Chapter 14

Logistic Regression, Poisson Regression, and Generalized Linear Models

In Chapter 13 we considered nonlinear regression models where the error terms are normally distributed. In this chapter, we take up nonlinear regression models for two important cases where the response outcomes are discrete and the error terms are not normally distributed. First, we consider the logistic nonlinear regression model for use when the response variable is qualitative with two possible outcomes, such as financial status of firm (sound status, headed toward insolvency) or blood pressure status (high blood pressure, not high blood pressure). We then extend this model so that it can be applied when the response variable is a qualitative variable having more than two possible outcomes; for instance, blood pressure status might be classified as high, normal, or low.

Next we take up the Poisson regression model for use when the response variable is a count where large counts are rare events, such as the number of tornadoes in an upper Midwest locality during a year. Finally, we explain that nearly all of the nonlinear regression models discussed in Chapter 13 and in this chapter, as well as the normal error linear models discussed earlier, belong to a family of regression models called generalized linear models.

The nonlinear regression models presented in this chapter are appropriate for analyzing data arising from either observational studies or from experimental studies.

14.1 Regression Models with Binary Response Variable

In a variety of regression applications, the response variable of interest has only two possible qualitative outcomes, and therefore can be represented by a binary indicator variable taking on values 0 and 1.

1. In an analysis of whether or not business firms have an industrial relations department, according to size of firm, the response variable was defined to have the two possible

outcomes: firm has industrial relations department, firm does not have industrial relations department. These outcomes may be coded 1 and 0, respectively (or vice versa).

2. In a study of labor force participation of married women, as a function of age, number of children, and husband's income, the response variable Y was defined to have the two possible outcomes: married woman in labor force, married woman not in labor force. Again, these outcomes may be coded 1 and 0, respectively.

3. In a study of liability insurance possession, according to age of head of household, amount of liquid assets, and type of occupation of head of household, the response variable Y was defined to have the two possible outcomes: household has liability insurance, household does not have liability insurance. These outcomes again may be coded 1 and 0, respectively.

4. In a longitudinal study of coronary heart disease as a function of age, gender, smoking history, cholesterol level, percent of ideal body weight, and blood pressure, the response variable Y was defined to have the two possible outcomes: person developed heart disease during the study, person did not develop heart disease during the study. These outcomes again may be coded 1 and 0, respectively.

These examples show the wide range of applications in which the response variable is binary and hence may be represented by an indicator variable. A binary response variable, taking on the values 0 and 1, is said to involve *binary responses* or *dichotomous responses*. We consider first the meaning of the response function when the outcome variable is binary, and then we take up some special problems that arise with this type of response variable.

Meaning of Response Function when Outcome Variable Is Binary

Consider the simple linear regression model:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \qquad Y_i = 0, 1 \tag{14.1}$$

where the outcome Y_i is binary, taking on the value of either 0 or 1. The expected response $E\{Y_i\}$ has a special meaning in this case. Since $E\{\varepsilon_i\} = 0$ we have:

$$E\{Y_i\} = \beta_0 + \beta_1 X_i \tag{14.2}$$

Consider Y_i to be a Bernoulli random variable for which we can state the probability distribution as follows:

Y_i	Probability
1	$P(Y_i = 1) = \pi_i$
0	$P(Y_i = 0) = 1 - \pi_i$

Thus, π_i is the probability that $Y_i = 1$, and $1 - \pi_i$ is the probability that $Y_i = 0$. By the definition of expected value of a random variable in (A.12), we obtain:

$$E\{Y_i\} = 1(\pi_i) + 0(1 - \pi_i) = \pi_i = P(Y_i = 1) \tag{14.3}$$

Equating (14.2) and (14.3), we thus find:

$$E\{Y_i\} = \beta_0 + \beta_1 X_i = \pi_i \tag{14.4}$$

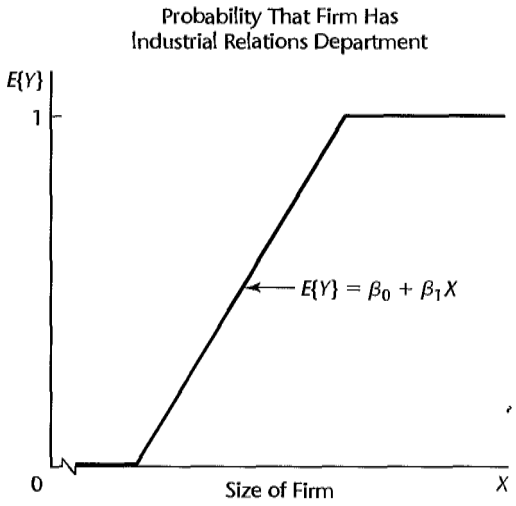


FIGURE 14.1
Illustration of Response Function when Response Variable Is Binary—Industrial Relations Department Example.

The mean response $E\{Y_i\} = \beta_0 + \beta_1 X_i$ as given by the response function is therefore simply the probability that $Y_i = 1$ when the level of the predictor variable is X_i. This interpretation of the mean response applies whether the response function is a simple linear one, as here, or a complex multiple regression one. The mean response, when the outcome variable is a 0, 1 indicator variable, always represents the probability that $Y = 1$ for the given levels of the predictor variables. Figure 14.1 illustrates a simple linear response function for an indicator outcome variable. Here, the indicator variable Y refers to whether or not a firm has an industrial relations department, and the predictor variable X is size of firm. The response function in Figure 14.1 shows the probability that firms of given size have an industrial relations department.

Special Problems when Response Variable Is Binary

Special problems arise, unfortunately, when the response variable is an indicator variable. We consider three of these now, using a simple linear regression model as an illustration.

1. *Nonnormal Error Terms.* For a binary 0, 1 response variable, each error term $\varepsilon_i = Y_i - (\beta_0 + \beta_1 X_i)$ can take on only two values:

$$\text{When } Y_i = 1: \quad \varepsilon_i = 1 - \beta_0 - \beta_1 X_i \tag{14.5a}$$

$$\text{When } Y_i = 0: \quad \varepsilon_i = -\beta_0 - \beta_1 X_i \tag{14.5b}$$

Clearly, normal error regression model (2.1), which assumes that the ε_i are normally distributed, is not appropriate.

2. *Nonconstant Error Variance.* Another problem with the error terms ε_i is that they do not have equal variances when the response variable is an indicator variable. To see this, we shall obtain $\sigma^2\{Y_i\}$ for the simple linear regression model (14.1), utilizing (A.15):

$$\sigma^2\{Y_i\} = E\{(Y_i - E\{Y_i\})^2\} = (1 - \pi_i)^2 \pi_i + (0 - \pi_i)^2 (1 - \pi_i)$$

or:

$$\sigma^2\{Y_i\} = \pi_i(1 - \pi_i) = (E\{Y_i\})(1 - E\{Y_i\}) \tag{14.6}$$

The variance of ε_i is the same as that of Y_i because $\varepsilon_i = Y_i - \pi_i$ and π_i is a constant:

$$\sigma^2\{\varepsilon_i\} = \pi_i(1 - \pi_i) = (E\{Y_i\})(1 - E\{Y_i\}) \tag{14.7}$$

or:

$$\sigma^2\{\varepsilon_i\} = (\beta_0 + \beta_1 X_i)(1 - \beta_0 - \beta_1 X_i) \tag{14.7a}$$

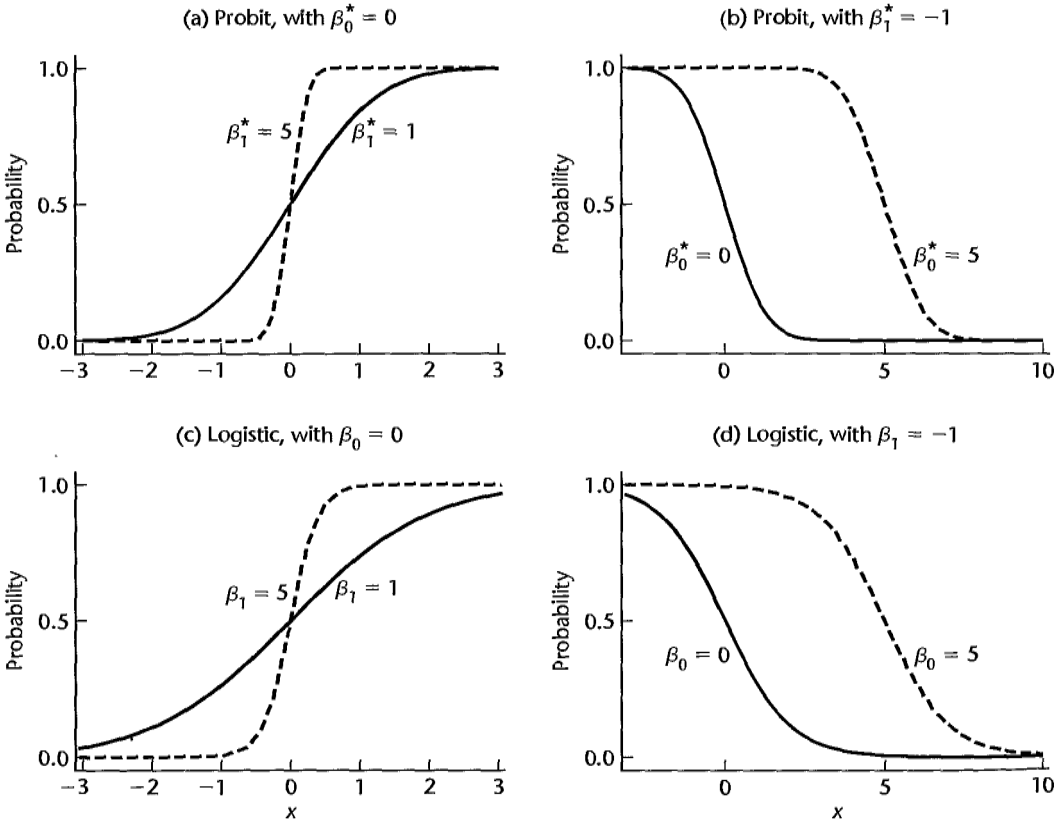
Note from (14.7a) that $\sigma^2\{\varepsilon_i\}$ depends on X_i. Hence, the error variances will differ at different levels of X, and ordinary least squares will no longer be optimal.

3. *Constraints on Response Function.* Since the response function represents probabilities when the outcome variable is a 0, 1 indicator variable, the mean responses should be constrained as follows:

$$0 \leq E\{Y\} = \pi \leq 1 \tag{14.8}$$

Many response functions do not automatically possess this constraint. A linear response function, for instance, may fall outside the constraint limits within the range of the predictor variable in the scope of the model.

FIGURE 14.2 Examples of Probit and Logistic Mean Response Functions.



The difficulties created by the need for the restriction in (14.8) on the response function are the most serious. One could use weighted least squares to handle the problem of unequal error variances. In addition, with large sample sizes the method of least squares provides estimators that are asymptotically normal under quite general conditions, even if the distribution of the error terms is far from normal. However, the constraint on the mean responses to fall between 0 and 1 frequently will rule out a linear response function. In the industrial relations department example, for instance, use of a linear response function subject to the constraints on the mean response might require a probability of 0 for the mean response for all small firms and a probability of 1 for the mean response for all large firms, as illustrated in Figure 14.1. Such a model would often be considered unreasonable. Instead, a model where the probabilities 0 and 1 are reached asymptotically, as illustrated by each of the S-shaped curves in Figure 14.2, would usually be more appropriate.

14.2 Sigmoidal Response Functions for Binary Responses

In this section, we introduce three response functions for modeling binary responses. These functions are bounded between 0 and 1, have a characteristic *sigmoidal-* or *S*-shape, and approach 0 and 1 asymptotically. These functions arise naturally when the binary response variable results from a zero-one recoding (or dichotomization) of an underlying continuous response variable, and they are often appropriate for discrete binary responses as well.

