch=2, col='green')
  } )
#   Figure 10.8
plot(Predict(f, cad.dur, fun=plogis), ylab='P',
     ylim=c(.2, .8), addpanel=w)
```

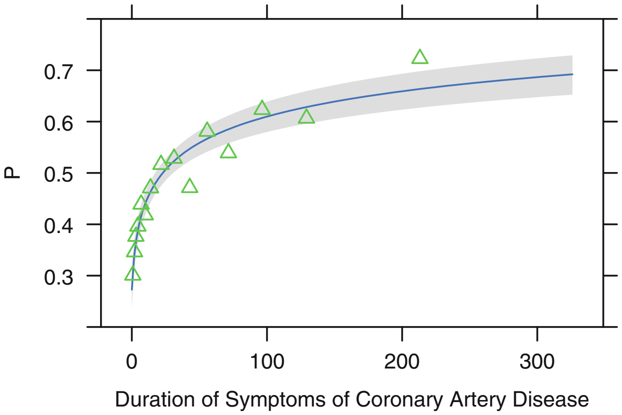

Fig. 10.8 Fitted linear logistic model in $\log_{10}(\text{duration} + 1)$, with subgroup estimates using groups of 150 patients. Fitted equation is $\text{logit}(\texttt{tvdlm}) = -.9809 + .7122 \log_{10}(\text{months} + 1)$.

If the model contains two continuous predictors, they may both be expanded with spline functions in order to test linearity or to describe nonlinear relationships. Testing interaction is more difficult here. If X_1 is continuous, one might temporarily group X_1 into quantile groups. Consider the subset of 2258 (1490 with disease) of the 3504 patients used in Figure 10.6 who have serum cholesterol measured. A logistic model for predicting significant coronary disease was fitted with age in tertiles (modeled with two dummy variables), sex, age × sex interaction, four-knot restricted cubic spline in cholesterol, and age tertile × cholesterol interaction. Except for the sex adjustment this model is equivalent to fitting three separate spline functions in cholesterol, one for each age tertile. The fitted model is shown in Figure 10.9 for cholesterol and age tertile against logit of significant disease. Significant age × cholesterol interaction is apparent from the figure and is suggested by the Wald χ^2 statistic (10.03) that follows. Note that the test for linearity of the interaction with respect to cholesterol is very insignificant ($\chi^2 = 2.40$ on 4 d.f.), but we retain it for now. The fitted function is

```
acath ← transform(acath,
                  cholesterol = choleste,
                  age.tertile = cut2(age,g=3),
                  sx = as.integer(acath$sex) - 1)
# sx for loess, need to code as numeric
dd ← datadist(acath); options(datadist='dd')

# First model stratifies age into tertiles to get more
# empirical estimates of age x cholesterol interaction

f ← lrm(sigdz ~ age.tertile*(sex + rcs(cholesterol,4)),
        data=acath)
print(f, latex=TRUE)
```

Logistic Regression Model

```
lrm(formula = sigdz ~ age.tertile * (sex + rcs(cholesterol, 4)),
    data = acath)
```

Frequencies of Missing Values Due to Each Variable

```
sigdz age.tertile         sex cholesterol
    0           0           0        1246
```

		Model Likelihood Ratio Test		Discrimination Indexes		Rank Discrim. Indexes	
Obs	2258	LR χ^2	533.52	R^2	0.291	C	0.780
0	768	d.f.	14	g	1.316	D_{xy}	0.560
1	1490	Pr$(>\chi^2)$	< 0.0001	g_r	3.729	γ	0.562
max $\left\|\frac{\partial \log L}{\partial \beta}\right\|$	2×10^{-8}			g_p	0.252	τ_a	0.251
				Brier	0.173		

	Coef	S.E.	Wald Z	Pr$(>\|Z\|)$
Intercept	-0.4155	1.0987	-0.38	0.7053
age.tertile=[49,58)	0.8781	1.7337	0.51	0.6125
age.tertile=[58,82]	4.7861	1.8143	2.64	0.0083
sex=female	-1.6123	0.1751	-9.21	< 0.0001
cholesterol	0.0029	0.0060	0.48	0.6347
cholesterol'	0.0384	0.0242	1.59	0.1126
cholesterol"	-0.1148	0.0768	-1.49	0.1350
age.tertile=[49,58) * sex=female	-0.7900	0.2537	-3.11	0.0018
age.tertile=[58,82] * sex=female	-0.4530	0.2978	-1.52	0.1283
age.tertile=[49,58) * cholesterol	0.0011	0.0095	0.11	0.9093

	Coef	S.E.	Wald Z	Pr$(>\lvert Z\rvert)$
age.tertile=[58,82] * cholesterol	-0.0158	0.0099	-1.59	0.1111
age.tertile=[49,58) * cholesterol'	-0.0183	0.0365	-0.50	0.6162
age.tertile=[58,82] * cholesterol'	0.0127	0.0406	0.31	0.7550
age.tertile=[49,58) * cholesterol''	0.0582	0.1140	0.51	0.6095
age.tertile=[58,82] * cholesterol''	-0.0092	0.1301	-0.07	0.9436

```
ltx(f)
```

$$X\hat{\beta} = -0.415 + 0.878[\text{age.tertile} \in [49,58)] + 4.79[\text{age.tertile} \in [58,82]] - 1.61[\text{female}] + 0.00287\text{cholesterol} + 1.52\times10^{-6}(\text{cholesterol} - 160)_+^3 - 4.53\times10^{-6}(\text{cholesterol} - 208)_+^3 + 3.44\times10^{-6}(\text{cholesterol} - 243)_+^3 - 4.28\times10^{-7}(\text{cholesterol} - 319)_+^3 + [\text{female}][-0.79[\text{age.tertile} \in [49,58)] - 0.453[\text{age.tertile} \in [58,82]]] + [\text{age.tertile} \in [49,58)][0.00108\text{cholesterol} - 7.23\times10^{-7}(\text{cholesterol} - 160)_+^3 + 2.3\times10^{-6}(\text{cholesterol} - 208)_+^3 - 1.84\times10^{-6}(\text{cholesterol} - 243)_+^3 + 2.69\times10^{-7}(\text{cholesterol} - 319)_+^3] + [\text{age.tertile} \in [58,82]][-0.0158\text{cholesterol} + 5\times10^{-7}(\text{cholesterol} - 160)_+^3 - 3.64\times10^{-7}(\text{cholesterol} - 208)_+^3 - 5.15\times10^{-7}(\text{cholesterol} - 243)_+^3 + 3.78\times10^{-7}(\text{cholesterol} - 319)_+^3].$$

```
# Table 10.5:
latex(anova(f), file='', size='smaller',
      caption='Crudely categorizing age into tertiles',
      label='tab:anova-tertiles')
```

```
yl ← c(-1,5)
plot(Predict(f, cholesterol, age.tertile),
     adj.subtitle=FALSE, ylim=yl)    # Figure 10.9
```

Table 10.5 Crudely categorizing age into tertiles

	χ^2	d.f.	P
age.tertile (Factor+Higher Order Factors)	120.74	10	< 0.0001
All Interactions	21.87	8	0.0052
sex (Factor+Higher Order Factors)	329.54	3	< 0.0001
All Interactions	9.78	2	0.0075
cholesterol (Factor+Higher Order Factors)	93.75	9	< 0.0001
All Interactions	10.03	6	0.1235
Nonlinear (Factor+Higher Order Factors)	9.96	6	0.1263
age.tertile × sex (Factor+Higher Order Factors)	9.78	2	0.0075
age.tertile × cholesterol (Factor+Higher Order Factors)	10.03	6	0.1235
Nonlinear	2.62	4	0.6237
Nonlinear Interaction : f(A,B) vs. AB	2.62	4	0.6237
TOTAL NONLINEAR	9.96	6	0.1263
TOTAL INTERACTION	21.87	8	0.0052
TOTAL NONLINEAR + INTERACTION	29.67	10	0.0010
TOTAL	410.75	14	< 0.0001

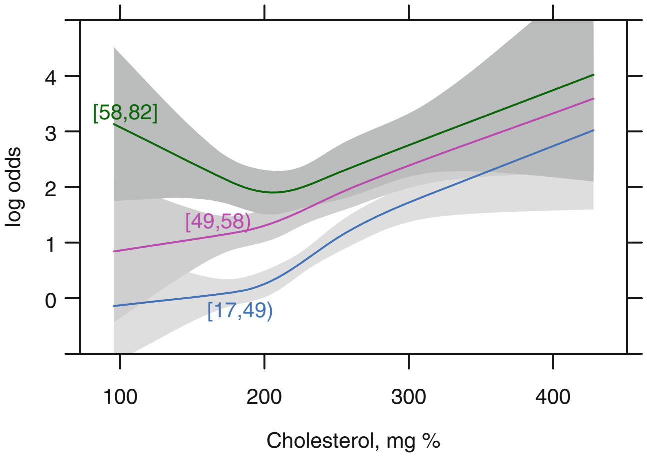

Fig. 10.9 Log odds of significant coronary artery disease modeling age with two dummy variables

Before fitting a parametric model that allows interaction between age and cholesterol, let us use the local regression model of Cleveland et al.[96] discussed in Section 2.4.7. This nonparametric smoothing method is not meant to handle binary Y, but it can still provide useful graphical displays in the binary case. Figure 10.10 depicts the fit from a local regression model predicting $Y = 1 =$ significant coronary artery disease. Predictors are sex (modeled parametrically with a dummy variable), age, and cholesterol, the last two fitted nonparametrically. The effect of not explicitly modeling a probability is seen in the figure, as the predicted probabilities exceeded 1. Because of this we do not take the logit transformation but leave the predicted values in raw form. However, the overall shape is in agreement with Figure 10.10.

```
# Re-do model with continuous age
f ← loess(sigdz ∼ age * (sx + cholesterol), data=acath,
           parametric="sx", drop.square="sx")
ages   ← seq(25,   75, length=40)
chols  ← seq(100, 400, length=40)
g ← expand.grid(cholesterol=chols, age=ages, sx=0)
# drop sex dimension of grid since held to 1 value
p ← drop(predict(f, g))
p[p < 0.001] ← 0.001
p[p > 0.999] ← 0.999
zl ← c(-3, 6)    # Figure 10.10
wireframe(qlogis(p) ∼ cholesterol*age,
          xlab=list(rot=30), ylab=list(rot=-40),
          zlab=list(label='log odds', rot=90), zlim=zl,
          scales = list(arrows = FALSE), data=g)
```

Chapter 2 discussed linear splines, which can be used to construct linear spline surfaces by adding all cross-products of the linear variables and spline terms in the model. With a sufficient number of knots for each predictor, the linear spline surface can fit a wide variety of patterns. However, it requires a large number of parameters to be estimated. For the age–sex–cholesterol example, a linear spline surface is fitted for age and cholesterol, and a sex × age spline interaction is also allowed. Figure 10.11 shows a fit that placed knots at quartiles of the two continuous variables[c]. The algebraic form of the fitted model is shown below.

```
f ← lrm(sigdz ~ lsp(age,c(46,52,59)) *
          (sex + lsp(cholesterol,c(196,224,259))),
          data=acath)
ltx(f)
```

$$X\hat{\beta} = -1.83 + 0.0232\,\text{age} + 0.0759(\text{age}-46)_+ - 0.0025(\text{age}-52)_+ + 2.27(\text{age}-59)_+ + 3.02[\text{female}] - 0.0177\,\text{cholesterol} + 0.114(\text{cholesterol}-196)_+ - 0.131(\text{cholesterol}-224)_+ + 0.0651(\text{cholesterol}-259)_+ + [\text{female}][-0.112\,\text{age} + 0.0852\,(\text{age}-46)_+ - 0.0302\,(\text{age}-52)_+ + 0.176\,(\text{age}-59)_+] + \text{age}\,[0.000577\,\text{cholesterol} - 0.00286\,(\text{cholesterol}-196)_+ + 0.00382\,(\text{cholesterol}-224)_+ - 0.00205\,(\text{cholesterol}-259)_+] + (\text{age}-46)_+[-0.000936\,\text{cholesterol} + 0.00643(\text{cholesterol}-196)_+ - 0.0115(\text{cholesterol}-224)_+ + 0.00756(\text{cholesterol}-259)_+] + (\text{age}-52)_+[0.000433\,\text{cholesterol} - 0.0037\,(\text{cholesterol}-196)_+ + 0.00815\,(\text{cholesterol}-224)_+ - 0.00715\,(\text{cholesterol}-259)_+] + (\text{age}-59)_+[-0.0124\,\text{cholesterol} + 0.015\,(\text{cholesterol}-196)_+ - 0.0067\,(\text{cholesterol}-224)_+ + 0.00752\,(\text{cholesterol}-259)_+].$$

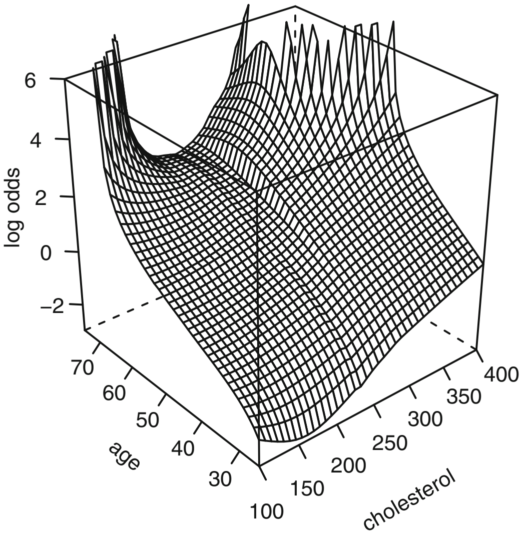

Fig. 10.10 Local regression fit for the logit of the probability of significant coronary disease vs. age and cholesterol for males, based on the `loess` function.

[c] In the wireframe plots that follow, predictions for cholesterol–age combinations for which fewer than 5 exterior points exist are not shown, so as to not extrapolate to regions not supported by at least five points beyond the data perimeter.

```
latex(anova(f), caption='Linear spline surface', file='',
      size='smaller', label='tab:anova-lsp')   # Table 10.6
```

```
perim ← with(acath,
             perimeter(cholesterol, age, xinc=20, n=5))
zl ← c(-2, 4)   # Figure 10.11
bplot(Predict(f, cholesterol, age, np=40), perim=perim,
      lfun=wireframe, zlim=zl, adj.subtitle=FALSE)
```

Table 10.6 Linear spline surface

	χ^2	d.f.	P
age (Factor+Higher Order Factors)	164.17	24	< 0.0001
All Interactions	42.28	20	0.0025
Nonlinear (Factor+Higher Order Factors)	25.21	18	0.1192
sex (Factor+Higher Order Factors)	343.80	5	< 0.0001
All Interactions	23.90	4	0.0001
cholesterol (Factor+Higher Order Factors)	100.13	20	< 0.0001
All Interactions	16.27	16	0.4341
Nonlinear (Factor+Higher Order Factors)	16.35	15	0.3595
age × sex (Factor+Higher Order Factors)	23.90	4	0.0001
Nonlinear	12.97	3	0.0047
Nonlinear Interaction : f(A,B) vs. AB	12.97	3	0.0047
age × cholesterol (Factor+Higher Order Factors)	16.27	16	0.4341
Nonlinear	11.45	15	0.7204
Nonlinear Interaction : f(A,B) vs. AB	11.45	15	0.7204
f(A,B) vs. Af(B) + Bg(A)	9.38	9	0.4033
Nonlinear Interaction in age vs. Af(B)	9.99	12	0.6167
Nonlinear Interaction in cholesterol vs. Bg(A)	10.75	12	0.5503
TOTAL NONLINEAR	33.22	24	0.0995
TOTAL INTERACTION	42.28	20	0.0025
TOTAL NONLINEAR + INTERACTION	49.03	26	0.0041
TOTAL	449.26	29	< 0.0001

Chapter 2 also discussed a tensor spline extension of the restricted cubic spline model to fit a smooth function of two predictors, $f(X_1, X_2)$. Since this function allows for general interaction between X_1 and X_2, the two-variable cubic spline is a powerful tool for displaying and testing interaction, assuming the sample size warrants estimating $2(k-1)+(k-1)^2$ parameters for a rectangular grid of $k \times k$ knots. Unlike the linear spline surface, the cubic surface is smooth. It also requires fewer parameters in most situations. The general cubic model with $k = 4$ (ignoring the sex effect here) is

$$\begin{aligned} & \beta_0 + \beta_1 X_1 + \beta_2 X_1' + \beta_3 X_1'' + \beta_4 X_2 + \beta_5 X_2' + \beta_6 X_2'' + \beta_7 X_1 X_2 \\ + & \quad \beta_8 X_1 X_2' + \beta_9 X_1 X_2'' + \beta_{10} X_1' X_2 + \beta_{11} X_1' X_2' \\ + & \quad + \beta_{12} X_1' X_2'' + \beta_{13} X_1'' X_2 + \beta_{14} X_1'' X_2' + \beta_{15} X_1'' X_2'', \end{aligned} \tag{10.31}$$

where X_1', X_1'', X_2', and X_2'' are restricted cubic spline component variables for X_1 and X_2 for $k = 4$. A general test of interaction with 9 d.f. is $H_0 : \beta_7 = \ldots = \beta_{15} = 0$. A test of adequacy of a simple product form interaction is $H_0 : \beta_8 = \ldots = \beta_{15} = 0$ with 8 d.f. A 13 d.f. test of linearity and additivity is $H_0 : \beta_2 = \beta_3 = \beta_5 = \beta_6 = \beta_7 = \beta_8 = \beta_9 = \beta_{10} = \beta_{11} = \beta_{12} = \beta_{13} = \beta_{14} = \beta_{15} = 0$.

Figure 10.12 depicts the fit of this model. There is excellent agreement with Figures 10.9 and 10.11, including an increased (but probably insignificant) risk with low cholesterol for age ≥ 57.

```
f ← lrm(sigdz ~ rcs(age,4)*(sex + rcs(cholesterol,4)),
          data=acath, tol=1e-11)
ltx(f)
```

$X\hat{\beta} = -6.41 + 0.166\text{age} - 0.00067(\text{age} - 36)_+^3 + 0.00543(\text{age} - 48)_+^3 - 0.00727(\text{age} - 56)_+^3 + 0.00251(\text{age} - 68)_+^3 + 2.87[\text{female}] + 0.00979\text{cholesterol} + 1.96 \times 10^{-6}(\text{cholesterol} - 160)_+^3 - 7.16 \times 10^{-6}(\text{cholesterol} - 208)_+^3 + 6.35 \times 10^{-6}(\text{cholesterol} - 243)_+^3 - 1.16 \times 10^{-6}(\text{cholesterol} - 319)_+^3 + [\text{female}][-0.109\text{age} + 7.52 \times 10^{-5}(\text{age} - 36)_+^3 + 0.00015(\text{age} - 48)_+^3 - 0.00045(\text{age} - 56)_+^3 + 0.000225(\text{age} - 68)_+^3] + \text{age}[-0.00028\text{cholesterol} + 2.68 \times 10^{-9}(\text{cholesterol} - 160)_+^3 + 3.03 \times 10^{-8}(\text{cholesterol} - 208)_+^3 - 4.99 \times 10^{-8}(\text{cholesterol} - 243)_+^3 + 1.69 \times 10^{-8}(\text{cholesterol} - 319)_+^3] + \text{age}'[0.00341\text{cholesterol} - 4.02 \times 10^{-7}(\text{cholesterol} - 160)_+^3 + 9.71 \times 10^{-7}(\text{cholesterol} - 208)_+^3 - 5.79 \times 10^{-7}(\text{cholesterol} - 243)_+^3 + 8.79 \times 10^{-9}(\text{cholesterol} - 319)_+^3] + \text{age}''[-0.029\text{cholesterol} + 3.04 \times 10^{-6}(\text{cholesterol} -$

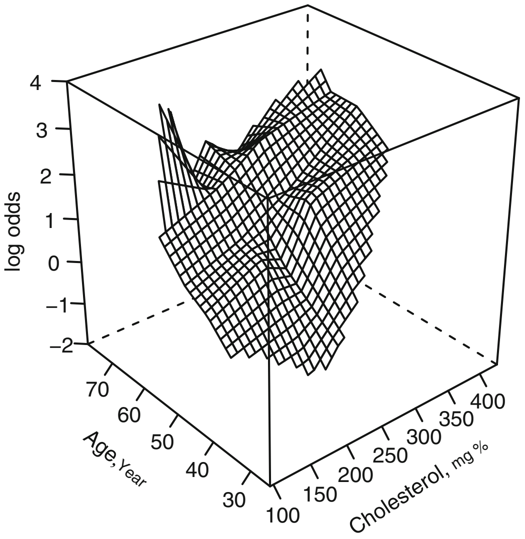

Fig. 10.11 Linear spline surface for males, with knots for age at 46, 52, 59 and knots for cholesterol at 196, 224, and 259 (quartiles).

$160)^3_+ - 7.34\times10^{-6}(\text{cholesterol} - 208)^3_+ + 4.36\times10^{-6}(\text{cholesterol} - 243)^3_+ - 5.82\times10^{-8}(\text{cholesterol} - 319)^3_+]$.