Probit Mean Response Function

Consider a health researcher studying the effect of a mother's use of alcohol (X—an index of degree of alcohol use during pregnancy) on the duration of her pregnancy (Y^c). Here we use the superscript c to emphasize that the response variable, pregnancy duration, is a continuous response. This can be represented by a simple linear regression model:

$$Y_i^c = \beta_0^c + \beta_1^c X_i + \varepsilon_i^c \tag{14.9}$$

and we will assume that ε_i^c is normally distributed with mean zero and variance σ_c^2.

If the continuous response variable, pregnancy duration, were available, we might proceed with the usual simple linear regression analysis. However, in this instance, researchers coded each pregnancy duration as preterm or full term using the following rule:

$$Y_i = \begin{cases} 1 & \text{if } Y_i^c \le 38 \text{ weeks (preterm)} \\ 0 & \text{if } Y_i^c > 38 \text{ weeks (full term)} \end{cases}$$

It follows from (14.3) and (14.9) that:

$$P(Y_i = 1) = \pi_i = P\left(Y_i^c \le 38\right) \tag{14.10a}$$

$$= P\left(\beta_0^c + \beta_1^c X_i + \varepsilon_i^c \le 38\right) \tag{14.10b}$$

$$= P\left(\varepsilon_i^c \le 38 - \beta_0^c - \beta_1^c X_i\right) \tag{14.10c}$$

$$= P\left(\frac{\varepsilon_i^c}{\sigma_c} \le \frac{38 - \beta_0^c}{\sigma_c} - \frac{\beta_1^c}{\sigma_c} X_i\right) \tag{14.10d}$$

$$= P(Z \le \beta_0^* + \beta_1^* X_i) \tag{14.10e}$$

where $\beta_0^* = (38 - \beta_0^c)/\sigma_c$, $\beta_1^* = -\beta_1^c/\sigma_c$, and $Z = \varepsilon_i^c/\sigma_c$ follows a standard normal distribution. If we let $P(Z \leq z) = \Phi(z)$, we have, from (14.10a–e):

$$P(Y_i = 1) = \Phi(\beta_0^* + \beta_1^* X_i) \tag{14.11}$$

Equations (14.3) and (14.11) together yield the nonlinear regression function known as the *probit mean response function:*

$$E\{Y_i\} = \pi_i = \Phi(\beta_0^* + \beta_1^* X_i) \tag{14.12}$$

The inverse function, Φ^{-1}, of the standard normal cumulative distribution function Φ, is sometimes called the probit transformation. We solve for the linear predictor, $\beta_0^* + \beta_1^* X_i$ in (14.12) by applying the probit transformation to both sides of the expression, obtaining:

$$\Phi^{-1}(\pi_i) = \pi_i' = \beta_0^* + \beta_1^* X_i \tag{14.13}$$

The resulting expression, $\pi_i' = \beta_0^* + \beta_1^* X_i$, is called the *probit response function*, or more generally, the *linear predictor.*

Plots of the probit mean response function (14.12) for various values of β_0^* and β_1^* are shown in Figures 14.2a and 14.2b. Some characteristics of this response function are:

1. The probit mean response function is bounded between 0 and 1, and it approaches these limits asymptotically.
2. As β_1^* increases (for $\beta_1^* > 0$), the mean function becomes more S-shaped, changing more rapidly in the center. Figure 14.2a shows two probit mean response functions, where both intercept coefficients are 0, and the slope coefficients are 1 and 5. Notice that the curve has a more pronounced S-shape with $\beta_1^* = 5$.
3. Changing the sign of β_1^* from positive to negative changes the mean response function from a monotone increasing function to a monotone decreasing function. The probit mean response functions plotted in Figure 14.2a have positive slope coefficients while those in Figure 14.2b have negative slope coefficients.
4. Increasing or decreasing the intercept β_0^* shifts the mean response function horizontally. (The direction of the shift depends on the signs of both β_0^* and β_1^*.) Figure 14.2b shows two probit mean response functions, where both slope coefficients are -1, and the intercept coefficients are 0 and 5. Notice that the curve has shifted to the right as β_0^* changes from 0 to 5.
5. Finally, we note the following *symmetry property* of the probit response function. If the response variable is recoded using $Y_i' = 1 - Y_i$, that is, by changing the 1s to 0s and the 0s to 1s—the signs of all of the coefficients are reversed. This follows easily from the symmetry of the standard normal distribution: since $\Phi(Z) = 1 - \Phi(-Z)$, it follows that $P(Y_i' = 1) = P(Y_i = 0) = 1 - \Phi(\beta_0^* + \beta_1^* X_i) = \Phi(-\beta_0^* - \beta_1^* X_i)$.

Logistic Mean Response Function

We have seen that the assumption of normally distributed errors for the underlying continuous response variable in (14.9) led to the use of the standard normal cumulative distribution function, Φ, to model π_i. An alternative error distribution that is very similar to the normal distribution is the logistic distribution. Figure 14.3 presents plots of the standard normal density function and the logistic density function, each with mean zero and variance one. The plots are nearly indistinguishable, although the logistic distribution has slightly heavier

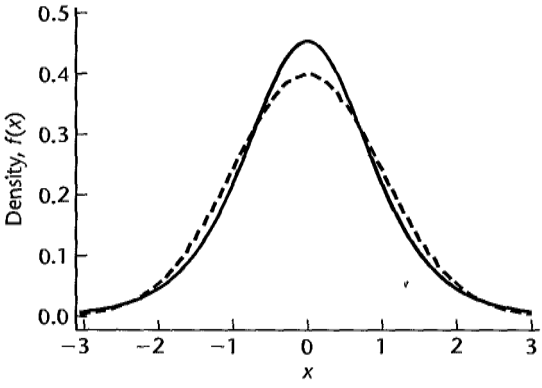


FIGURE 14.3 Plots of Normal Density (dashed line) and Logistic Density (solid line), Each Having Mean 0 and Variance 1.

tails. The density of a logistic random variable ε_L having mean zero and standard deviation $\sigma = \pi/\sqrt{3}$ has a simple form:

$$f_L(\varepsilon_L) = \frac{\exp(\varepsilon_L)}{[1 + \exp(\varepsilon_L)]^2} \tag{14.14a}$$

Its cumulative distribution function is:

$$F_L(\varepsilon_L) = \frac{\exp(\varepsilon_L)}{1 + \exp(\varepsilon_L)} \tag{14.14b}$$

Suppose now that ε_i^c in (14.9) has a logistic distribution with mean zero and standard deviation σ_c. Then, from (14.10d) we have:

$$P(Y_i = 1) = P\left(\frac{\varepsilon_i^c}{\sigma_c} \leq \beta_0^* + \beta_1^* X_i\right)$$

where ε_i^c/σ_c follows a logistic distribution with mean zero and standard deviation one. Multiplying both sides of the inequality inside the probability statement on the right by $\pi/\sqrt{3}$ does not change the probability; therefore:

$$P(Y_i = 1) = \pi_i = P\left(\frac{\pi}{\sqrt{3}}\frac{\varepsilon_i^c}{\sigma_c} \leq \frac{\pi}{\sqrt{3}}\beta_0^* + \frac{\pi}{\sqrt{3}}\beta_1^* X_i\right) \tag{14.15a}$$

$$= P(\varepsilon_L \leq \beta_0 + \beta_1 X_i) \tag{14.15b}$$

$$= F_L(\beta_0 + \beta_1 X_i) \tag{14.15c}$$

$$= \frac{\exp(\beta_0 + \beta_1 X_i)}{1 + \exp(\beta_0 + \beta_1 X_i)} \tag{14.15d}$$

where $\beta_0 = (\pi/\sqrt{3})\beta_0^*$ and $\beta_1 = (\pi/\sqrt{3})\beta_1^*$ denote the logistic regression parameters. To summarize, the *logistic mean response function* is:

$$E\{Y_i\} = \pi_i = F_L(\beta_0 + \beta_1 X_i) = \frac{\exp(\beta_0 + \beta_1 X_i)}{1 + \exp(\beta_0 + \beta_1 X_i)} \tag{14.16}$$

Straightforward algebra shows that an equivalent form of (14.16) is given by:

$$E\{Y_i\} = \pi_i = [1 + \exp(-\beta_0 - \beta_1 X_i)]^{-1} \tag{14.17}$$

Applying the inverse of the cumulative distribution function F_L to the two middle terms in (14.16) yields:

$$F_L^{-1}(\pi_i) = \beta_0 + \beta_1 X_i = \pi_i' \tag{14.18}$$

The transformation $F_L^{-1}(\pi_i)$ is called the *logit transformation of the probability* π_i, and is given by:

$$F_L^{-1}(\pi_i) = \log_e\left(\frac{\pi_i}{1-\pi_i}\right) \tag{14.18a}$$

where the ratio $\pi_i/(1-\pi_i)$ in (14.18a) is called the *odds*. The linear predictor in (14.18) is referred to as the *logit response function.*

Figures 14.2c and 14.2d each show two logistic mean response functions, where the parameters correspond to those in Figures 14.2a and 14.2b for the probit mean response function. It is clear from the plots that these logistic mean response functions are qualitatively similar to the corresponding probit mean response functions. The five properties of the probit mean response function, listed earlier, are also true for the logistic mean response function. The observed differences in logistic and probit mean response functions are largely due to the differences in the scaling of the parameters mentioned previously. Note that the symmetry property for the probit mean response function also holds for the logistic mean response function.

Complementary Log-Log Response Function

A third mean response function is sometimes used when the error distribution of ε^c is not symmetric. The density function $f_G(\varepsilon)$ of the *extreme value* or *Gumbel* probability distribution having mean zero and variance one is shown in Figure 14.4, along with the comparable standard normal and logistic densities discussed earlier. Notice that this density is skewed to the right and clearly distinct from the standard normal and logistic densities. It can be shown that use of the Gumbel error distribution for ε^c in (14.9) leads to the mean response function:

$$\pi_i = 1 - \exp\left(-\exp\left(\beta_0^G + \beta_1^G X_i\right)\right) \tag{14.19}$$

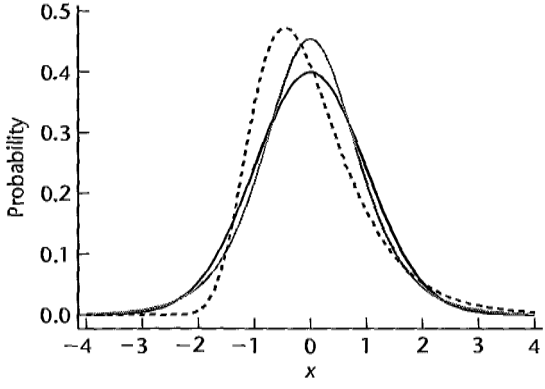


FIGURE 14.4 Plots of Gumbel (dashed line), Normal (black line), and Logistic (gray line) Density Functions, Each Having Mean 0 and Variance 1.

Solving for the linear predictor $\beta_0^G + \beta_1^G X_i$, we obtain the *complementary log-log* response model:

$$\pi_i' = \log[-\log(1 - \pi(X_i))] = \beta_0^G + \beta_1^G X_i \qquad \textbf{(14.19a)}$$

The symmetry property discussed on page 560 for the logit and probit models does not hold for (14.19).

For the remainder of this chapter, we focus on the use of the logistic mean response function. This is currently the most widely used model for two reasons: (1) we shall see that the regression parameters have relatively simple and useful interpretations, and (2) statistical software is widely available for analysis of logistic regression models. In the next two sections we consider in detail the fitting of simple and multiple logistic regression models to binary data.

Comment

Our development of the logistic and probit mean response functions assumed that the binary response Y_i was obtained from an explicit dichotomization of an observed continuous response Y_i^c, but this is not required. These response functions often work well for binary responses that do not arise from such a dichotomization. In addition, binary responses frequently can be interpreted as having arisen from a dichotomization of an unobserved, or latent, continuous response. ■

14.3 Simple Logistic Regression

We shall use the method of maximum likelihood to estimate the parameters of the logistic response function. This method is well suited to deal with the problems associated with the responses Y_i being binary. As explained in Section 1.8, we first need to develop the joint probability function of the sample observations. Instead of using the normal distribution for the Y observations as was done earlier in (1.26), we now need to utilize the Bernoulli distribution for a binary random variable.

Simple Logistic Regression Model

First, we require a formal statement of the simple logistic regression model. Recall that when the response variable is binary, taking on the values 1 and 0 with probabilities π and $1 - \pi$, respectively, Y is a Bernoulli random variable with parameter $E\{Y\} = \pi$. We could state the simple logistic regression model in the usual form:

$$Y_i = E\{Y_i\} + \varepsilon_i$$

Since the distribution of the error term ε_i depends on the Bernoulli distribution of the response Y_i, it is preferable to state the simple logistic regression model in the following fashion:

Y_i are independent Bernoulli random variables with expected values $E\{Y_i\} = \pi_i$, where:

$$E\{Y_i\} = \pi_i = \frac{\exp(\beta_0 + \beta_1 X_i)}{1 + \exp(\beta_0 + \beta_1 X_i)} \qquad \textbf{(14.20)}$$

The X observations are assumed to be known constants. Alternatively, if the X observations are random, $E\{Y_i\}$ is viewed as a conditional mean, given the value of X_i.

Likelihood Function

Since each Y_i observation is an ordinary Bernoulli random variable, where:

$$P(Y_i = 1) = \pi_i$$
$$P(Y_i = 0) = 1 - \pi_i$$

we can represent its probability distribution as follows:

$$f_i(Y_i) = \pi_i^{Y_i}(1 - \pi_i)^{1-Y_i} \qquad Y_i = 0, 1; \quad i = 1, \ldots, n \tag{14.21}$$

Note that $f_i(1) = \pi_i$ and $f_i(0) = 1 - \pi_i$. Hence, $f_i(Y_i)$ simply represents the probability that $Y_i = 1$ or 0.

Since the Y_i observations are independent, their joint probability function is:

$$g(Y_1, \ldots, Y_n) = \prod_{i=1}^{n} f_i(Y_i) = \prod_{i=1}^{n} \pi_i^{Y_i}(1 - \pi_i)^{1-Y_i} \tag{14.22}$$

Again, it will be easier to find the maximum likelihood estimates by working with the logarithm of the joint probability function:

$$\begin{aligned} \log_e g(Y_1, \ldots, Y_n) &= \log_e \prod_{i=1}^{n} \pi_i^{Y_i}(1 - \pi_i)^{1-Y_i} \\ &= \sum_{i=1}^{n} [Y_i \log_e \pi_i + (1 - Y_i) \log_e (1 - \pi_i)] \\ &= \sum_{i=1}^{n} \left[Y_i \log_e \left(\frac{\pi_i}{1 - \pi_i} \right) \right] + \sum_{i=1}^{n} \log_e (1 - \pi_i) \end{aligned} \tag{14.23}$$

Since $E\{Y_i\} = \pi_i$ for a binary variable, it follows from (14.16) that:

$$1 - \pi_i = [1 + \exp(\beta_0 + \beta_1 X_i)]^{-1} \tag{14.24}$$

Furthermore, from (14.18a), we obtain:

$$\log_e \left(\frac{\pi_i}{1 - \pi_i} \right) = \beta_0 + \beta_1 X_i \tag{14.25}$$

Hence, (14.23) can be expressed as follows:

$$\log_e L(\beta_0, \beta_1) = \sum_{i=1}^{n} Y_i(\beta_0 + \beta_1 X_i) - \sum_{i=1}^{n} \log_e [1 + \exp(\beta_0 + \beta_1 X_i)] \tag{14.26}$$

where $L(\beta_0, \beta_1)$ replaces $g(Y_1, \ldots, Y_n)$ to show explicitly that we now view this function as the likelihood function of the parameters to be estimated, given the sample observations.

Maximum Likelihood Estimation

The maximum likelihood estimates of β_0 and β_1 in the simple logistic regression model are those values of β_0 and β_1 that maximize the log-likelihood function in (14.26). No closed-form solution exists for the values of β_0 and β_1 in (14.26) that maximize the log-likelihood function. Computer-intensive numerical search procedures are therefore required

to find the maximum likelihood estimates b_0 and b_1. There are several widely used numerical search procedures; one of these employs iteratively reweighted least squares, which we shall explain in Section 14.4. Reference 14.1 provides a discussion of several numerical search procedures for finding maximum likelihood estimates. We shall rely on standard statistical software programs specifically designed for logistic regression to obtain the maximum likelihood estimates b_0 and b_1.

Once the maximum likelihood estimates b_0 and b_1 are found, we substitute these values into the response function in (14.20) to obtain the fitted response function. We shall use $\hat{\pi}_i$ to denote the fitted value for the ith case:

$$\hat{\pi}_i = \frac{\exp(b_0 + b_1 X_i)}{1 + \exp(b_0 + b_1 X_i)} \tag{14.27}$$

The fitted logistic response function is as follows:

$$\hat{\pi} = \frac{\exp(b_0 + b_1 X)}{1 + \exp(b_0 + b_1 X)} \tag{14.28}$$

If we utilize the logit transformation in (14.18), we can express the fitted response function in (14.28) as follows:

$$\hat{\pi}' = b_0 + b_1 X \tag{14.29}$$

where:

$$\hat{\pi}' = \log_e\left(\frac{\hat{\pi}}{1 - \hat{\pi}}\right) \tag{14.29a}$$

We call (14.29) the *fitted logit response function.*

Once the fitted logistic response function has been obtained, the usual next steps are to examine the appropriateness of the fitted response function and, if the fit is good, to make a variety of inferences and predictions. We shall postpone a discussion of how to examine the goodness of fit of a logistic response function and how to make inferences and predictions until we have considered the multiple logistic regression model with a number of predictor variables.

Example

A systems analyst studied the effect of computer programming experience on ability to complete within a specified time a complex programming task, including debugging. Twenty-five persons were selected for the study. They had varying amounts of programming experience (measured in months of experience), as shown in Table 14.1a, column 1. All persons were given the same programming task, and the results of their success in the task are shown in column 2. The results are coded in binary fashion: $Y = 1$ if the task was completed successfully in the allotted time, and $Y = 0$ if the task was not completed successfully. Figure 14.5 contains a scatter plot of the data. This plot is not too informative because of the nature of the response variable, other than to indicate that ability to complete the task successfully appears to increase with amount of experience. A lowess nonparametric response curve was fitted to the data and is also shown in Figure 14.5. A sigmoidal S-shaped response function is clearly suggested by the nonparametric lowess fit. It was therefore decided to fit the logistic regression model (14.20).

A standard logistic regression package was run on the data. The results are contained in Table 14.1b. Since $b_0 = -3.0597$ and $b_1 = .1615$, the estimated logistic regression

TABLE 14.1 **Data and Maximum Likelihood Estimates—Programming Task Example.**

(a) Data

Person i	(1) Months of Experience X_i	(2) Task Success Y_i	(3) Fitted Value $\hat{\pi}_i$
1	14	0	.310
2	29	0	.835
3	6	0	.110
...	...	...	...
23	28	1	.812
24	22	1	.621
25	8	1	.146

(b) Maximum Likelihood Estimates

Regression Coefficient	Estimated Regression Coefficient	Estimated Standard Deviation
β_0	−3.0597	1.259
β_1	.1615	.0650

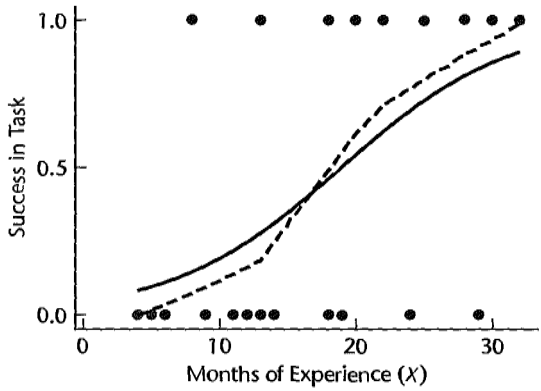


FIGURE 14.5 **Scatter Plot, Lowess Curve (dashed line), and Estimated Logistic Mean Response Function (solid line)—Programming Task Example.**

function (14.28) is:

$$\hat{\pi} = \frac{\exp(-3.0597 + .1}{1 + \exp(-3.0597 +}$$

The fitted values are given in Table 14.1a, column response for $i = 1$, where $X_1 = 14$, is:

$$\hat{\pi}_1 = \frac{\exp[-3.0597 + .1615}{1 + \exp[-3.0597 + .16}$$

This fitted value is the estimated probability that a person with 14 months experience will successfully complete the programming task. In addition to the lowess fit, Figure 14.5 also contains a plot of the fitted logistic response function, $\hat{\pi}(x)$.

Interpretation of b_1

The interpretation of the estimated regression coefficient b_1 in the fitted logistic response function (14.30) is not the straightforward interpretation of the slope in a linear regression model. The reason is that the effect of a unit increase in X varies for the logistic regression model according to the location of the starting point on the X scale. An interpretation of b_1 is found in the property of the fitted logistic function that the estimated odds $\hat{\pi}/(1-\hat{\pi})$ are multiplied by $\exp(b_1)$ for any unit increase in X.

To see this, we consider the value of the fitted logit response function (14.29) at $X = X_j$:

$$\hat{\pi}'(X_j) = b_0 + b_1 X_j$$

The notation $\hat{\pi}'(X_j)$ indicates specifically the X level associated with the fitted value. We also consider the value of the fitted logit response function at $X = X_j + 1$:

$$\hat{\pi}'(X_j + 1) = b_0 + b_1(X_j + 1)$$

The difference between the two fitted values is simply:

$$\hat{\pi}'(X_j + 1) - \hat{\pi}'(X_j) = b_1$$

Now according to (14.29a), $\hat{\pi}'(X_j)$ is the logarithm of the estimated odds when $X = X_j$; we shall denote it by $\log_e(\text{odds}_1)$. Similarly, $\hat{\pi}'(X_j + 1)$ is the logarithm of the estimated odds when $X = X_j + 1$; we shall denote it by $\log_e(\text{odds}_2)$. Hence, the difference between the two fitted logit response values can be expressed as follows:

$$\log_e(\text{odds}_2) - \log_e(\text{odds}_1) = \log_e\left(\frac{\text{odds}_2}{\text{odds}_1}\right) = b_1$$

Taking antilogs of each side, we see that the estimated ratio of the odds, called the *odds ratio* and denoted by $\widehat{OR}$, equals $\exp(b_1)$:

$$\widehat{OR} = \frac{\text{odds}_2}{\text{odds}_1} = \exp(b_1) \qquad \textbf{(14.31)}$$

mple For the programming task example, we see from Figure 14.5 that the probability of success increases sharply with experience. Specifically, Table 14.1b shows that the odds ratio is $\widehat{OR} = \exp(b_1) = \exp(.1615) = 1.175$, so that the odds of completing the task increase by 17.5 percent with each additional month of experience.

Since a unit increase of one month is quite small, the estimated odds ratio of 1.175 may not adequately show the change in odds for a longer difference in time. In general, the estimated odds ratio when there is a difference of c units of X is $\exp(cb_1)$. For example, should we wish to compare individuals with relatively little experience to those with extensive experience, say 10 months versus 25 months so that $c = 15$, then the odds ratio would be estimated to be $\exp[15(.1615)] = 11.3$. This indicates that the odds of completing the task increase over 11-fold for experienced persons compared to relatively inexperienced persons.

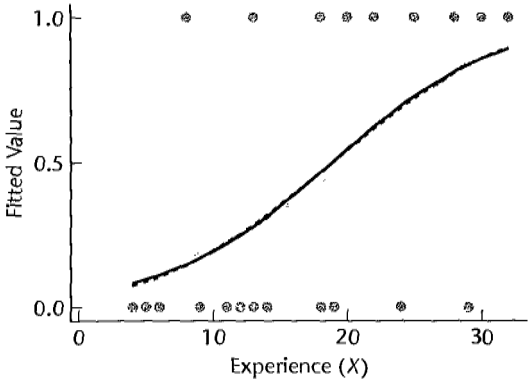


FIGURE 14.6 Logistic (solid line), Probit (dashed line), and Complementary Log-Log (gray line) Fits—Programming Task Example.