```
latex(anova(f), caption='Cubic spline surface', file='',
      size='smaller', label='tab:anova-rcs') #Table 10.7
```

```
# Figure 10.12:
bplot(Predict(f, cholesterol, age, np=40), perim=perim,
      lfun=wireframe, zlim=zl, adj.subtitle=FALSE)
```

Table 10.7 Cubic spline surface

	χ^2	d.f.	P
age (Factor+Higher Order Factors)	165.23	15	< 0.0001
All Interactions	37.32	12	0.0002
Nonlinear (Factor+Higher Order Factors)	21.01	10	0.0210
sex (Factor+Higher Order Factors)	343.67	4	< 0.0001
All Interactions	23.31	3	< 0.0001
cholesterol (Factor+Higher Order Factors)	97.50	12	< 0.0001
All Interactions	12.95	9	0.1649
Nonlinear (Factor+Higher Order Factors)	13.62	8	0.0923
age × sex (Factor+Higher Order Factors)	23.31	3	< 0.0001
Nonlinear	13.37	2	0.0013
Nonlinear Interaction : f(A,B) vs. AB	13.37	2	0.0013
age × cholesterol (Factor+Higher Order Factors)	12.95	9	0.1649
Nonlinear	7.27	8	0.5078
Nonlinear Interaction : f(A,B) vs. AB	7.27	8	0.5078
f(A,B) vs. Af(B) + Bg(A)	5.41	4	0.2480
Nonlinear Interaction in age vs. Af(B)	6.44	6	0.3753
Nonlinear Interaction in cholesterol vs. Bg(A)	6.27	6	0.3931
TOTAL NONLINEAR	29.22	14	0.0097
TOTAL INTERACTION	37.32	12	0.0002
TOTAL NONLINEAR + INTERACTION	45.41	16	0.0001
TOTAL	450.88	19	< 0.0001

Statistics for testing age × cholesterol components of this fit are above. None of the nonlinear interaction components is significant, but we again retain them.

The general interaction model can be restricted to be of the form

$$f(X_1, X_2) = f_1(X_1) + f_2(X_2) + X_1 g_2(X_2) + X_2 g_1(X_1) \tag{10.32}$$

by removing the parameters $\beta_{11}, \beta_{12}, \beta_{14}$, and β_{15} from the model. The previous table of Wald statistics included a test of adequacy of this reduced form ($\chi^2 = 5.41$ on 4 d.f., $P = .248$). The resulting fit is in Figure 10.13.

```
f ← lrm(sigdz ∼ sex*rcs(age,4) + rcs(cholesterol,4) +
          rcs(age,4) %ia% rcs(cholesterol,4), data=acath)
latex(anova(f), file='', size='smaller',
      caption='Singly nonlinear cubic spline surface',
      label='tab:anova-ria') #Table 10.8
```

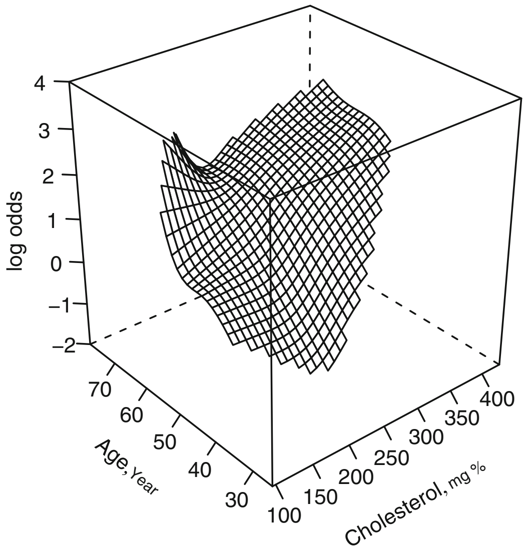

Fig. 10.12 Restricted cubic spline surface in two variables, each with $k = 4$ knots

Table 10.8 Singly nonlinear cubic spline surface

	χ^2	d.f.	P
sex (Factor+Higher Order Factors)	343.42	4	< 0.0001
All Interactions	24.05	3	< 0.0001
age (Factor+Higher Order Factors)	169.35	11	< 0.0001
All Interactions	34.80	8	< 0.0001
Nonlinear (Factor+Higher Order Factors)	16.55	6	0.0111
cholesterol (Factor+Higher Order Factors)	93.62	8	< 0.0001
All Interactions	10.83	5	0.0548
Nonlinear (Factor+Higher Order Factors)	10.87	4	0.0281
age × cholesterol (Factor+Higher Order Factors)	10.83	5	0.0548
Nonlinear	3.12	4	0.5372
Nonlinear Interaction : f(A,B) vs. AB	3.12	4	0.5372
Nonlinear Interaction in age vs. Af(B)	1.60	2	0.4496
Nonlinear Interaction in cholesterol vs. Bg(A)	1.64	2	0.4400
sex × age (Factor+Higher Order Factors)	24.05	3	< 0.0001
Nonlinear	13.58	2	0.0011
Nonlinear Interaction : f(A,B) vs. AB	13.58	2	0.0011
TOTAL NONLINEAR	27.89	10	0.0019
TOTAL INTERACTION	34.80	8	< 0.0001
TOTAL NONLINEAR + INTERACTION	45.45	12	< 0.0001
TOTAL	453.10	15	< 0.0001

```
# Figure 10.13:
bplot(Predict(f, cholesterol, age, np=40), perim=perim,
      lfun=wireframe, zlim=zl, adj.subtitle=FALSE)
ltx(f)
```

Table 10.9 Linear interaction surface

	χ^2	d.f.	P
age (Factor+Higher Order Factors)	167.83	7	< 0.0001
All Interactions	31.03	4	< 0.0001
Nonlinear (Factor+Higher Order Factors)	14.58	4	0.0057
sex (Factor+Higher Order Factors)	345.88	4	< 0.0001
All Interactions	22.30	3	0.0001
cholesterol (Factor+Higher Order Factors)	89.37	4	< 0.0001
All Interactions	7.99	1	0.0047
Nonlinear	10.65	2	0.0049
age × cholesterol (Factor+Higher Order Factors)	7.99	1	0.0047
age × sex (Factor+Higher Order Factors)	22.30	3	0.0001
Nonlinear	12.06	2	0.0024
Nonlinear Interaction : f(A,B) vs. AB	12.06	2	0.0024
TOTAL NONLINEAR	25.72	6	0.0003
TOTAL INTERACTION	31.03	4	< 0.0001
TOTAL NONLINEAR + INTERACTION	43.59	8	< 0.0001
TOTAL	452.75	11	< 0.0001

$X\hat{\beta} = -7.2 + 2.96[\text{female}] + 0.164\text{age} + 7.23\times10^{-5}(\text{age}-36)_+^3 - 0.000106(\text{age}-48)_+^3 - 1.63\times10^{-5}(\text{age}-56)_+^3 + 4.99\times10^{-5}(\text{age}-68)_+^3 + 0.0148\text{cholesterol} + 1.21\times10^{-6}(\text{cholesterol}-160)_+^3 - 5.5\times10^{-6}(\text{cholesterol}-208)_+^3 + 5.5\times10^{-6}(\text{cholesterol}-243)_+^3 - 1.21\times10^{-6}(\text{cholesterol}-319)_+^3 + \text{age}[-0.00029\,\text{cholesterol} + 9.28\times10^{-9}(\text{cholesterol}-160)_+^3 + 1.7\times10^{-8}(\text{cholesterol}-208)_+^3 - 4.43\times10^{-8}(\text{cholesterol}-243)_+^3 + 1.79\times10^{-8}(\text{cholesterol}-319)_+^3] + \text{cholesterol}[2.3\times10^{-7}(\text{age}-36)_+^3 + 4.21\times10^{-7}(\text{age}-48)_+^3 - 1.31\times10^{-6}(\text{age}-56)_+^3 + 6.64\times10^{-7}(\text{age}-68)_+^3] + [\text{female}][-0.111\text{age} + 8.03\times10^{-5}(\text{age}-36)_+^3 + 0.000135(\text{age}-48)_+^3 - 0.00044(\text{age}-56)_+^3 + 0.000224(\text{age}-68)_+^3].$

The fit is similar to the former one except that the climb in risk for low-cholesterol older subjects is less pronounced. The test for nonlinear interaction is now more concentrated ($P = .54$ with 4 d.f.). Figure 10.14 accordingly depicts a fit that allows age and cholesterol to have nonlinear main effects, but restricts the interaction to be a product between (untransformed) age and cholesterol. The function agrees substantially with the previous fit.

```
f ← lrm(sigdz ~ rcs(age,4)*sex + rcs(cholesterol,4) +
         age %ia% cholesterol, data=acath)
latex(anova(f), caption='Linear interaction surface', file='',
      size='smaller', label='tab:anova-lia') #Table 10.9
```

```
# Figure 10.14:
bplot(Predict(f, cholesterol, age, np=40), perim=perim,
      lfun=wireframe, zlim=zl, adj.subtitle=FALSE)
f.linia ← f  # save linear interaction fit for later
ltx(f)
```

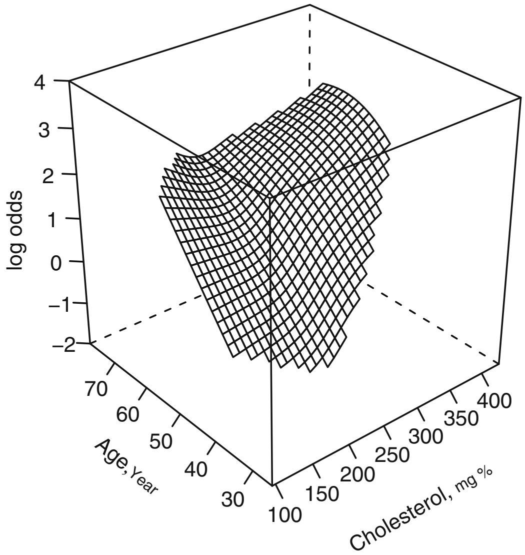

Fig. 10.13 Restricted cubic spline fit with age × spline(cholesterol) and cholesterol × spline(age)

$X\hat{\beta} = -7.36 + 0.182\text{age} - 5.18\times10^{-5}(\text{age} - 36)_+^3 + 8.45\times10^{-5}(\text{age} - 48)_+^3 - 2.91\times 10^{-6}(\text{age} - 56)_+^3 - 2.99\times10^{-5}(\text{age} - 68)_+^3 + 2.8[\text{female}] + 0.0139\text{cholesterol} + 1.76\times10^{-6}(\text{cholesterol} - 160)_+^3 - 4.88\times10^{-6}(\text{cholesterol} - 208)_+^3 + 3.45\times 10^{-6}(\text{cholesterol} - 243)_+^3 - 3.26\times10^{-7}(\text{cholesterol} - 319)_+^3 - 0.00034\ \text{age} \times \text{cholesterol} + [\text{female}][-0.107\text{age} + 7.71\times10^{-5}(\text{age} - 36)_+^3 + 0.000115(\text{age} - 48)_+^3 - 0.000398(\text{age} - 56)_+^3 + 0.000205(\text{age} - 68)_+^3]$.

The Wald test for age × cholesterol interaction yields $\chi^2 = 7.99$ with 1 d.f., $P = .005$. These analyses favor the nonlinear model with simple product interaction in Figure 10.14 as best representing the relationships among cholesterol, age, and probability of prognostically severe coronary artery disease. A nomogram depicting this model is shown in Figure 10.21.

Using this simple product interaction model, Figure 10.15 displays predicted cholesterol effects at the mean age within each age tertile. Substantial agreement with Figure 10.9 is apparent.