Comment

The odds ratio interpretation of the estimated regression coefficient b_1 makes the logistic regression model especially attractive for modeling and interpreting epidemiologic studies. ■

Use of Probit and Complementary Log-Log Response Functions

As we discussed earlier in Section 14.2, alternative sigmoidal shaped response functions, such as the probit or complementary log-log functions, can be utilized as well. For example, it is interesting to fit the programming task data in Table 14.1 to these alternative response functions. Figure 14.6 shows the scatter plot of the data and the fitted logistic, probit, and complementary log-log mean response functions. The logistic and probit fits are very similar, whereas the complementary log-log fit differs slightly, having a less pronounced S-shape.

Repeat Observations—Binomial Outcomes

In some cases, particularly for designed experiments, a number of repeat observations are obtained at several levels of the predictor variable X. For instance, a pricing experiment involved showing a new product to 1,000 consumers, providing information about it, and then asking each consumer whether he or she would buy the product at a given price. Five prices were studied, and 200 persons were randomly selected for each price level. The response variable here is binary (would purchase, would not purchase); the predictor variable is price and has five levels.

When repeat observations are present, the log-likelihood function in (14.26) can be simplified. We shall adopt the notation used for replicate observations in our discussion of the F test for lack of fit in Section 3.7. We denote the X levels at which repeat observations are obtained by $X_1, \ldots X_c$ and we assume that there are n_j binary responses at level X_j. Then the observed value of the ith binary response at X_j is denoted by Y_{ij}, where $i = 1, \ldots, n_j$, and $j = 1, \ldots, c$. The number of 1s at level X_j is denoted by $Y_{\cdot j}$:

$$Y_{\cdot j} = \sum_{i=1}^{n_j} Y_{ij} \tag{14.32a}$$

and the proportion of 1s at level X_j is denoted by p_j:

$$p_j = \frac{Y_{.j}}{n_j} \tag{14.32b}$$

The random variable $Y_{.j}$ has a *binomial distribution* given by:

$$f(Y_{.j}) = \binom{n_j}{Y_{.j}} \pi_j^{Y_{.j}} (1-\pi_j)^{n_j - Y_{.j}} \tag{14.33}$$

where:

$$\binom{n_j}{Y_{.j}} = \frac{n_j!}{(Y_{.j})!(n_j - Y_{.j})!}$$

and the factorial notation $a!$ represents $a(a-1)(a-2)\cdots 1$. The binomial random variable $Y_{.j}$ has mean $n_j\pi_j$ and variance $n_j\pi_j(1-\pi_j)$. The log-likelihood function then can be stated as follows:

$$\log_e L(\beta_0, \beta_1) = \sum_{j=1}^{c} \left\{ \log_e \binom{n_j}{Y_{.j}} + Y_{.j}(\beta_0 + \beta_1 X_j) - n_j \log_e[1 + \exp(\beta_0 + \beta_1 X_j)] \right\} \tag{14.34}$$

Example

In a study of the effectiveness of coupons offering a price reduction on a given product, 1,000 homes were selected at random. A packet containing advertising material and a coupon for the product were mailed to each home. The coupons offered different price reductions (5, 10, 15, 20, and 30 dollars), and 200 homes were assigned at random to each of the price reduction categories. The predictor variable X in this study is the amount of price reduction, and the response variable Y is a binary variable indicating whether or not the coupon was redeemed within a six-month period.

Table 14.2 contains the data for this study. X_j denotes the price reduction offered by a coupon, n_j the number of households that received a coupon with price reduction X_j, $Y_{.j}$ the number of these households that redeemed the coupon, and p_j the proportion of households receiving a coupon with price reduction X_j that redeemed the coupon. The logistic regression model (14.20) was fitted by a logistic regression package and the fitted

TABLE 14.2 Data—Coupon Effectiveness Example.

Level j	(1) Price Reduction X_j	(2) Number of Households n_j	(3) Number of Coupons Redeemed $Y_{.j}$	(4) Proportion of Coupons Redeemed p_j	(5) Model-Based Estimate $\hat{\pi}_j$
1	5	200	30	.150	.1736
2	10	200	55	.275	.2543
3	15	200	70	.350	.3562
4	20	200	100	.500	.4731
5	30	200	137	.685	.7028

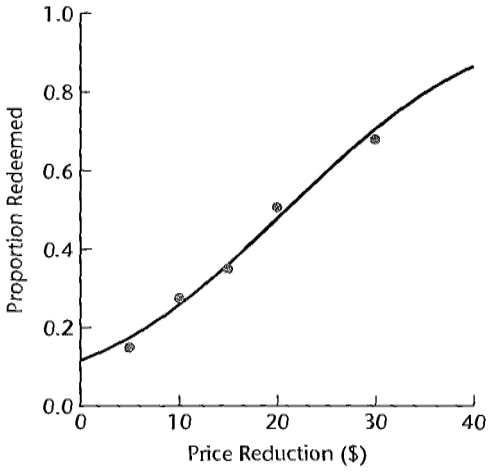


FIGURE 14.7
Plot of Proportions of Coupons Redeemed and Fitted Logistic Response Function—Coupon Effectiveness Example.

response function was found to be:

$$\hat{\pi} = \frac{\exp(-2.04435 + .096834X)}{1 + \exp(-2.04435 + .096834X)} \tag{14.35}$$

Fitted values are given in column 5 of Table 14.2. Figure 14.7 shows the fitted response function, as well as the proportions of coupons redeemed at each of the X_j levels. The logistic response function appears to provide a very good fit. The odds ratio here is:

$$\widehat{OR} = \exp(b_1) = \exp(.096834) = 1.102$$

Hence, the odds of a coupon being redeemed are estimated to increase by 10.2 percent with each one dollar increase in the coupon value, that is, with each one dollar reduction in price.

14.4 Multiple Logistic Regression

Multiple Logistic Regression Model

The simple logistic regression model (14.20) is easily extended to more than one predictor variable. In fact, several predictor variables are usually required with logistic regression to obtain adequate description and useful predictions.

In extending the simple logistic regression model, we simply replace $\beta_0 + \beta_1 X$ in (14.16) by $\beta_0 + \beta_1 X_1 + \cdots + \beta_{p-1} X_{p-1}$. To simplify the formulas, we shall use matrix notation and the following three vectors:

$$\underset{p\times 1}{\boldsymbol{\beta}} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{bmatrix} \qquad \underset{p\times 1}{\mathbf{X}} = \begin{bmatrix} 1 \\ X_1 \\ X_2 \\ \vdots \\ X_{p-1} \end{bmatrix} \qquad \underset{p\times 1}{\mathbf{X}_i} = \begin{bmatrix} 1 \\ X_{i1} \\ X_{i2} \\ \vdots \\ X_{i,p-1} \end{bmatrix} \tag{14.36}$$

We then have:

$$\mathbf{X}'\boldsymbol{\beta} = \beta_0 + \beta_1 X_1 + \cdots + \beta_{p-1} X_{p-1} \tag{14.37a}$$

$$\mathbf{X}_i'\boldsymbol{\beta} = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_{p-1} X_{i,p-1} \tag{14.37b}$$

With this notation, the simple logistic response function (14.20) extends to the multiple logistic response function as follows:

$$E\{Y\} = \frac{\exp(\mathbf{X}'\boldsymbol{\beta})}{1 + \exp(\mathbf{X}'\boldsymbol{\beta})} \tag{14.38}$$

and the equivalent simple logistic response form (14.17) extends to:

$$E\{Y\} = [1 + \exp(-\mathbf{X}'\boldsymbol{\beta})]^{-1} \tag{14.38a}$$

Similarly, the logit transformation (14.18a):

$$\pi' = \log_e\left(\frac{\pi}{1-\pi}\right) \tag{14.39}$$

now leads to the logit response function, or linear predictor:

$$\pi' = \mathbf{X}'\boldsymbol{\beta} \tag{14.40}$$

The multiple logistic regression model can therefore be stated as follows:

Y_i are independent Bernoulli random variables with expected values $E\{Y_i\} = \pi_i$, where:

$$E\{Y_i\} = \pi_i = \frac{\exp(\mathbf{X}_i'\boldsymbol{\beta})}{1 + \exp(\mathbf{X}_i'\boldsymbol{\beta})} \tag{14.41}$$

Again, the X observations are considered to be known constants. Alternatively, if the X variables are random, $E\{Y_i\}$ is viewed as a conditional mean, given the values of $X_{i1}, \ldots, X_{i,p-1}$.

Like the simple logistic response function (14.16), the multiple logistic response function (14.41) is monotonic and sigmoidal in shape with respect to $\mathbf{X}'\boldsymbol{\beta}$ and is almost linear when π is between .2 and .8. The X variables may be different predictor variables, or some may represent curvature and/or interaction effects. Also, the predictor variables may be quantitative, or they may be qualitative and represented by indicator variables. This flexibility makes the multiple logistic regression model very attractive.

Comment

When the logistic regression model contains only qualitative variables, it is often referred to as a log-linear model. See Reference 14.2 for an in-depth discussion of the analysis of log-linear models. ■

Fitting of Model

Again, we shall utilize the method of maximum likelihood to estimate the parameters of the multiple logistic response function (14.41). The log-likelihood function for simple logistic regression in (14.26) extends directly for multiple logistic regression:

$$\log_e L(\boldsymbol{\beta}) = \sum_{i=1}^{n} Y_i(\mathbf{X}_i'\boldsymbol{\beta}) - \sum_{i=1}^{n} \log_e[1 + \exp(\mathbf{X}_i'\boldsymbol{\beta})] \tag{14.42}$$

Numerical search procedures are used to find the values of $\beta_0, \beta_1, \ldots, \beta_{p-1}$ that maximize $\log_e L(\boldsymbol{\beta})$. These maximum likelihood estimates will be denoted by $b_0, b_1, \ldots, b_{p-1}$. Let $\mathbf{b}$ denote the vector of the maximum likelihood estimates:

$$\underset{p\times 1}{\mathbf{b}} = \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_{p-1} \end{bmatrix} \tag{14.43}$$

The fitted logistic response function and fitted values can then be expressed as follows:

$$\hat{\pi} = \frac{\exp(\mathbf{X}'\mathbf{b})}{1+\exp(\mathbf{X}'\mathbf{b})} = [1+\exp(-\mathbf{X}'\mathbf{b})]^{-1} \tag{14.44a}$$

$$\hat{\pi}_i = \frac{\exp(\mathbf{X}_i'\mathbf{b})}{1+\exp(\mathbf{X}_i'\mathbf{b})} = [1+\exp(-\mathbf{X}_i'\mathbf{b})]^{-1} \tag{14.44b}$$

where:

$$\mathbf{X}'\mathbf{b} = b_0 + b_1 X_1 + \cdots + b_{p-1} X_{p-1} \tag{14.44c}$$

$$\mathbf{X}_i'\mathbf{b} = b_0 + b_1 X_{i1} + \cdots + b_{p-1} X_{i,p-1} \tag{14.44d}$$

Geometric interpretation. Recall that when fitting a standard multiple regression model with two predictors, the estimated regression surface is a plane in three-dimensional space, as shown in Figure 6.7 on page 240 for the Dwaine Studios example. A multiple logistic regression fit based on two continuous predictors can also be represented by a surface in three-dimensional space, but the surface follows the characteristic S-shape that we saw for simple logistic models. For example, Figure 14.8 displays a three-dimensional plot of a logistic response function that depicts the relationship between the development of coronary disease (Y, the binary outcome) and two continuous predictors, cholesterol level (X_1) and age (X_2). This surface increases in an approximately linear fashion for larger values of

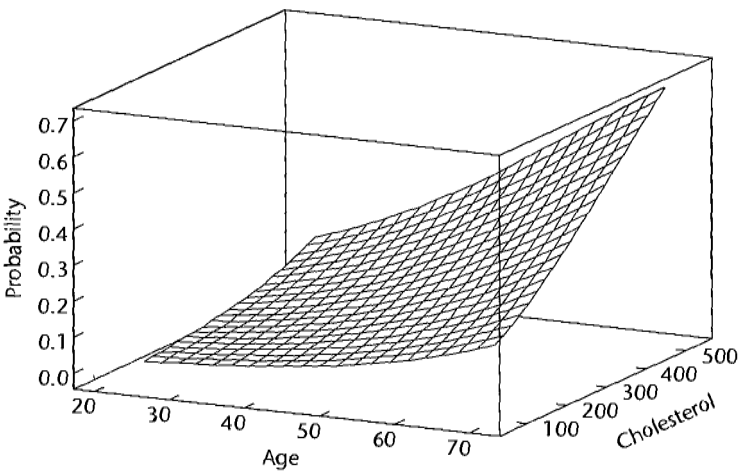


FIGURE 14.8
Three-Dimensional Fitted Logistic Response Surface—Coronary Heart Disease Example.

cholesterol level and age, but levels off and is nearly horizontal for small values of these predictors.

We shall rely on standard statistical packages for logistic regression to conduct the numerical search procedures for obtaining the maximum likelihood estimates. We therefore proceed directly to an example to illustrate the fitting and interpretation of a multiple logistic regression model.

Example

In a health study to investigate an epidemic outbreak of a disease that is spread by mosquitoes, individuals were randomly sampled within two sectors in a city to determine if the person had recently contracted the disease under study. This was ascertained by the interviewer, who asked pertinent questions to assess whether certain specific symptoms associated with the disease were present during the specified period. The response variable Y was coded 1 if this disease was determined to have been present, and 0 if not.

Three predictor variables were included in the study, representing known or potential risk factors. They are age, socioeconomic status of household, and sector within city. Age (X_1) is a quantitative variable. Socioeconomic status is a categorical variable with three levels. It is represented by two indicator variables (X_2 and X_3), as follows:

Class	X_2	X_3
Upper	0	0
Middle	1	0
Lower	0	1

City sector is also a categorical variable. Since there were only two sectors in the study, one indicator variable (X_4) was used, defined so that $X_4 = 0$ for sector 1 and $X_4 = 1$ for sector 2.

The reason why the upper socioeconomic class was chosen as the reference class (i.e., the class for which the indicator variables X_2 and X_3 are coded 0) is that it was expected that this class would have the lowest disease rate among the socioeconomic classes. By making this class the reference class, the odds ratios associated with regression coefficients β_2 and β_3 would then be expected to be greater than 1, facilitating their interpretation. For the same reason, sector 1, where the epidemic was less severe, was chosen as the reference class for the sector indicator variable X_4.

The data for 196 individuals in the sample are given in the disease outbreak data set in Appendix C.10. The first 98 cases were selected for fitting the model. The remaining 98 cases were saved to serve as a validation data set. Table 14.3 in columns 1–5 contains the data for a portion of the 98 cases used for fitting the model. Note the use of the indicator variables as just explained for the two categorical variables. The primary purpose of the study was to assess the strength of the association between each of the predictor variables and the probability of a person having contracted the disease.

A first-order multiple logistic regression model with the three predictor variables was considered *a priori* to be reasonable:

$$E\{Y\} = [1 + \exp(-\mathbf{X}'\boldsymbol{\beta})]^{-1} \tag{14.45}$$

TABLE 14.3 Portion of Model-Building Data Set—Disease Outbreak Example.

	Case	(1) Age	(2) Socioeconomic Status	(3) Socioeconomic Status	(4) City Sector	(5) Disease Status	(6) Fitted Value
	i	X_{i1}	X_{i2}	X_{i3}	X_{i4}	Y_i	$\hat{\pi}_i$
	1	33	0	0	0	0	.209
(Coded)	2	35	0	0	0	0	.219
	3	6	0	0	0	0	.106
	4	60	0	0	0	0	.371
	5	18	0	1	0	1	.111
	6	26	0	1	0	0	.136
	...	...	...	...	...	...	...
	98	35	0	1	0	0	.171

TABLE 14.4 Maximum Likelihood Estimates of Logistic Regression Function (14.45)—Disease Outbreak Example.

(a) Estimated Coefficients, Standard Deviations, and Odds Ratios

Regression Coefficient	Estimated Regression Coefficient	Estimated Standard Deviation	Estimated Odds Ratio
β_0	−3.8877	.9955	—
β_1	.02975	.01350	1.030
β_2	.4088	.5990	1.505
β_3	−.30525	.6041	.737
β_4	1.5747	.5016	4.829

(b) Estimated Approximate Variance-Covariance Matrix

$$\mathbf{s}^2\{\mathbf{b}\} = \begin{array}{c} \begin{matrix} b_0 & b_1 & b_2 & b_3 & b_4 \end{matrix} \\ \begin{bmatrix} .4129 & -.0057 & -.1836 & -.2010 & -.1632 \\ -.0057 & .00018 & .00115 & .00073 & .00034 \\ -.1836 & .00115 & .3588 & .1482 & .0129 \\ -.2010 & .00073 & .1482 & .3650 & .0623 \\ -.1632 & .00034 & .0129 & .0623 & .2516 \end{bmatrix} \end{array}$$

where:

$$\mathbf{X}'\boldsymbol{\beta} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 \qquad (14.45a)$$

This model was fitted by the method of maximum likelihood to the data for the 98 cases. The results are summarized in Table 14.4a. The estimated logistic response function is:

$$\hat{\pi} = [1 + \exp(3.8877 - .02975X_1 - .4088X_2 + .30525X_3 - 1.5747X_4]^{-1} \qquad (14.46)$$

The interpretation of the estimated regression coefficients in the fitted first-order multiple logistic response function parallels that for the simple logistic response function: $\exp(b_k)$ is the estimated odds ratio for predictor variable X_k. The only difference in interpretation for multiple logistic regression is that the estimated odds ratio for predictor variable X_k

assumes that all other predictor variables are held constant. The levels at which they are held constant does not matter in a first-order model. We see from Table 14.4a, for instance, that the odds of a person having contracted the disease increase by about 3.0 percent with each additional year of age (X_1), for given socioeconomic status and city sector location. Also, the odds of a person in sector 2 (X_4) having contracted the disease are almost five times as great as for a person in sector 1, for given age and socioeconomic status. These are point estimates, to be sure, and we shall need to consider how precise these estimates are.

Table 14.3, column 6, contains the fitted values $\hat{\pi}_i$. These are calculated as usual. For instance, the estimated mean response for case $i = 1$, where $X_{11} = 33$, $X_{12} = 0$, $X_{13} = 0$, $X_{14} = 0$, is:

$$\hat{\pi}_1 = \{1 + \exp[2.3129 - .02975(33) - .4088(0) + .30525(0) - 1.5747(0)]\}^{-1} = .209$$

Polynomial Logistic Regression

Occasionally, the first-order logistic model may not provide an adequate fit to the data and a more complicated model may be needed. One such model is the kth-order polynomial logistic regression model, with logit response function:

$$\pi'(x) = \beta_0 + \beta_{11}x + \beta_{22}x^2 + \cdots + \beta_{kk}x^k \qquad \textbf{(14.47)}$$

where x denotes the centered predictor, $X - \bar{X}$. This model for the logit is still linear in the β parameters. For simplicity, we will use a second-order polynomial:

$$\pi'(x) = \beta_0 + \beta_{11}x + \beta_{22}x^2$$

to demonstrate the procedure.

Example

A study of 482 initial public offering companies (IPOs) was conducted to determine the characteristics of companies that attract venture capital. Here, the response of interest is whether or not the company was financed by venture capital funds. Several potential predictors are: the face value of the company; the number of shares offered; and whether or not the company was a leveraged buyout. The IPO data set is listed in Appendix C.11. In this example we consider just one predictor, the face value of the company.

Figure 14.9a contains a plot of venture capital involvement (Y) versus the the natural logarithm of the face value of the company (X) with a lowess smooth and the fitted

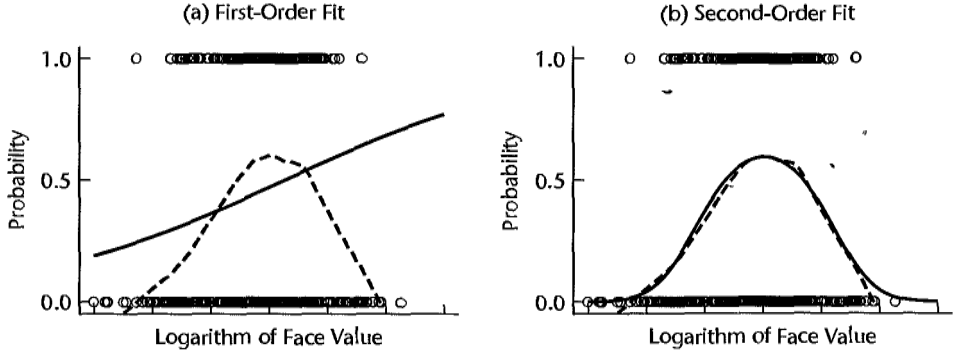


FIGURE 14.9
First- and second-order logistic regression fits (solid lines), the lowess smooths (dashed lines)—IPO example.

TABLE 14.5 Logistic Regression Output for Second-Order Model—IPO Example.

Predictor	Estimated Coefficient	Estimated Standard Error	z^*	*P*-value
Constant	$b_0 = 0.3005$	0.1240	2.42	0.015
x	$b_{11} = 0.5516$	0.1385	3.98	0.000
x^2	$b_{22} = -0.8615$	0.1404	−6.14	0.000

first-order logistic regression fit superimposed. (Here we chose to analyze the natural logarithm of face value because face value ranges over several orders of magnitude, with a highly skewed distribution.) The lowess smooth clearly suggests a mound-shaped relationship: for small and large companies, the likelihood of venture capital involvement is near zero, but for midsized companies it is over .5. The first-order logistic regression fit is unable to capture the characteristic mound shape of the mean response function and is clearly inadequate. Table 14.5 shows the fitted second-order response function:

$$\hat{\pi}' = .3005 + .5516x - .8615x^2$$

where $x = X - \bar{X}$. Also shown in Table 14.5 are three quantities to be discussed in Section 14.5, namely, the estimated standard error of each coefficient, a statistic, z^*, for testing the hypothesis that the coefficient is zero, and the resulting *P*-value. We simply note for now that the *P*-value for b_{22} is .000, confirming the need for a second-order term. Figure 14.9b plots the data, the lowess smooth, and the second-order polynomial logistic regression fit. Note that the second-order polynomial fit tracks the lowess smooth closely.

The above example demonstrated the use of polynomial regression for a single predictor. For multiple logistic regression, higher order polynomial terms and cross-products may be added to improve the fit of a model, as discussed in Section 8.1 in the context of multiple linear regression models.

Comments

1. The maximum likelihood estimates of the parameters $\boldsymbol{\beta}$ for the logistic regression model can be obtained by iteratively reweighted least squares. The procedure is straightforward, although it involves intensive use of a computer.

a. Obtain starting values for the regression parameters, to be denoted by $\mathbf{b}(0)$. Often, reasonable starting values can be obtained by ordinary least squares regression of Y on the predictor variables $X_1, \ldots, X_{p-1}$, using a first-order linear model.

b. Using these starting values, obtain:

$$\hat{\pi}_i'(0) = \mathbf{X}_i'[\mathbf{b}(0)] \tag{14.48a}$$

$$\hat{\pi}_i(0) = \frac{\exp[\hat{\pi}_i'(0)]}{1 + \exp[\hat{\pi}_i'(0)]} \tag{14.48b}$$

c. Calculate the new response variable:

$$Y_i'(0) = \hat{\pi}_i'(0) + \frac{Y_i - \hat{\pi}_i(0)}{\hat{\pi}_i(0)[1 - \hat{\pi}_i(0)]} \tag{14.49a}$$

and the weights:

$$w_i(0) = \hat{\pi}_i(0)[1 - \hat{\pi}_i(0)] \tag{14.49b}$$

d. Regress $Y'(0)$ in (14.49a) on the predictor variables $X_1, \ldots, X_{p-1}$ using a first-order linear model with weights in (14.49b) to obtain revised estimated regression coefficients, denoted by $\mathbf{b}(1)$.

e. Repeat steps b through d, making revisions in (14.48) and (14.49) by using the latest revised estimated regression coefficients until there is little if any change in the estimated coefficients. Often three or four iterations are sufficient to obtain convergence.