```
# Make estimates of cholesterol effects for mean age in
# tertiles corresponding to initial analysis
mean.age ←
  with(acath,
       as.vector(tapply(age, age.tertile, mean, na.rm=TRUE)))
plot(Predict(f, cholesterol, age=round(mean.age,2),
              sex="male"),
     adj.subtitle=FALSE, ylim=yl) #3 curves, Figure 10.15
```

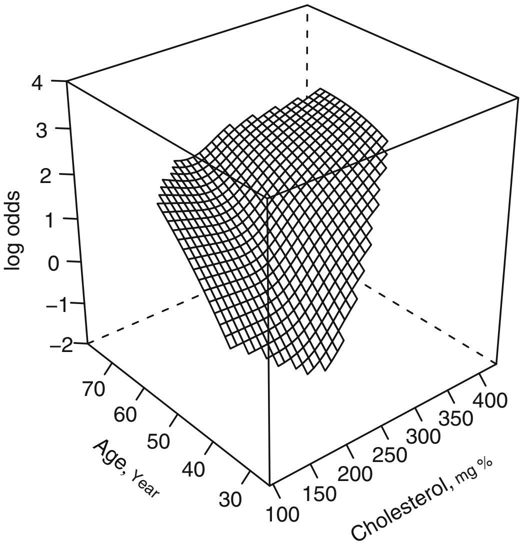

Fig. 10.14 Spline fit with nonlinear effects of cholesterol and age and a simple product interaction

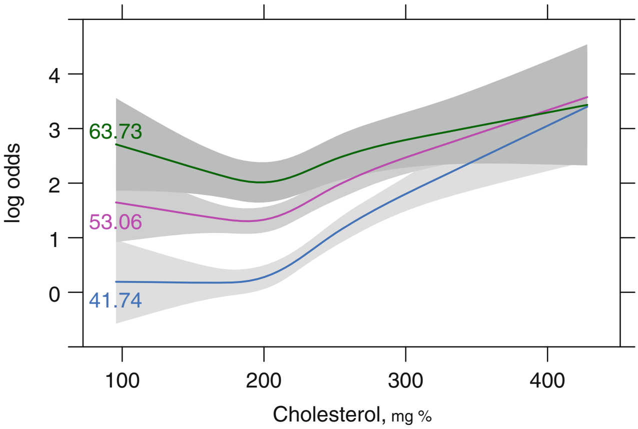

Fig. 10.15 Predictions from linear interaction model with mean age in tertiles indicated.

The partial residuals discussed in Section 10.4 can be used to check logistic model fit (although it may be difficult to deal with interactions). As an example, reconsider the "duration of symptoms" fit in Figure 10.7. Figure 10.16 displays "loess smoothed" and raw partial residuals for the original and log-transformed variable. The latter provides a more linear relationship, especially where the data are most dense.

Table 10.10 Merits of Methods for Checking Logistic Model Assumptions

Method	Choice Required	Assumes Additivity	Uses Ordering of X	Low Variance	Good Resolution on X
Stratification	Intervals				
Smoother on X_1 stratifying on X_2	Bandwidth		x (not on X_2)	x (if min. strat.)	x (X_1)
Smooth partial residual plot	Bandwidth	x	x	x	x
Spline model for all Xs	Knots	x	x	x	x

```
f ← lrm(tvdlm ∼ cad.dur, data=dz, x=TRUE, y=TRUE)
resid(f, "partial", pl="loess", xlim=c(0,250), ylim=c(-3,3))
scat1d(dz$cad.dur)
log.cad.dur ← log10(dz$cad.dur + 1)
f ← lrm(tvdlm ∼ log.cad.dur, data=dz, x=TRUE, y=TRUE)
resid(f, "partial", pl="loess", ylim=c(-3,3))
scat1d(log.cad.dur)   # Figure 10.16
```

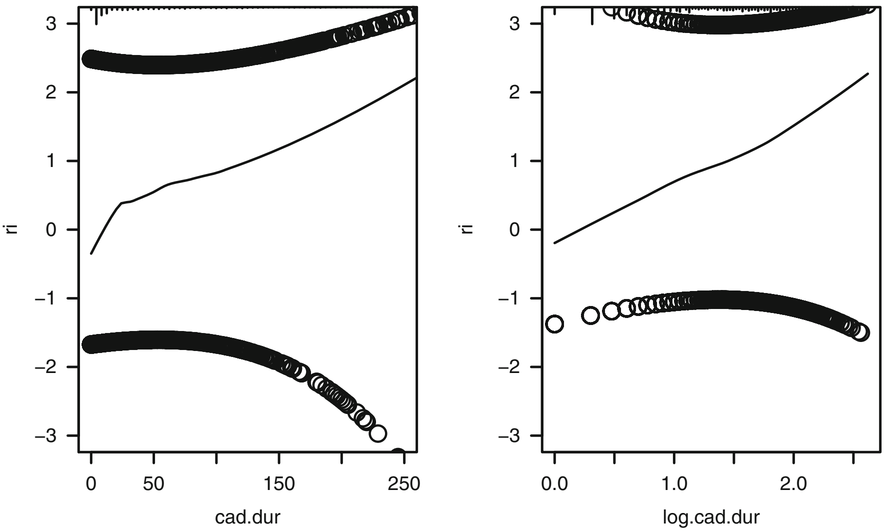

Fig. 10.16 Partial residuals for duration and $\log_{10}$(duration+1). Data density shown at top of each plot.

Table 10.10 summarizes the relative merits of stratification, nonparametric smoothers, and regression splines for determining or checking binary logistic model fits.

10.6 Collinearity

The variance inflation factors (VIFs) discussed in Section 4.6 can apply to any regression fit.[147, 654] These VIFs allow the analyst to isolate which variable(s) are responsible for highly correlated parameter estimates. Recall that, in general, collinearity is not a large problem compared with nonlinearity and overfitting.

10.7 Overly Influential Observations

Pregibon[511] developed a number of regression diagnostics that apply to the family of regression models of which logistic regression is a member. Influence statistics based on the "leave-out-one" method use an approximation to avoid having to refit the model n times for n observations. This approximation uses the fit and covariance matrix at the last iteration and assumes that the "weights" in the weighted least squares fit can be kept constant, yielding a computationally feasible one-step estimate of the leave-out-one regression coefficients.

Hosmer and Lemeshow [305, pp. 149–170] discuss many diagnostics for logistic regression and show how the final fit can be used in any least squares program that provides diagnostics. A new dependent variable to be used in that way is

$$Z_i = X\hat{\beta} + \frac{Y_i - \hat{P}_i}{V_i}, \tag{10.33}$$

where $V_i = \hat{P}_i(1-\hat{P}_i)$, and $\hat{P}_i = [1+\exp -X\hat{\beta}]^{-1}$ is the predicted probability that $Y_i = 1$. The $V_i, i = 1, 2, \ldots, n$ are used as weights in an ordinary weighted least squares fit of X against Z. This least squares fit will provide regression coefficients identical to b. The new standard errors will be off from the actual logistic model ones by a constant.

As discussed in Section 4.9, the standardized change in the regression coefficients upon leaving out each observation in turn (DFBETAS) is one of the most useful diagnostics, as these can pinpoint which observations are influential on each part of the model. After carefully modeling predictor transformations, there should be no lack of fit due to improper transformations. However, as the white blood count example in Section 4.9 indicates, it is commonly the case that extreme predictor values can still have too much influence on the estimates of coefficients involving that predictor.

In the age–sex–response example of Section 10.1.3, both DFBETAS and DFFITS identified the same influential observations. The observation given by age = 48 sex = female response = 1 was influential for both age and sex, while the observation age = 34 sex = male response = 1 was influential for age and the observation age = 50 sex = male response = 0 was influential for sex. It can readily be seen from Figure 10.3 that these points do not fit the overall trends in the data. However, as these data were simulated from a

Table 10.11 Example influence statistics

Females				Males			
DFBETAS			DFFITS	DFBETAS			DFFITS
Intercept	Age	Sex		Intercept	Age	Sex	
0.0	0.0	0.0	0	0.5	-0.5	-0.2	2
0.0	0.0	0.0	0	0.2	-0.3	0.0	1
0.0	0.0	0.0	0	-0.1	0.1	0.0	-1
0.0	0.0	0.0	0	-0.1	0.1	0.0	-1
-0.1	0.1	0.1	0	-0.1	0.1	-0.1	-1
-0.1	0.1	0.1	0	0.0	0.0	0.1	0
0.7	-0.7	-0.8	3	0.0	0.0	0.1	0
-0.1	0.1	0.1	0	0.0	0.0	0.1	0
-0.1	0.1	0.1	0	0.0	0.0	-0.2	-1
-0.1	0.1	0.1	0	0.1	-0.1	-0.2	-1
-0.1	0.1	0.1	0	0.0	0.0	0.1	0
-0.1	0.0	0.1	0	-0.1	0.1	0.1	0
-0.1	0.0	0.1	0	-0.1	0.1	0.1	0
0.1	0.0	-0.2	1	0.3	-0.3	-0.4	-2
0.0	0.0	0.1	-1	-0.1	0.1	0.1	0
0.1	-0.2	0.0	-1	-0.1	0.1	0.1	0
-0.1	0.2	0.0	1	-0.1	0.1	0.1	0
-0.2	0.2	0.0	1	0.0	0.0	0.0	0
-0.2	0.2	0.0	1	0.0	0.0	0.0	0
-0.2	0.2	0.1	1	0.0	0.0	0.0	0

population model that is truly linear in age and additive in age and sex, the apparent influential observations are just random occurrences. It is unwise to assume that in real data all points will agree with overall trends. Removal of such points would bias the results, making the model apparently more
11 predictive than it will be prospectively. See Table 10.11.