2. When the multiple logistic regression model is not a first-order model and contains quadratic or higher-power terms for the predictor variables and/or cross-product terms for interaction effects, the estimated regression coefficients b_k no longer have a simple interpretation.

3. When the assumptions of a monotonic sigmoidal relation between π and $\mathbf{X}'\boldsymbol{\beta}$, required for the multiple logistic regression model, are not appropriate, an alternative is to convert all predictor variables to categorical variables and employ a log-linear model. In the disease outbreak example, for instance, age could be converted into a categorical variable with three classes 0–18, 19–50, and 51–75. Reference 14.2 describes the use of log-linear models for binary response variables when the predictor variables are categorical.

4. Convergence difficulties in the numerical search procedures for finding the maximum likelihood estimates of the multiple logistic regression function may be encountered when the predictor variables are highly correlated or when there is a large number of predictor variables. Another instance that causes convergence problems occurs when a collection of the predictors either completely or nearly perfectly separates the outcome groups. Indication of this problem often can be detected by noting large estimated parameters and large estimated standard errors, similar to what occurs with multicollinearity problems. When convergence problems occur, it may be necessary to reduce the number of predictor variables in order to obtain convergence. ■

14.5 Inferences about Regression Parameters

The same types of inferences are of interest in logistic regression as for linear regression models—inferences about the regression coefficients, estimation of mean responses, and predictions of new observations.

The inference procedures that we shall present rely on large sample sizes. For large samples, under generally applicable conditions, maximum likelihood estimators for logistic regression are approximately normally distributed, with little or no bias, and with approximate variances and covariances that are functions of the second-order partial derivatives of the logarithm of the likelihood function.

Specifically, let $\mathbf{G}$ denote the matrix of second-order partial derivatives of the log-likelihood function in (14.42), the derivatives being taken with regard to the parameters $\beta_0, \beta_1, \ldots, \beta_{p-1}$:

$$\underset{p\times p}{\mathbf{G}} = [g_{ij}] \qquad i = 0, 1, \ldots, p-1;\ j = 0, 1, \ldots, p-1 \tag{14.50}$$

where:

$$g_{00} = \frac{\partial^2 \log_e L(\boldsymbol{\beta})}{\partial \beta_0^2}$$

$$g_{01} = \frac{\partial^2 \log_e L(\boldsymbol{\beta})}{\partial \beta_0 \partial \beta_1}$$

etc.

This matrix is called the *Hessian* matrix. When the second-order partial derivatives in the Hessian matrix are evaluated at $\boldsymbol{\beta} = \mathbf{b}$, that is, at the maximum likelihood estimates, the estimated approximate variance-covariance matrix of the estimated regression coefficients for logistic regression can be obtained as follows:

$$\mathbf{s}^2\{\mathbf{b}\} = ([-g_{ij}]_{\boldsymbol{\beta}=\mathbf{b}})^{-1} \tag{14.51}$$

The estimated approximate variances and covariances in (14.51) are routinely provided by most logistic regression computer packages.

Inferences about the regression coefficients for the simple logistic regression model (14.20) or the multiple logistic regression model (14.41) are based on the following approximate result when the sample size is large:

$$\frac{b_k - \beta_k}{s\{b_k\}} \sim z \qquad k = 0, 1, \ldots, p-1 \tag{14.52}$$

where z is a standard normal random variable and $s\{b_k\}$ is the estimated approximate standard deviation of b_k obtained from (14.51).

Test Concerning a Single β_k: Wald Test

A large-sample test of a single regression parameter can be constructed based on (14.52). For the alternatives:

$$\begin{aligned} H_0&: \beta_k = 0 \\ H_a&: \beta_k \neq 0 \end{aligned} \tag{14.53a}$$

an appropriate test statistic is:

$$z^* = \frac{b_k}{s\{b_k\}} \tag{14.53b}$$

and the decision rule is:

$$\begin{aligned} &\text{If } |z^*| \leq z(1-\alpha/2), \text{ conclude } H_0 \\ &\text{If } |z^*| > z(1-\alpha/2), \text{ conclude } H_a \end{aligned} \tag{14.53c}$$

One-sided alternatives will involve a one-sided decision rule. The testing procedure in (14.53) is commonly referred to as the Wald test. On occasion, the square of z^* is used instead, and the test is then based on a chi-square distribution with 1 degree of freedom. This is also referred to as the Wald test.

Example

In the programming task example, β_1 was expected to be positive. The alternatives of interest therefore are:

$$\begin{aligned} H_0&: \beta_1 \leq 0 \\ H_a&: \beta_1 > 0 \end{aligned}$$

Test statistic (14.53b), using the results in Table 14.1b, is:

$$z^* = \frac{.1615}{.0650} = 2.485$$

For $\alpha = .05$, we require $z(.95) = 1.645$. The decision rule therefore is:

$$\text{If } z^* \leq 1.645, \text{ conclude } H_0$$
$$\text{If } z^* > 1.645, \text{ conclude } H_a$$

Since $z^* = 2.485 > 1.645$, we conclude H_a, that β_1 is positive, as expected. The one-sided P-value of this test is .0065.

Interval Estimation of a Single β_k

From (14.52), we obtain directly the approximate $1 - \alpha$ confidence limits for β_k:

$$b_k \pm z(1 - \alpha/2)s\{b_k\} \tag{14.54}$$

where $z(1 - \alpha/2)$ is the $(1 - \alpha/2)100$ percentile of the standard normal distribution.

The corresponding confidence limits for the odds ratio $\exp(\beta_k)$ are:

$$\exp[b_k \pm z(1 - \alpha/2)s\{b_k\}] \tag{14.55}$$

Example

For the programming task example, it is desired to estimate β_1 with an approximate 95 percent confidence interval. We require $z(.975) = 1.960$, as well as the estimates $b_1 = .1615$ and $s\{b_1\} = .0650$ which are given in Table 14.1b. Hence, the confidence limits are $.1615 \pm 1.960(.0650)$, and the approximate 95 percent confidence interval for β_1 is:

$$.0341 \leq \beta_1 \leq .2889$$

Thus, we can conclude with approximately 95 percent confidence that β_1 is between .0341 and .2889. The corresponding 95 percent confidence limits for the odds ratio are $\exp(.0341) = 1.03$ and $\exp(.2889) = 1.33$.

To examine whether the large-sample inference procedures are applicable here when $n = 25$, bootstrap sampling can be employed, as described in Chapter 13. Alternatively, estimation procedures have been developed for logistic regression that do not depend on any large-sample approximations. LogXact (Reference 14.3) was run on the data and produced 95 percent confidence limits for β_1 of .041 and .296. The large-sample limits of .034 and .289 are reasonably close to the LogXact limits, confirming the applicability of large-sample theory here.

If we wish to consider the odds ratio for persons whose experience differs by, say, five months, the point estimate of this odds ratio would be $\exp(5b_1) = \exp[5(.1615)] = 2.242$, and the 95 percent confidence limits would be obtained from the confidence limits for b_1 as follows: $\exp[5(.0341)] = 1.186$ and $\exp[5(.2889)] = 4.240$. Thus, with 95 percent confidence we estimate that the odds of success increase by between 19 percent and 324 percent with an additional five months of experience.

Comments

1. If the large-sample conditions for inferences are not met, the bootstrap procedure can be employed to obtain confidence limits for the regression coefficients. The bootstrap here requires generating Bernoulli random variables as discussed in Section 14.8 for the construction of simulated envelopes.

2. We are using the z approximation here for large-sample inferences rather than the t approximation used in Chapter 13 for nonlinear regression. This choice is conventional for logistic regression.

For large sample sizes, there is little difference between the t distribution and the standard normal distribution.

3. Approximate joint confidence intervals for several logistic regression parameters can be developed by the Bonferroni procedure. If g parameters are to be estimated with family confidence coefficient of approximately $1 - \alpha$, the joint Bonferroni confidence limits are:

$$b_k \pm Bs\{b_k\} \tag{14.56}$$

where:

$$B = z(1 - \alpha/2g) \tag{14.56a}$$

4. For power and sample size considerations in logistic regression modeling, see Reference 14.4

■

Test whether Several $\beta_k = 0$: Likelihood Ratio Test

Frequently there is interest in determining whether a subset of the X variables in a multiple logistic regression model can be dropped, that is, in testing whether the associated regression coefficients β_k equal zero. The test procedure we shall employ is a general one for use with maximum likelihood estimation, and is analogous to the general linear test procedure for linear models. The test is called the *likelihood ratio test*, and, like the general linear test, is based on a comparison of full and reduced models. The test is valid for large sample sizes.

We begin with the full logistic model with response function:

$$\pi = [1 + \exp(-\mathbf{X}'\boldsymbol{\beta}_F)]^{-1} \qquad \text{Full model} \tag{14.57}$$

where:

$$\mathbf{X}'\boldsymbol{\beta}_F = \beta_0 + \beta_1 X_1 + \cdots + \beta_{p-1} X_{p-1}$$

We then find the maximum likelihood estimates for the full model, now denoted by $\mathbf{b}_F$, and evaluate the likelihood function $L(\boldsymbol{\beta})$ when $\boldsymbol{\beta}_F = \mathbf{b}_F$. We shall denote this value of the likelihood function for the full model by $L(F)$.

The hypothesis we wish to test is:

$$\begin{aligned} &H_0\text{: } \beta_q = \beta_{q+1} = \cdots = \beta_{p-1} = 0 \\ &H_a\text{: not all of the } \beta_k \text{ in } H_0 \text{ equal zero} \end{aligned} \tag{14.58}$$

where, for convenience, we arrange the model so that the last $p - q$ coefficients are those tested. The reduced logistic model therefore has the response function:

$$\pi = [1 + \exp(-\mathbf{X}'\boldsymbol{\beta}_R)]^{-1} \qquad \text{Reduced model} \tag{14.59}$$

where:

$$\mathbf{X}'\boldsymbol{\beta}_R = \beta_0 + \beta_1 X_1 + \cdots + \beta_{q-1} X_{q-1}$$

Now we obtain the maximum likelihood estimates $\mathbf{b}_R$ for the reduced model and evaluate the likelihood function for the reduced model containing q parameters when $\boldsymbol{\beta}_R = \mathbf{b}_R$. We shall denote this value of the likelihood function for the reduced model by $L(R)$. It can be shown that $L(R)$ cannot exceed $L(F)$ since one cannot obtain a larger maximum for the likelihood function using a subset of the parameters.

The actual test statistic for the likelihood ratio test, denoted by G^2, is:

$$G^2 = -2\log_e\left[\frac{L(R)}{L(F)}\right] = -2[\log_e L(R) - \log_e L(F)] \qquad \textbf{(14.60)}$$

Note that if the ratio $L(R)/L(F)$ is small, indicating H_a is the appropriate conclusion, then G^2 is large. Thus, large values of G^2 lead to conclusion H_a.

Large-sample theory states that when n is large, G^2 is distributed approximately as $\chi^2(p-q)$ when H_0 in (14.58) holds. The degrees of freedom correspond to $df_R - df_F = (n-q) - (n-p) = p - q$. The appropriate decision rule therefore is:

$$\begin{aligned} &\text{If } G^2 \leq \chi^2(1-\alpha; p-q), \text{ conclude } H_0 \\ &\text{If } G^2 > \chi^2(1-\alpha; p-q), \text{ conclude } H_a \end{aligned} \qquad \textbf{(14.61)}$$

Example

In the disease outbreak example, the model building began with the three predictor variables that were considered *a priori* to be key explanatory variables—age, socioeconomic status, and city sector. A logistic regression model was fitted containing these three predictor variables and the log-likelihood for this model was obtained. Then tests were conducted to see whether a variable could be dropped from the model. First, age (X_1) was dropped from the logistic model and the log-likelihood for this reduced model was obtained. The results were:

$$L(F) = L(b_0, b_1, b_2, b_3, b_4) = -50.527 \qquad L(R) = L(b_0, b_2, b_3, b_4) = -53.102$$

Hence the required test statistic is:

$$G^2 = -2[\log_e L(R) - \log_e L(F)] = -2[-53.102 - (-50.527)] = 5.150$$

For $\alpha = .05$, we require $\chi^2(.95; 1) = 3.84$. Hence to test H_0: $\beta_1 = 0$, H_a: $\beta_1 \neq 0$, the appropriate decision rule is:

$$\begin{aligned} &\text{If } G^2 \leq 3.84, \text{ conclude } H_0 \\ &\text{If } G^2 > 3.84, \text{ conclude } H_a \end{aligned}$$

Since $G^2 = 5.15 \geq 3.84$, we conclude H_a, that X_1 should not be dropped from the model. The P-value of this test is .023.

Similar tests for socioeconomic status (X_2, X_3) and city sector (X_4) led to P-values of .55 and .001. The P-value for socioeconomic status suggests that it can be dropped from the model containing the other two predictor variables. However, since this variable was considered *a priori* to be important, additional analyses were conducted. When socioeconomic status is the only predictor in the logistic regression model, the P-value for the test whether this predictor variable is helpful is .16, suggesting marginal importance for this variable. In addition, the estimated regression coefficients for age and city sector and their estimated standard deviations are not appreciably affected by whether or not socioeconomic status is in the regression model. Hence, it was decided to keep socioeconomic status in the logistic regression model in view of its *a priori* importance.

The next question of concern was whether any two-factor interaction terms are required in the model. The full model now includes all possible two-factor interactions, in addition

to the main effects, so that $\mathbf{X}'\boldsymbol{\beta}_F$ for this model is as follows:

$$\mathbf{X}'\boldsymbol{\beta}_F = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_1 X_2 + \beta_6 X_1 X_3 + \beta_7 X_1 X_4 + \beta_8 X_2 X_4 + \beta_9 X_3 X_4 \qquad \text{Full model}$$

We wish to test:

$$H_0\text{: } \beta_5 = \beta_6 = \beta_7 = \beta_8 = \beta_9 = 0$$
$$H_a\text{: not all } \beta_k \text{ in } H_0 \text{ equal zero}$$

so that $\mathbf{X}'\boldsymbol{\beta}_R$ for the reduced model is:

$$\mathbf{X}'\boldsymbol{\beta}_R = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 \qquad \text{Reduced model}$$

A computer run of a multiple logistic regression package yielded:

$$L(F) = -46.998$$
$$L(R) = -50.527$$
$$G^2 = -2[\log_e(R) - \log_e(F)] = 7.058$$

If H_0 holds, G^2 follows approximately the chi-square distribution with 5 degrees of freedom. For $\alpha = .05$, we require $\chi^2(.95; 5) = 11.07$. Since $G^2 = 7.058 < 11.07$, we conclude H_0, that the two-factor interactions are not needed in the logistic regression model. The P-value of this test is .22. We note again that a logistic regression model without interaction terms is desirable, because otherwise $\exp(\beta_k)$ no longer can be interpreted as the odds ratio.

Thus, the fitted logistic regression model (14.46) was accepted as the model to be checked diagnostically and, finally, to be validated.

Comment

The Wald test for a single regression parameter in (14.53) is more versatile than the likelihood ratio test in (14.60). The latter can only be used to test H_0: $\beta_k = 0$, whereas the former can be used also for one-sided tests and for testing whether β_k equals some specified value other than zero. When testing H_0: $\beta_k = 0$, the two tests are not identical and may occasionally lead to different conclusions. For example, the Wald test P-value for dropping age when socioeconomic status and sector are in the model for the disease data set example is .0275; the P-value for the likelihood ratio test is .023. ■

14.6 Automatic Model Selection Methods

Several automatic model selection methods are available for building logistic regression models. These include all-possible-regressions and stepwise procedures. We begin with a discussion of criteria for model selection.

Model Selection Criteria

In the context of multiple linear regression models, we discussed the use of the following model selection criteria in Chapter 9: R_p^2, $R_{a,p}^2$, C_p, AIC_p, SBC_p, and $PRESS_p$. For logistic regression modeling, the AIC_p and SBC_p criteria are easily adapted and are generally available in commercial software. For these reasons we will focus on the use of these t

criteria. The modifications are as follows:

$$AIC_p = -2\log_e L(\mathbf{b}) + 2p \quad \textbf{(14.62)}$$

$$SBC_p = -2\log_e L(\mathbf{b}) + p\log_e(n) \quad \textbf{(14.63)}$$

where $\log_e L(\mathbf{b})$ is the log-likelihood expression in (14.42). Promising models will yield relatively small values for these criteria. A third criterion that is frequently provided by software packages is -2 times the log-likelihood, or $-2\log_e L(\mathbf{b})$. For this criterion, we also seek models giving small values. A drawback of this third criterion is that $-2\log_e L(\mathbf{b})$ will never increase as terms are added to the model, because there is no penalty for adding predictors. This is analogous to the use of SSE_p or R_p^2 in multiple linear regression. It is easily seen from (14.62) and (14.63) that AIC_p and SBC_p also involve $-2\log_e L(\mathbf{b})$, but penalties are added based on the number of terms p. This penalty is $2p$ for AIC_p and $p\log_e(n)$ for SBC_p.

Best Subsets Procedures

"Best" subsets procedures were discussed in Section 9.4 in the context of multiple linear regression. Recall that these procedures identify a group of subset models that give the best values of a specified criterion. As long as the number of parameters is not too large (typically less than 30 or 40) these procedures can be useful. As we noted in Section 9.4, time-saving algorithms have been developed that can identify the most promising models, without having to evaluate all 2^{p-1} candidates. These procedures are similarly applicable in the context of logistic regression. We now illustrate the use of the the best subsets procedure based on the AIC_p and SBC_p criteria.

Example

For the disease outbreak example, there are four predictors, age (X_1), socioeconomic status (X_2 and X_3) and city sector (X_4). Normally, it is advantageous to tie the two indicators for the qualitative predictor socioeconomic status together; that is, a model should either have both predictors, or neither. Since very few statistical software packages follow this convention, we will allow them to be independently included. This leads to the $2^4 = 16$ possible regression models listed in columns 2–5 of Table 14.6a. The AIC_p, SBC_p, and $-2\log_e L(\mathbf{b})$ criterion values for each of the 16 models are listed in columns 6–8 of Table 14.6a and are plotted against p in Figures 14.10a–c, respectively.

As shown in Figures 14.10a and 14.10b, both AIC_p and SBC_p are minimized for $p = 3$. Inspection of Table 14.6b reveals that the best two-predictor model for both criteria is based on X_1 (age) and X_4 (city sector). Other models that appear promising on the basis of the AIC_p criterion are the three-predictor subsets based on X_1, X_2, and X_4 and X_1, X_3, and X_4, and the full model based on all four predictors. SBC_p also identifies the two three-predictor subset models just noted, as well as the one-predictor model based on X_4. The tendency of SBC_p to favor smaller models is evident in this example.

The plot of $-2\log_e L(\mathbf{b})$ in Figure 14.10c also points to a two- or three-predictor subset. The additional reduction in $-2\log_e L(\mathbf{b})$ from moving from the best two-predictor model to the best three-predictor model are small, and the returns continue to diminish as we move from three predictors to the full, four-predictor model.

wise Model Selection

As we noted in Chapter 9 in the context of model selection for multiple linear regression, when the number of predictors is large (i.e., 40 or more) the use of all-possible-regression

TABLE 14.6 Best Subsets Results—Disease Outbreak Example.

(a) Results for All Possible Models ($X_{ij} = 1$ if X_j in model i; $X_{ij} = 0$ otherwise)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
			Socioeconomic Status					
Model i	Parameters p	Age X_{i1}	X_{i2}	X_{i3}	City Sector X_{i4}	AIC_p	SBC_p	$-2\log_e L(b)$
1	1	0	0	0	0	124.318	126.903	122.318
2	2	1	0	0	0	118.913	124.083	114.913
3	2	0	1	0	0	124.882	130.052	120.882
4	2	0	0	1	0	122.229	127.399	118.229
5	2	0	0	0	1	111.534	116.704	107.534
6	3	1	1	0	0	119.109	126.864	113.109
7	3	1	0	1	0	117.968	125.723	111.968
8	3	1	0	0	1	108.259	116.014	102.259
9	3	0	1	1	0	124.085	131.840	118.085
10	3	0	1	0	1	112.881	120.636	106.881
11	3	0	0	1	1	112.371	120.126	106.371
12	4	1	1	1	0	119.502	129.842	111.502
13	4	1	1	0	1	109.310	119.650	101.310
14	4	1	0	1	1	109.521	119.861	101.521
15	4	0	1	1	1	114.204	124.543	106.204
16	5	1	1	1	1	111.054	123.979	101.054

(b) Best Four Models for Each Criterion

	AIC_p Criterion		SBC_p Criterion	
Rank	Predictors	AIC_p	Predictors	SBC_p
1	X_1, X_4	108.259	X_1, X_4	116.014
2	X_1, X_2, X_4	109.310	X_4	116.704
3	X_1, X_3, X_4	109.521	X_1, X_2, X_4	119.650
4	X_1, X_2, X_3, X_4	111.054	X_1, X_3, X_4	119.861

procedures for model selection may not be feasible. In such cases, stepwise selection procedures are generally employed. The stepwise procedures discussed in Section 9.4 for multiple linear regression are easily adapted for use in logistic regression. The only change required concerns the decision rule for adding or deleting a predictor. For multiple linear regression this decision is based on t_k, the t-value associated with b_k, and its P-value. For logistic regression, we obtain an analogous procedure by basing the decision on the Wald statistic z^* in (14.53b) for the kth estimated regression parameter, and its P-value. With this change, implementation of the various stepwise variants, such as the forward stepwise, forward selection, and backward elimination algorithms is straightforward. We illustrate the use of forward stepwise selection for the disease outbreak data.

Example

Figure 14.11 provides partial output from the SPSS forward stepwise selection procedure for the disease outbreak example. This routine will add a predictor only if the P-value associated with its Wald test statistic is less than 0.05. In step one, city sector (X_4)

FIGURE 14.10 **Plots of AIC_p, SBC_p, and $-2\log_e L$(b)—Disease Outbreak Example.**

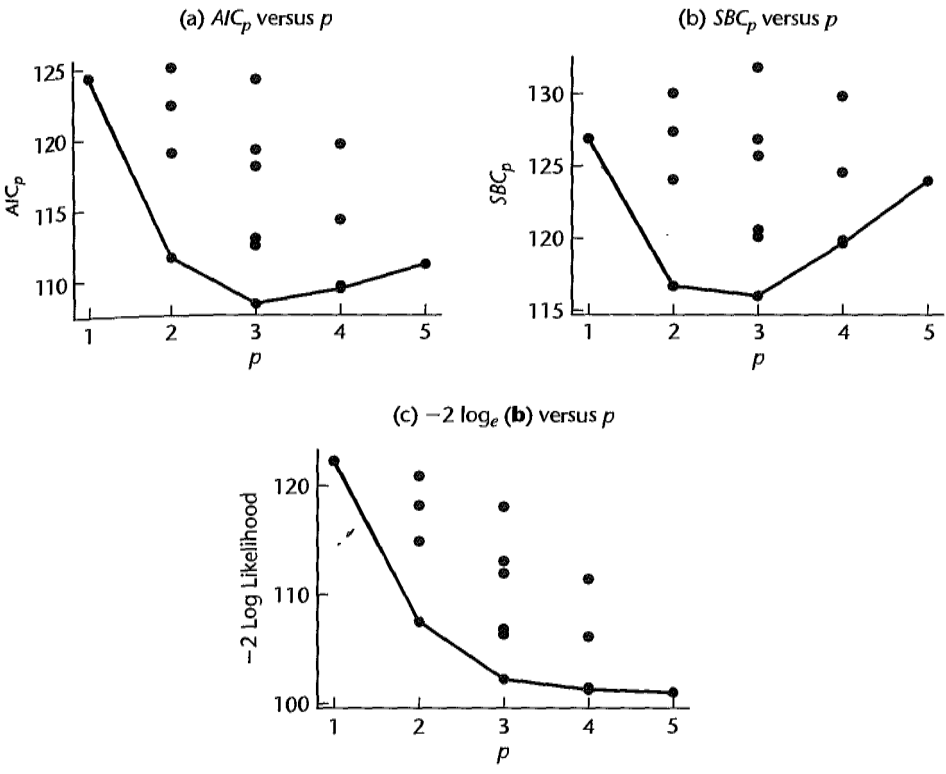


FIGURE 14.11 Partial Output from SPSS Logistic Stepwise Selection Procedure—Disease Outbreak Example.