```
f ← update(fasr, x=TRUE, y=TRUE)
which.influence(f, .4)    # Table 10.11
```

10.8 Quantifying Predictive Ability

The test statistics discussed above allow one to test whether a factor or set of factors is related to the response. If the sample is sufficiently large, a factor that grades risk from .01 to .02 may be a significant risk factor. However, that factor is not very useful in predicting the response for an individual subject. There is controversy regarding the appropriateness of R^2 from ordinary least
12 squares in this setting.[136,424] The generalized R^2_{N} index of Nagelkerke[471] and Cragg and Uhler[137], Maddala[431], and Magee[432] described in Section 9.8.3 can be useful for quantifying the predictive strength of a model:

$$R^2_{\mathrm{N}} = \frac{1 - \exp(-\mathrm{LR}/n)}{1 - \exp(-L^0/n)}, \tag{10.34}$$

where LR is the global log likelihood ratio statistic for testing the importance of all p predictors in the model and L^0 is the -2 log likelihood for the null model. 13

Tjur[613] coined the term "coefficient of discrimination" D, defined as the average $\hat{P}$ when $Y = 1$ minus the average $\hat{P}$ when $Y = 0$, and showed how it ties in with sum of squares–based R^2 measures. D has many advantages as an index of predictive power[d].

Linnet[416] advocates quadratic and logarithmic probability scoring rules for measuring predictive performance for probability models. Linnet shows how to bootstrap such measures to get bias-corrected estimates and how to use bootstrapping to compare two correlated scores. The quadratic scoring rule is Brier's score, frequently used in judging meteorologic forecasts[30, 73]:

$$B = \frac{1}{n}\sum_{i=1}^{n}(\hat{P}_i - Y_i)^2, \tag{10.35}$$

where $\hat{P}_i$ is the predicted probability and Y_i the corresponding observed response for the ith observation. 14

A unitless index of the strength of the rank correlation between predicted probability of response and actual response is a more interpretable measure of the fitted model's predictive discrimination. One such index is the probability of concordance, c, between predicted probability and response. The c index, which is derived from the Wilcoxon–Mann–Whitney two-sample rank test, is computed by taking all possible pairs of subjects such that one subject responded and the other did not. The index is the proportion of such pairs with the responder having a higher predicted probability of response than the nonresponder.

Bamber[39] and Hanley and McNeil[255] have shown that c is identical to a widely used measure of diagnostic discrimination, the area under a "receiver operating characteristic" (ROC) curve. A value of c of .5 indicates random predictions, and a value of 1 indicates perfect prediction (i.e., perfect separation of responders and nonresponders). A model having c greater than roughly .8 has some utility in predicting the responses of individual subjects. The concordance index is also related to another widely used index, Somers' D_{xy} rank correlation[579] between predicted probabilities and observed responses, by the identity

$$D_{xy} = 2(c - .5). \tag{10.36}$$

D_{xy} is the difference between concordance and discordance probabilities. When $D_{xy} = 0$, the model is making random predictions. When $D_{xy} = 1$,

[d] Note that D and B (below) and other indexes not related to c (below) do not work well in case-control studies because of their reliance on absolute probability estimates.

the predictions are perfectly discriminating. These rank-based indexes have 15 the advantage of being insensitive to the prevalence of positive responses.

A commonly used measure of predictive ability for binary logistic models is the fraction of correctly classified responses. Here one chooses a cutoff on the predicted probability of a positive response and then predicts that a response will be positive if the predicted probability exceeds this cutoff. There are a number of reasons why this measure should be avoided.

1. It's highly dependent on the cutpoint chosen for a "positive" prediction.
2. You can add a highly significant variable to the model and have the percentage classified correctly actually decrease. Classification error is a very insensitive and statistically inefficient measure[264, 633] since if the threshold for "positive" is, say 0.75, a prediction of 0.99 rates the same as one of 0.751.
3. It gets away from the purpose of fitting a logistic model. A logistic model is a model for the probability of an event, not a model for the occurrence of the event. For example, suppose that the event we are predicting is the probability of being struck by lightning. Without having any data, we would predict that you won't get struck by lightning. However, you might develop an interesting model that discovers real risk factors that yield probabilities of being struck that range from 0.000000001 to 0.001.
4. If you make a classification rule from a probability model, you are being presumptuous. Suppose that a model is developed to assist physicians in diagnosing a disease. Physicians sometimes profess to desiring a binary decision model, but if given a probability they will rightfully apply different thresholds for treating different patients or for ordering other diagnostic tests. Even though the age of the patient may be a strong predictor of the probability of disease, the physician will often use a lower threshold of disease likelihood for treating a young patient. This usage is above and beyond how age affects the likelihood.
5. If a disease were present in only 0.02 of the population, one could be 0.98 accurate in diagnosing the disease by ruling that everyone is disease–free, i.e., by avoiding predictors. The proportion classified correctly fails to take the difficulty of the task into account.
6. van Houwelingen and le Cessie[633] demonstrated a peculiar property that occurs when you try to obtain an honest estimate of classification error using cross-validation. The cross-validated error rate corrects the apparent error rate only if the predicted probability is exactly $1/2$ or is $1/2 \pm 1/(2n)$. The cross-validation estimate of optimism is "zero for n even and negligibly small for n odd." Better measures of error rate such as the Brier score and logarithmic scoring rule do not have this problem. They also have the nice property of being maximized when the predicted probabilities are the 16 population probabilities.[416].

10.9 Validating the Fitted Model

The major cause of unreliable models is overfitting the data. The methods described in Section 5.3 can be used to assess the accuracy of models fairly. If a sample has been held out and never used to study associations with the response, indexes of predictive accuracy can now be estimated using that sample. More efficient is cross-validation, and bootstrapping is the most efficient validation procedure. As discussed earlier, bootstrapping does not require holding out any data, since all aspects of model development (stepwise variable selection, tests of linearity, estimation of coefficients, etc.) are revalidated on samples taken with replacement from the whole sample.

Cox[130] proposed and Harrell and Lee[267] and Miller et al.[457] further developed the idea of fitting a new binary logistic model to a new sample to estimate the relationship between the predicted probability and the observed outcome in that sample. This fit provides a simple calibration equation that can be used to quantify unreliability (lack of calibration) and to calibrate the predictions for future use. This logistic calibration also leads to indexes of unreliability (U), discrimination (D), and overall quality ($Q = D - U$) which are derived from likelihood ratio tests[267]. Q is a logarithmic scoring rule, which can be compared with Brier's index (Equation 10.35). See [633] for many more ideas.

With bootstrapping we do not have a separate validation sample for assessing calibration, but we can estimate the overoptimism in assuming that the final model needs no calibration, that is, it has overall intercept=0 and slope=1. As discussed in Section 5.3, refitting the model

$$P_c = \text{Prob}\{Y = 1|X\hat{\beta}\} = [1 + \exp -(\gamma_0 + \gamma_1 X\hat{\beta})]^{-1} \tag{10.37}$$

(where P_c denotes the calibrated probability and the original predicted probability is $\hat{P} = [1 + \exp(-X\hat{\beta})]^{-1}$) in the original sample will always result in $\gamma = (\gamma_0, \gamma_1) = (0, 1)$, since a logistic model will always "fit" the training sample when assessed overall. We thus estimate γ by using Efron's[172] method to estimate the overoptimism in $(0, 1)$ to obtain bias-corrected estimates of the true calibration. Simulations have shown this method produces an efficient estimate of γ.[259]

More stringent calibration checks can be made by running separate calibrations for different covariate levels. Smooth nonparametric curves described in Section 10.11 are more flexible than the linear-logit calibration method just described.

A good set of indexes to estimate for summarizing a model validation is the c or D_{xy} indexes and measures of calibration. In addition, the overoptimism in the indexes may be reported to quantify the amount of overfitting present. The estimate of γ can be used to draw a calibration curve by plotting $\hat{P}$ on the x-axis and $\hat{P}_c = [1 + \exp -(\gamma_0 + \gamma_1 L)]^{-1}$ on the y-axis, where $L = \text{logit}(\hat{P})$.[130, 267] An easily interpreted index of unreliability, E_{max}, follows immediately from this calibration model:

$$E_{max}(a,b) = \max_{a \le \hat{P} \le b} |\hat{P} - \hat{P}_c|, \tag{10.38}$$

the maximum error in predicted probabilities over the range $a \le \hat{P} \le b$. In some cases, we would compute the maximum absolute difference in predicted and calibrated probabilities over the entire interval, that is, use $E_{max}(0,1)$. The null hypothesis $H_0 : E_{max}(0,1) = 0$ can easily be tested by testing $H_0 : \gamma_0 = 0, \gamma_1 = 1$ as above. Since E_{max} does not weight the discrepancies by the actual distribution of predictions, it may be preferable to compute the average absolute discrepancy over the actual distribution of predictions (or to use a mean squared error, incorporating the same calibration function).

If stepwise variable selection is being done, a matrix depicting which factors are selected at each bootstrap sample will shed light on how arbitrary is the selection of "significant" factors. See Section 5.3 for reasons to compare full and stepwise model fits.

As an example using bootstrapping to validate the calibration and discrimination of a model, consider the data in Section 10.1.3. Using 150 samples with replacement, we first validate the additive model with age and sex forced into every model. The optimism-corrected discrimination and calibration statistics produced by `validate` (see Section 10.11) are in the table below.

```
d ← sex.age.response
dd ← datadist(d); options(datadist='dd')
f ← lrm(response ~ sex + age, data=d, x=TRUE, y=TRUE)
set.seed(3)  # for reproducibility
v1  ← validate(f, B=150)
```

```
latex(v1,
      caption='Bootstrap Validation, 2 Predictors Without
      Stepdown', digits=2, size='Ssize', file='')
```

Bootstrap Validation, 2 Predictors Without Stepdown

Index	Original Sample	Training Sample	Test Sample	Optimism	Corrected Index	n
D_{xy}	0.70	0.70	0.67	0.04	0.66	150
R^2	0.45	0.48	0.43	0.05	0.40	150
Intercept	0.00	0.00	0.01	−0.01	0.01	150
Slope	1.00	1.00	0.91	0.09	0.91	150
E_{max}	0.00	0.00	0.02	0.02	0.02	150
D	0.39	0.44	0.36	0.07	0.32	150
U	−0.05	−0.05	0.04	−0.09	0.04	150
Q	0.44	0.49	0.32	0.16	0.28	150
B	0.16	0.15	0.18	−0.03	0.19	150
g	2.10	2.49	1.97	0.52	1.58	150
g_p	0.35	0.35	0.34	0.01	0.34	150

Now we incorporate variable selection. The variables selected in the first 10 bootstrap replications are shown below. The apparent Somers' D_{xy} is 0.7, and the bias-corrected D_{xy} is 0.66. The slope shrinkage factor is 0.91. The maximum absolute error in predicted probability is estimated to be 0.02.

We next allow for step-down variable selection at each resample. For illustration purposes only, we use a suboptimal stopping rule based on significance of *individual* variables at the $\alpha = 0.10$ level. Of the 150 repetitions, both age and sex were selected in 137, and neither variable was selected in 3 samples. The validation statistics are in the table below.

```
v2 ← validate(f, B=150, bw=TRUE,
              rule='p', sls=.1, type='individual')
```

```
latex(v2,
      caption='Bootstrap Validation, 2 Predictors with Stepdown',
      digits=2, B=15, file='', size='Ssize')
```

Bootstrap Validation, 2 Predictors with Stepdown

Index	Original Sample	Training Sample	Test Sample	Optimism	Corrected Index	n
D_{xy}	0.70	0.70	0.64	0.07	0.63	150
R^2	0.45	0.49	0.41	0.09	0.37	150
Intercept	0.00	0.00	−0.04	0.04	−0.04	150
Slope	1.00	1.00	0.84	0.16	0.84	150
$E_{\max}$	0.00	0.00	0.05	0.05	0.05	150
D	0.39	0.45	0.34	0.11	0.28	150
U	−0.05	−0.05	0.06	−0.11	0.06	150
Q	0.44	0.50	0.28	0.22	0.22	150
B	0.16	0.14	0.18	−0.04	0.20	150
g	2.10	2.60	1.88	0.72	1.38	150
g_p	0.35	0.35	0.33	0.02	0.33	150

Factors Retained in Backwards Elimination
First 15 Resamples

sex	age
●	●
●	●
●	●
●	●
●	●
●	●
●	●
●	●
●	●
●	●
●	●
●	●
●	●
●	●
●	

Frequencies of Numbers of Factors Retained

0	1	2
3	10	137

The apparent Somers' D_{xy} is 0.7 for the original stepwise model (which actually retained both age and sex), and the bias-corrected D_{xy} is 0.63, slightly worse than the more correct model which forced in both variables. The calibration was also slightly worse as reflected in the slope correction factor estimate of 0.84 versus 0.91.

Next, five additional candidate variables are considered. These variables are random uniform variables, $x1, \ldots, x5$ on the $[0, 1]$ interval, and have no association with the response.

```
set.seed(133)
n  ← nrow(d)
x1 ← runif(n)
x2 ← runif(n)
x3 ← runif(n)
x4 ← runif(n)
x5 ← runif(n)
f  ← lrm(response ~ age + sex + x1 + x2 + x3 + x4 + x5,
         data=d, x=TRUE, y=TRUE)
v3 ← validate(f, B=150, bw=TRUE,
              rule='p', sls=.1, type='individual')
```

```
k ← attr(v3, 'kept')
# Compute number of x1-x5 selected
nx ← apply(k[,3:7], 1, sum)
# Get selections of age and sex
v ← colnames(k)
as ← apply(k[,1:2], 1,
             function(x) paste(v[1:2][x], collapse=', '))
table(paste(as, ' ', nx, 'Xs'))
```

```
          0 Xs            1 Xs       age    2 Xs  age, sex    0 Xs
            50               3                1                 34
age, sex    1 Xs  age, sex    2 Xs  age, sex    3 Xs  age, sex    4 Xs
            17              11                7                  1
     sex    0 Xs       sex    1 Xs
            12               3
```

```
latex(v3,
 caption='Bootstrap Validation with 5 Noise Variables and
 Stepdown', digits=2, B=15, size='Ssize', file='')
```

Bootstrap Validation with 5 Noise Variables and Stepdown

Index	Original Sample	Training Sample	Test Sample	Optimism	Corrected Index	n
D_{xy}	0.70	0.47	0.38	0.09	0.60	139
R^2	0.45	0.34	0.23	0.11	0.34	139
Intercept	0.00	0.00	0.03	−0.03	0.03	139
Slope	1.00	1.00	0.78	0.22	0.78	139
$E_{\max}$	0.00	0.00	0.06	0.06	0.06	139
D	0.39	0.31	0.18	0.13	0.26	139
U	−0.05	−0.05	0.07	−0.12	0.07	139
Q	0.44	0.36	0.11	0.25	0.19	139
B	0.16	0.17	0.22	−0.04	0.20	139
g	2.10	1.81	1.06	0.75	1.36	139
g_p	0.35	0.23	0.19	0.04	0.31	139

Factors Retained in Backwards Elimination
First 15 Resamples

age	sex	x1	x2	x3	x4	x5
●	●		●	●	●	●
●	●	●				●
●	●					
●	●				●	●
●	●	●				
●	●					
●	●					
●	●		●			
●	●			●		

Frequencies of Numbers of Factors Retained

0	1	2	3	4	5	6
50	15	37	18	11	7	1

Using step-down variable selection with the same stopping rule as before, the "final" model on the original sample correctly deleted $x1, \ldots, x5$. Of the 150 bootstrap repetitions, 11 samples yielded a singularity or non-convergence either in the full-model fit or after step-down variable selection. Of the 139 successful repetitions, the frequencies of the number of factors selected, as well as the frequency of variable combinations selected, are shown above. Validation statistics are also shown above.

Figure 10.17 depicts the calibration (reliability) curves for the three strategies using the corrected intercept and slope estimates in the above tables as γ_0 and γ_1, and the logistic calibration model $P_c = [1 + \exp -(\gamma_0 + \gamma_1 L)]^{-1}$, where P_c is the "actual" or calibrated probability, L is logit($\hat{P}$), and $\hat{P}$ is the predicted probability. The shape of the calibration curves (driven by slopes < 1) is typical of overfitting—low predicted probabilities are too low and high predicted probabilities are too high. Predictions near the overall prevalence of the outcome tend to be calibrated even when overfitting is present.

```
g ← function(v) v[c('Intercept','Slope'),'index.corrected']
k ← rbind(g(v1), g(v2), g(v3))
co ← c(2,5,4,1)
plot(0, 0, ylim=c(0,1), xlim=c(0,1),
     xlab="Predicted Probability",
     ylab="Estimated Actual Probability", type="n")
legend(.45,.35,c("age, sex", "age, sex stepdown",
                 "age, sex, x1-x5", "ideal"),
       lty=1, col=co, cex=.8, bty="n")
probs ← seq(0, 1, length=200); L ← qlogis(probs)
for(i in 1:3) {
  P ← plogis(k[i,'Intercept'] + k[i,'Slope'] * L)
  lines(probs, P, col=co[i], lwd=1)
}
abline(a=0, b=1, col=co[4], lwd=1)   # Figure 10.17
```

"Honest" calibration curves may also be estimated using nonparametric smoothers in conjunction with bootstrapping and cross-validation (see Section 10.11).

10.10 Describing the Fitted Model

Once the proper variables have been modeled and all model assumptions have been met, the analyst needs to present and interpret the fitted model. There are at least three ways to proceed. The coefficients in the model may be interpreted. For each variable, the change in log odds for a sensible change in the variable value (e.g., interquartile range) may be computed. Also, the odds

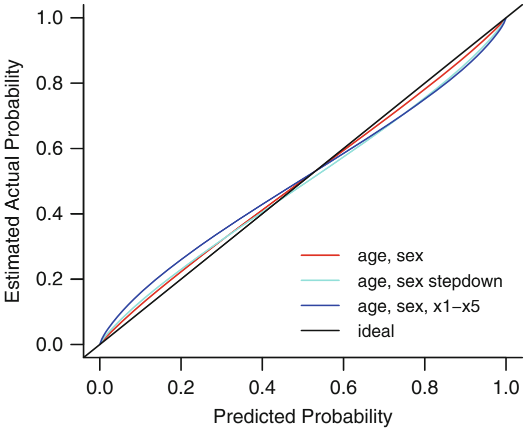

Fig. 10.17 Estimated logistic calibration (reliability) curves obtained by bootstrapping three modeling strategies.

Table 10.12 Effects Response : sigdz

	Low	High	Δ	Effect	S.E.	Lower 0.95	Upper 0.95
age	46	59	13	0.90629	0.18381	0.546030	1.26650
Odds Ratio	46	59	13	2.47510		1.726400	3.54860
cholesterol	196	259	63	0.75479	0.13642	0.487410	1.02220
Odds Ratio	196	259	63	2.12720		1.628100	2.77920
sex — female:male	1	2		-2.42970	0.14839	-2.720600	-2.13890
Odds Ratio	1	2		0.08806		0.065837	0.11778

ratio or factor by which the odds increases for a certain change in a predictor, holding all other predictors constant, may be displayed. Table 10.12 contains such summary statistics for the linear age × cholesterol interaction surface fit described in Section 10.5.

```
s ← summary(f.linia)    # Table 10.12
latex(s, file='', size='Ssize',
      label='tab:lrm-cholxage-confbar')
```

```
plot(s)    # Figure 10.18
```

The outer quartiles of age are 46 and 59 years, so the “half-sample” odds ratio for age is 2.47, with 0.95 confidence interval $[1.63, 3.74]$ when sex is male and cholesterol is set to its median. The effect of increasing cholesterol from 196 (its lower quartile) to 259 (its upper quartile) is to increase the log odds by 0.79 or to increase the odds by a factor of 2.21. Since there are interactions allowed between age and sex and between age and cholesterol, each odds ratio in the above table depends on the setting of at least one other factor. The

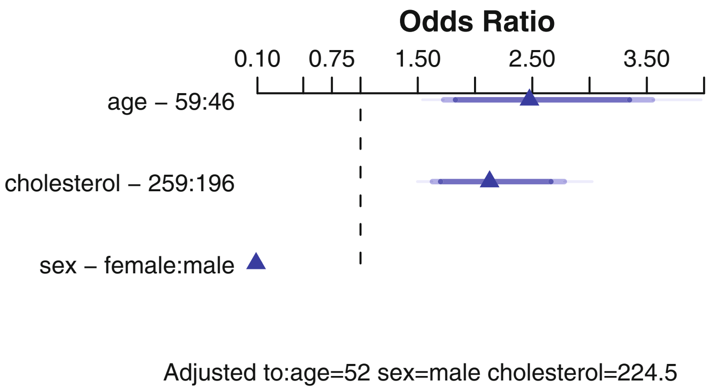

Fig. 10.18 Odds ratios and confidence bars, using quartiles of age and cholesterol for assessing their effects on the odds of coronary disease

results are shown graphically in Figure 10.18. The shaded confidence bars show various levels of confidence and do not pin the analyst down to, say, the 0.95 level.

For those used to thinking in terms of odds or log odds, the preceding description may be sufficient. Many prefer instead to interpret the model in terms of predicted probabilities instead of odds. If the model contains only a single predictor (even if several spline terms are required to represent that predictor), one may simply plot the predictor against the predicted response. Such a plot is shown in Figure 10.19 which depicts the fitted relationship between age of diagnosis and the probability of acute bacterial meningitis (ABM) as opposed to acute viral meningitis (AVM), based on an analysis of 422 cases from Duke University Medical Center.[580] The data may be found on the web site. A linear spline function with knots at 1, 2, and 22 years was used to model this relationship.

When the model contains more than one predictor, one may graph the predictor against log odds, and barring interactions, the shape of this relationship will be independent of the level of the other predictors. When displaying the model on what is usually a more interpretable scale, the probability scale, a difficulty arises in that unlike log odds the relationship between one predictor and the probability of response depends on the levels of all other factors. For example, in the model

$$\text{Prob}\{Y = 1|X\} = \{1 + \exp[-(\beta_0 + \beta_1 X_1 + \beta_2 X_2)]\}^{-1} \qquad (10.39)$$

there is no way to factor out X_1 when examining the relationship between X_2 and the probability of a response. For the two-predictor case one can plot X_2 versus predicted probability for each level of X_1. When it is uncertain whether to include an interaction in this model, consider presenting graphs for two models (with and without interaction terms included) as was done in [658].

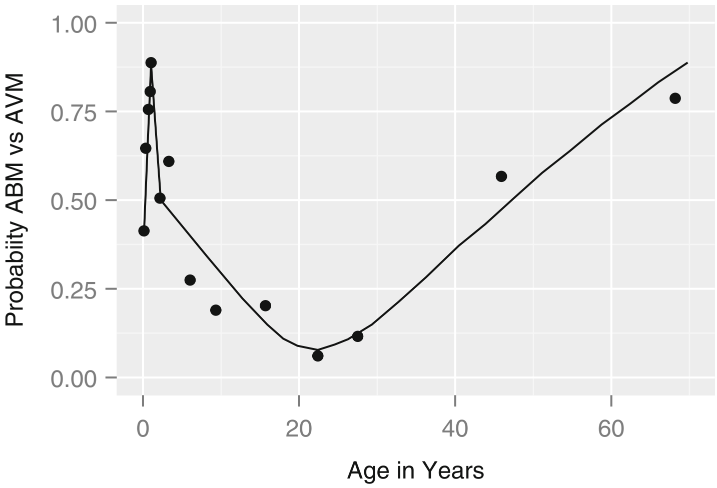

Fig. 10.19 Linear spline fit for probability of bacterial versus viral meningitis as a function of age at onset[580]. Points are simple proportions by age quantile groups.

When three factors are present, one could draw a separate graph for each level of X_3, a separate curve on each graph for each level of X_1, and vary X_2 on the x-axis. Instead of this, or if more than three factors are present, a good way to display the results may be to plot "adjusted probability estimates" as a function of one predictor, adjusting all other factors to constants such as the mean. For example, one could display a graph relating serum cholesterol to probability of myocardial infarction or death, holding age constant at 55, sex at 1 (male), and systolic blood pressure at 120 mmHg.

The final method for displaying the relationship between several predictors and probability of response is to construct a nomogram.[40, 254] A nomogram not only sheds light on how the effect of one predictor on the probability of response depends on the levels of other factors, but it allows one to quickly estimate the probability of response for individual subjects. The nomogram in Figure 10.20 allows one to predict the probability of acute bacterial meningitis (given the patient has either viral or bacterial meningitis) using the same sample as in Figure 10.19. Here there are four continuous predictor values, none of which are linearly related to log odds of bacterial meningitis: age at admission (expressed as a linear spline function), month of admission (expressed as $|\text{month} - 8|$), cerebrospinal fluid glucose/blood glucose ratio (linear effect truncated at .6; that is, the effect is the glucose ratio if it is $\leq .6$, and .6 if it exceeded .6), and the cube root of the total number of polymorphonuclear leukocytes in the cerebrospinal fluid. 17

The model associated with Figure 10.14 is depicted in what could be called a "precision nomogram" in Figure 10.21. Discrete cholesterol levels were required because of the interaction between two continuous variables.

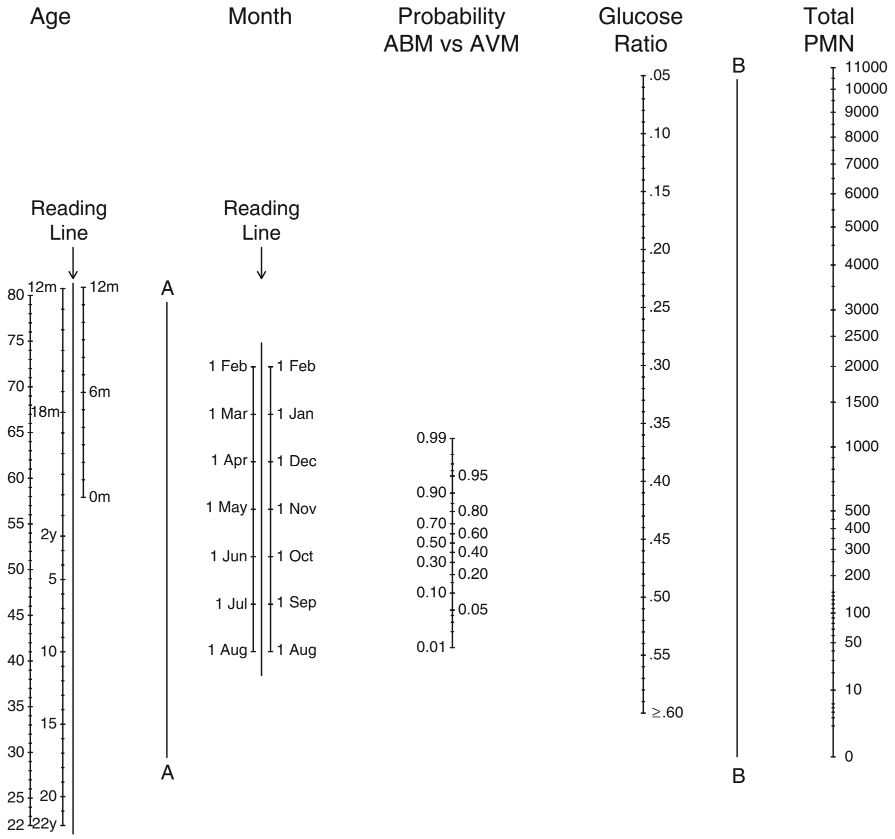

Fig. 10.20 Nomogram for estimating probability of bacterial (ABM) versus viral (AVM) meningitis. Step 1, place ruler on reading lines for patient's age and month of presentation and mark intersection with line A; step 2, place ruler on values for glucose ratio and total polymorphonuclear leukocyte (PMN) count in cerebrospinal fluid and mark intersection with line B; step 3, use ruler to join marks on lines A and B, then read off the probability of ABM versus AVM.[580]

```
# Draw a nomogram that shows examples of confidence intervals
nom ← nomogram(f.linia, cholesterol=seq(150, 400, by=50),
                interact=list(age=seq(30, 70, by=10)),
                lp.at=seq(-2, 3.5, by=.5),
                conf.int=TRUE, conf.lp="all",
                fun=function(x)1/(1+exp(-x)),  # or plogis
                funlabel="Probability of CAD",
                fun.at=c(seq(.1, .9, by=.1), .95, .99)
                )   # Figure 10.21
plot(nom, col.grid = gray(c(0.8, 0.95)),
     varname.label=FALSE, ia.space=1, xfrac=.46, lmgp=.2)
```

10.11 R Functions

The general R statistical modeling functions[96] described in Section 6.2 work with the author's `lrm` function for fitting binary and ordinal logistic regression models. `lrm` has several options for doing penalized maximum likelihood estimation, with special treatment of categorical predictors so as to shrink all estimates (including the reference cell) to the mean. The following example [18] fits a logistic model containing predictors `age`, `blood.pressure`, and `sex`, with `age` fitted with a smooth five-knot restricted cubic spline function and a different shape of the age relationship for males and females.

```
fit ← lrm(death ~ blood.pressure + sex * rcs(age,5))
anova(fit)
plot(Predict(fit, age, sex))
```

The `pentrace` function makes it easy to check the effects of a sequence of penalties. The following code fits an unpenalized model and plots the AIC and Schwarz BIC for a variety of penalties so that approximately the best cross-validating model can be chosen (and so we can learn how the penalty relates to the effective degrees of freedom). Here we elect to only penalize the nonlinear or non-additive parts of the model.

```
f ← lrm(death ~ rcs(age,5)*treatment + lsp(sbp,c(120,140)),
         x=TRUE, y=TRUE)
plot(pentrace(f,
              penalty=list(nonlinear=seq(.25,10,by=.25))) )
```

See Sections 9.8.1 and 9.10 for more information. [19]

The `residuals` function for `lrm` and the `which.influence` function can be used to check predictor transformations as well as to analyze overly influential observations in binary logistic regression. See Figure 10.16 for one application. The `residuals` function will also perform the unweighted sum of squares test for global goodness of fit described in Section 10.5.

The `validate` function when used on an object created by `lrm` does resampling validation of a logistic regression model, with or without backward step-down variable deletion. It provides bias-corrected Somers' D_{xy} rank correlation, R^2_{N} index, the intercept and slope of an overall logistic calibration equation, the maximum absolute difference in predicted and calibrated probabilities E_{max}, the discrimination index D [(model L.R. $\chi^2 - 1)/n$], the unreliability index $U =$ (difference in -2 log likelihood between uncalibrated $X\beta$ and $X\beta$ with overall intercept and slope calibrated to test sample)$/n$, and the overall quality index $Q = D - U$.[267] The "corrected" slope can be thought of as a shrinkage factor that takes overfitting into account. See `predab.resample` in Section 6.2 for the list of resampling methods.

The `calibrate` function produces bootstrapped or cross-validated calibration curves for logistic and linear models. The "apparent" calibration accuracy is estimated using a nonparametric smoother relating predicted probabilities

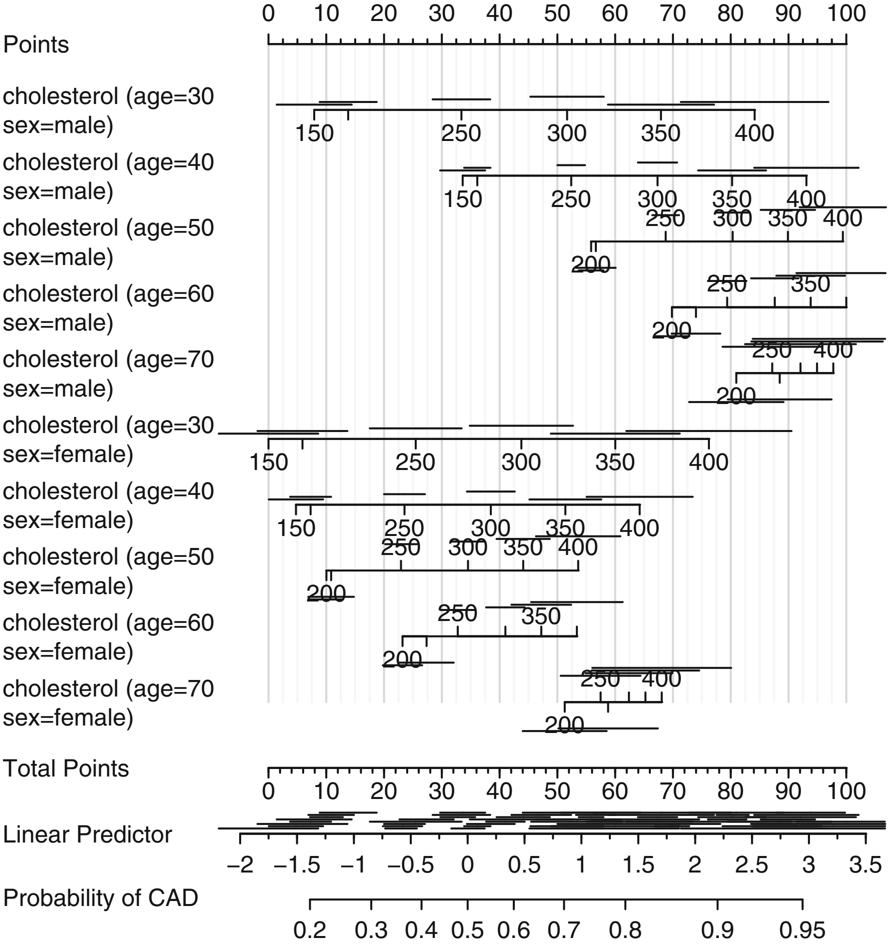

Fig. 10.21 Nomogram relating age, sex, and cholesterol to the log odds and to the probability of significant coronary artery disease. Select one axis corresponding to sex and to age $\in \{30, 40, 50, 60, 70\}$. There is linear interaction between age and sex and between age and cholesterol. 0.70 and 0.90 confidence intervals are shown (0.90 in gray). Note that for the "Linear Predictor" scale there are various lengths of confidence intervals near the same value of $X\hat{\beta}$, demonstrating that the standard error of $X\hat{\beta}$ depends on the individual X values. Also note that confidence intervals corresponding to smaller patient groups (e.g., females) are wider.

to observed binary outcomes. The nonparametric estimate is evaluated at a sequence of predicted probability levels. Then the distances from the 45° line are compared with the differences when the current model is evaluated back on the whole sample (or omitted sample for cross-validation). The differences in the differences are estimates of overoptimism. After averaging over many replications, the predicted-value-specific differences are then subtracted from the apparent differences and an adjusted calibration curve is obtained. Unlike `validate`, `calibrate` does not assume a linear logistic calibration. For an example, see the end of Chapter 11. `calibrate` will print the mean absolute calibration error, the 0.9 quantile of the absolute error, and the mean squared error, all over the observed distribution of predicted values.

The `val.prob` function is used to compute measures of discrimination and calibration of predicted probabilities for a separate sample from the one used to derive the probability estimates. Thus `val.prob` is used in external validation and data-splitting. The function computes similar indexes as `validate` plus the Brier score and a statistic for testing for unreliability or $H_0 : \gamma_0 = 0, \gamma_1 = 1$.

In the following example, a logistic model is fitted on 100 observations simulated from the actual model given by

$$\text{Prob}\{Y = 1|X_1, X_2, X_3\} = [1 + \exp[-(-1 + 2X_1)]]^{-1}, \qquad (10.40)$$

where X_1 is a random uniform $[0, 1]$ variable. Hence X_2 and X_3 are irrelevant. After fitting a linear additive model in X_1, X_2, and X_3, the coefficients are used to predict $\text{Prob}\{Y = 1\}$ on a separate sample of 100 observations.

```
set.seed(13)
n ← 200
x1 ← runif(n)
x2 ← runif(n)
x3 ← runif(n)
logit ← 2*(x1-.5)
P ← 1/(1+exp(-logit))
y ← ifelse(runif(n) ≤ P, 1, 0)
d ← data.frame(x1, x2, x3, y)
f ← lrm(y ∼ x1 + x2 + x3, subset=1:100)
phat ← predict(f, d[101:200,], type='fitted')
# Figure 10.22
v ← val.prob(phat, y[101:200], m=20, cex=.5)
```

The output is shown in Figure 10.22.

The R built-in function `glm`, a very general modeling function, can fit binary logistic models. The response variable *must* be coded 0/1 for `glm` to work. `Glm` is a slight modification of the built-in `glm` function in the `rms` package that allows fits to use `rms` methods. This facilitates Poisson and several other types of regression analysis.

10.12 Further Reading

1 See [590] for modeling strategies specific to binary logistic regression.

2 See [632] for a nice review of logistic modeling. Agresti[6] is an excellent source for categorical Y in general.

3 Not only does discriminant analysis assume the same regression model as logistic regression, but it also assumes that the predictors are each normally distributed and that jointly the predictors have a multivariate normal distribution. These assumptions are unlikely to be met in practice, especially when

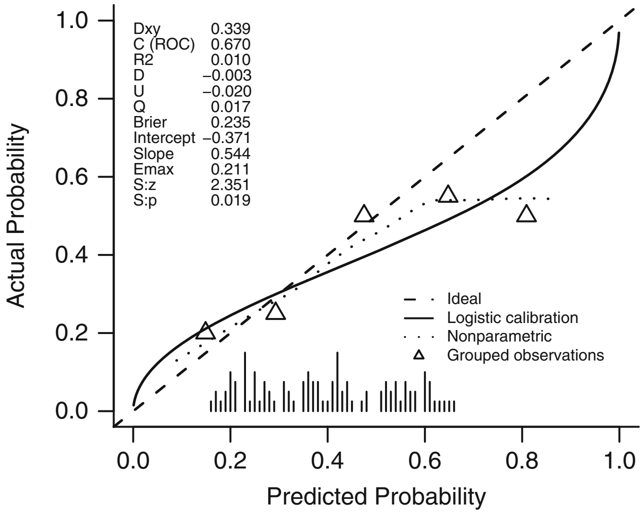

Fig. 10.22 Validation of a logistic model in a test sample of size $n = 100$. The calibrated risk distribution (histogram of logistic-calibrated probabilities) is shown.

one of the predictors is a discrete variable such as sex group. When discriminant analysis assumptions are violated, logistic regression yields more accurate estimates.[251, 514] Even when discriminant analysis is optimal (i.e., when all its assumptions are satisfied) logistic regression is virtually as accurate as the discriminant model.[264]

4 See [573] for a review of measures of effect for binary outcomes.

5 Cepedaet al.[95] found that propensity adjustment is better than covariate adjustment with logistic models when the number of events per variable is less than 8.

6 Pregibon[512] developed a modification of the log likelihood function that when maximized results in a fit that is resistant to overly influential and outlying observations.

7 See Hosmer and Lemeshow[306] for methods of testing for a difference in the observed event proportion and the predicted event probability (average of predicted probabilities) for a group of heterogeneous subjects.

8 See Hosmer and Lemeshow,[305] Kay and Little,[341] and Collett [115, Chap. 5]. Landwehr et al.[373] proposed the partial residual (see also Fowlkes[199]).

9 See Berk and Booth[51] for other partial-like residuals.

10 See [341] for an example comparing a smoothing method with a parametric logistic model fit.

11 See Collett [115, Chap. 5] and Pregibon[512] for more information about influence statistics. Pregibon's resistant estimator of β handles overly influential *groups* of observations and allows one to estimate the weight that an observation contributed to the fit after making the fit robust. Observations receiving low weight are partially ignored but are not deleted.

12 Buyse[86] showed that in the case of a single categorical predictor, the ordinary R^2 has a ready interpretation in terms of variance explained for binary responses. Menard[454] studied various indexes for binary logistic regression. He criticized R^2_{N} for being too dependent on the proportion of observations with $Y = 1$. Hu et al.[309] further studied the properties of variance-based R^2 measures for binary responses. Tjur[613] has a nice discussion discrimination graphics

and sum of squares–based R^2 measures for binary logistic regression, as well as a good discussion of "separation" and infinite regression coefficients. Sums of squares are approximated various ways.

13 Very little work has been done on developing adjusted R^2 measures in logistic regression and other non-linear model setups. Liao and McGee[406] developed one adjusted R^2 measure for binary logistic regression, but it uses simulation to adjust for the bias of overfitting. One might as well use the bootstrap to adjust any of the indexes discussed in this section.

14 [123, 633] have more pertinent discussion of probability accuracy scores.

15 Copas[121] demonstrated how ROC areas can be misleading when applied to different responses having greatly different prevalences. He proposed another approach, the logit rank plot. Newsom[473] is an excellent reference on D_{xy}. Newson[474] developed several generalizations to D_{xy} including a stratified version, and discussed the jackknife variance estimator for them. ROC areas are not very useful for comparing two models[118, 493] (but see[490]).

16 Gneiting and Raftery[219] have an excellent review of proper scoring rules. Hand[253] contains much information about assessing classification accuracy. Mittlböck and Schemper[461] have an excellent review of indexes of explained variation for binary logistic models. See also Korn and Simon[366] and Zheng and Agresti.[684].

17 Pryor et al.[515] presented nomograms for a 10-variable logistic model. One of the variables was sex, which interacted with some of the other variables. Evaluation of predicted probabilities was simplified by the construction of separate nomograms for females and males. Seven terms for discrete predictors were collapsed into one weighted point score axis in the nomograms, and age by risk factor interactions were captured by having four age scales.

18 Moons et al.[462] presents a case study in penalized binary logistic regression modeling.

19 The `rcspline.plot` function in the `Hmisc` R package does not allow for interactions as does `lrm`, but it can provide detailed output for checking spline fits. This function plots the estimated spline regression and confidence limits, placing summary statistics on the graph. If there are no adjustment variables, `rcspline.plot` can also plot two alternative estimates of the regression function: proportions or logit proportions on grouped data, and a nonparametric estimate. The nonparametric regression estimate is based on smoothing the binary responses and taking the logit transformation of the smoothed estimates, if desired. The smoothing uses the "super smoother" of Friedman[207] implemented in the R function `supsmu`.

10.13 Problems

1. Consider the age–sex–response example in Section 10.1.3. This dataset is available from the text's web site in the Datasets area.

   a. Duplicate the analyses done in Section 10.1.3.
   b. For the model containing both age and sex, test H_0 : logit response is linear in age versus H_a : logit response is quadratic in age. Use the best test statistic.
   c. Using a Wald test, test H_0 : no age × sex interaction. Interpret all parameters in the model.

d. Plot the estimated logit response as a function of age and sex, with and without fitting an interaction term.
e. Perform a likelihood ratio test of H_0 : the model containing only age and sex is adequate versus H_a : model is inadequate. Here, "inadequate" may mean nonlinearity (quadratic) in age or presence of an interaction.
f. Assuming no interaction is present, test H_0 : model is linear in age versus H_a : model is nonlinear in age. Allow "nonlinear" to be more general than quadratic. (Hint: use a restricted cubic spline function with knots at age=39, 45, 55, 64 years.)
g. Plot age against the estimated spline transformation of age (the transformation that would make age fit linearly). You can set the sex and intercept terms to anything you choose. Also plot Prob{response = 1 | age, sex} from this fitted restricted cubic spline logistic model.

2. Consider a binary logistic regression model using the following predictors: age (years), sex, race (white, African-American, Hispanic, Oriental, other), blood pressure (mmHg). The fitted model is given by

$$\text{logit Prob}[Y = 1|X] = X\hat{\beta} = -1.36 + .03(\text{race} = \text{African-American})$$
$$- .04(\text{race} = \text{hispanic}) + .05(\text{race} = \text{oriental}) - .06(\text{race} = \text{other})$$
$$+ .07|\text{blood pressure} - 110| + .3(\text{sex} = \text{male}) - .1\text{age} + .002\text{age}^2 +$$
$$(\text{sex} = \text{male})[.05\text{age} - .003\text{age}^2].$$

a. Compute the predicted logit (log odds) that $Y = 1$ for a 50-year-old female Hispanic with a blood pressure of 90 mmHg. Also compute the odds that $Y = 1$ (Prob$[Y = 1]$/Prob$[Y = 0]$) and the estimated probability that $Y = 1$.
b. Estimate odds ratios for each nonwhite race compared with the reference group (white), holding all other predictors constant. Why can you estimate the relative effect of race for all types of subjects without specifying their characteristics?
c. Compute the odds ratio for a blood pressure of 120 mmHg compared with a blood pressure of 105, holding age first to 30 years and then to 40 years.
d. Compute the odds ratio for a blood pressure of 120 mmHg compared with a blood pressure of 105, all other variables held to unspecified constants. Why is this relative effect meaningful without knowing the subject's age, race, or sex?
e. Compute the estimated risk difference in changing blood pressure from 105 mmHg to 120 mmHg, first for age = 30 then for age = 40, for a white female. Why does the risk difference depend on age?
f. Compute the relative odds for males compared with females, for age = 50 and other variables held constant.
g. Same as the previous question but for females : males instead of males : females.
h. Compute the odds ratio resulting from increasing age from 50 to 55 for males, and then for females, other variables held constant. What is wrong with the following question: What is the relative effect of changing age by one year?