Logistic Regression
Block 1: Method = Forward Stepwise (Wald)

Variables in the Equation

		B	S.E.	Wald	*df*	Sig.	Exp(*B*)
Step 1[a]	SECTOR	1.743	.473	13.593	1	.000	5.716
	Constant	−3.332	.765	18.990	1	.000	.036
Step 2[b]	AGE	.029	.013	4.946	1	.026	1.030
	SECTOR	1.673	.487	11.791	1	.001	5.331
	Constant	−4.009	.873	21.060	1	.000	.018

a. Variable(s) entered on step 1: SECTOR.

b. Variable(s) entered on step 2: AGE.

entered; its P-value .000. In Step 2, age (X_1) is entered, with a P-value of 0.026. At this point the procedure terminates, because no further predictors can be added with resulting P-values less than 0.05. Thus, the forward stepwise selection procedure has identified the same model favored by AIC_p and SBC_p. Notice that SPSS also prints the square of the Wald test statistics z^* from (14.53b) in the column labeled "Wald." As noted earlier, when $(z^*)^2$ is used, P-values are obtained from a chi-square distribution with 1 degree of freedom.

14.7 Tests for Goodness of Fit

The appropriateness of the fitted logistic regression model needs to be examined before it is accepted for use, as is the case for all regression models. In particular, we need to examine whether the estimated response function for the data is monotonic and sigmoidal in shape, key properties of the logistic response function. Goodness of fit tests provide an overall measure of the fit of the model, and are usually not sensitive when the fit is poor for just a few cases. Logistic regression diagnostics, which focus on individual cases, will be taken up in the next section.

Before discussing several goodness of fit tests, it is necessary to again distinguish between replicated and unreplicated binary data. In Sections 3.7 and 6.8, we discussed the F test for lack-of-fit for the simple and multiple linear regression models. For simple linear regression, the lack-of-fit test requires repeat observations at one or more levels of the single predictor X, and, for multiple regression, there must be multiple or repeat observations that have the same values for all of the predictors. This requirement also holds true for two of the goodness of fit tests that we will present for logistic regression, namely, the Pearson chi-square and the deviance goodness of fit tests. Then we present the Hosmer-Lemeshow test that is useful for unreplicated data sets or for data sets containing just a few replicated observations.

Pearson Chi-Square Goodness of Fit Test

The Pearson chi-square goodness of fit test assumes only that the Y_{ij} observations are independent and that replicated data of reasonable sample size are available. The test can detect major departures from a logistic response function, but is not sensitive to small departures from a logistic response function. The alternatives of interest are:

$$\begin{aligned} H_0&: E\{Y\} = [1 + \exp(-\mathbf{X}'\boldsymbol{\beta})]^{-1} \\ H_a&: E\{Y\} \neq [1 + \exp(-\mathbf{X}'\boldsymbol{\beta})]^{-1} \end{aligned} \tag{14.64}$$

As was the case with tests for lack-of-fit in simple and multiple linear regression, we shall denote the number of distinct combinations of the predictor variables by c, the ith binary response at predictor combination $\mathbf{X}_j$ by Y_{ij}, and the number of cases in the jth class ($j = 1, \ldots, c$) will be denoted by n_j. Recall from (14.32a) that:

$$\sum_{i=1}^{n_j} Y_{ij} = Y_{.j} \tag{14.65}$$

The number of cases in the jth class with outcome 1 will be denoted O_{j1} and the number of cases in the jth class with outcome 0 will be denoted by O_{j0}. Because the response variable Y_{ij} is a Bernoulli variable whose outcomes are 1 and 0, the number of cases O_{j1} and O_{j2} are given as follows:

$$O_{j1} = \sum_{i=1}^{n_j} Y_{ij} = Y_{.j} \tag{14.66}$$

$$O_{j0} = \sum_{i=1}^{n_j} (1 - Y_{ij}) = n_j - Y_{.j} = n_j - O_{j1} \tag{14.67}$$

for $j = 1, \ldots, c$.

If the logistic response function is appropriate, the expected value of Y_{ij} is given by:

$$E\{Y_{ij}\} = \pi_j = [1 + \exp(-\mathbf{X}_j'\beta)]^{-1} \tag{14.67}$$

and is estimated by the fitted value $\hat{\pi}_j$:

$$\hat{\pi}_j = [1 + \exp(-\mathbf{X}_j'\mathbf{b})]^{-1} \tag{14.68}$$

Consequently, if the logistic response function is appropriate, the expected numbers of cases with $Y_{ij} = 1$ and $Y_{ij} = 0$ for the jth class are estimated to be:

$$E_{j1} = n_j\hat{\pi}_j \tag{14.69a}$$

$$E_{j0} = n_j(1 - \hat{\pi}_j) = n_j - E_{j1} \tag{14.69b}$$

where E_{j1} denotes the estimated expected number of 1s in the jth class, and E_{j0} denotes the estimated expected number of 0s in the jth class.

The test statistic is the usual chi-square goodness of fit test statistic:

$$X^2 = \sum_{j=1}^{c}\sum_{k=0}^{1}\frac{(O_{jk} - E_{jk})^2}{E_{jk}} \tag{14.70}$$

If the logistic response function is appropriate, X^2 follows approximately a χ^2 distribution with $c - p$ degrees of freedom when n_j is large and $p < c$. As with other chi-square goodness of fit tests, it is advisable that most expected frequencies E_{jk} be moderately large, say 5 or greater, and none smaller than 1.

Large values of the test statistic X^2 indicate that the logistic response function is not appropriate. The decision rule for testing the alternatives in (14.64), when controlling the level of significance at α, therefore is:

$$\begin{aligned} &\text{If } X^2 \le \chi^2(1-\alpha; c-p), \text{ conclude } H_0 \\ &\text{If } X^2 > \chi^2(1-\alpha; c-p), \text{ conclude } H_a \end{aligned} \tag{14.71}$$

ample

For the coupon effectiveness example, we have five classes. Table 14.7 provides for each class j: n_j, the number of binary outcomes; $\hat{\pi}_j$, the model-based estimate of π_j; p_j, the observed proportion of 1s; O_{j0} and O_{j1}, the number of cases with $Y_{ij} = 0$ and $Y_{ij} = 1$ for each class; and finally, the estimated expected frequencies E_{j0} and E_{j1}, if the logistic regression model (14.35) is appropriate (calculations not shown).

E 14.7
ess of
or
e
on
l
eness

Class				Number of Coupons Not Redeemed		Number of Coupons Redeemed	
				Observed	Expected	Observed	Expected
j	n_j	$\hat{\pi}_j$	p_j	O_{j0}	E_{j0}	O_{j1}	E_{j1}
1	200	.1736	.150	170	165.3	30	34.7
2	200	.2543	.275	145	149.1	55	50.9
3	200	.3562	.350	130	128.8	70	71.2
4	200	.4731	.500	100	105.4	100	94.6
5	200	.7028	.685	63	59.4	137	140.6

Test statistic (14.76) is calculated as follows:

$$X^2 = \frac{(170 - 165.3)^2}{165.3} + \frac{(30 - 34.7)^2}{34.7} + \cdots + \frac{(137 - 140.6)^2}{140.6}$$
$$= 2.15$$

For $\alpha = 0.05$ and $c - p = 5 - 2 = 3$, we require $\chi^2(.95; 3) = 7.81$. Since $X^2 = 2.15 \leq 7.81$, we conclude H_0, that the logistic response function is appropriate. The P-value of the test is .54.

Deviance Goodness of Fit Test

The *deviance goodness of fit test* for logistic regression models is completely analogous to the F test for lack of fit for simple and multiple linear regression models. Like the F test for lack of fit and the Pearson chi-square goodness of fit test, we assume there are c unique combinations of the predictors denoted $X_1, \ldots, X_c$, the number of repeat binary observations at X_j is n_j, and the ith binary response at predictor combination X_j is denoted Y_{ij}.

The lack of fit test for standard regression was based on the general linear test of the reduced model $E\{Y_{ij}\} = \mathbf{X}_j'\boldsymbol{\beta}$ against the full model $E\{Y_{ij}\} = \mu_i$. In similar fashion, the deviance goodness of fit test is based on a likelihood ratio test of the reduced model:

$$E\{Y_{ij}\} = [1 + \exp(-\mathbf{X}_j'\boldsymbol{\beta})]^{-1} \qquad \text{Reduced model} \tag{14.72}$$

against the full model:

$$E\{Y_{ij}\} = \pi_j \qquad j = 1, \ldots, c \qquad \text{Full model} \tag{14.73}$$

where π_j are parameters, $j = 1, \ldots, c$. In the lack of fit test for standard regression, the full model allowed for a unique mean for each unique combination of the predictors, X_j. Similarly, the full model for the deviance goodness of fit test allows for a unique probability π_j for each predictor combination. This full model in the logistic regression case is usually referred to as the *saturated model*.

To carry out the likelihood ratio test in (14.60), we must obtain the values of the maximized likelihoods for the full and reduced models, namely $L(F)$ and $L(R)$. $L(R)$ is obtained by fitting the reduced model, and the maximum likelihood estimates of the c parameters in the full model are given by the sample proportions in (14.32b):

$$p_j = \frac{Y_{.j}}{n_j} \qquad j = 1, 2, \ldots, c \tag{14.74}$$

Letting $\hat{\pi}_j$ denote the reduced model estimate of π_j at X_j, $j = 1, \ldots, c$, it can be shown that likelihood ratio test statistic (14.60) is given by:

$$G^2 = -2[\log_e L(R) - \log_e L(F)]$$
$$= -2\sum_{j=1}^{c}\left[Y_{.j}\log_e\left(\frac{\hat{\pi}_j}{p_j}\right) + (n_j - Y_{.j})\log_e\left(\frac{1 - \hat{\pi}_j}{1 - p_j}\right)\right]$$
$$= DEV(X_0, X_1, \ldots, X_{p-1}) \tag{14.75}$$

The likelihood ratio test statistic in (14.75) is called the *deviance*, and we use $DEV(X_0, X_1, \ldots, X_{p-1})$ to denote the deviance for a logistic regression model based on predictors $X_0, X_1, \ldots, X_{p-1}$. The deviance measures the deviation, in terms of $-2\log_e L$, between the saturated model and the fitted reduced logistic regression model based on $X_0, X_1, \ldots, X_{p-1}$.

If the logistic response function is the correct response function and the sample sizes n_j are large, then the deviance will follow approximately a chi-square distribution with $c - p$ degrees of freedom. Large values of the deviance indicate that the fitted logistic model is not correct. Hence, to test the alternatives:

$$\begin{aligned} H_0&: E\{Y\} = [1 + \exp(-\mathbf{X}'\boldsymbol{\beta})]^{-1} \\ H_a&: E\{Y\} \neq [1 + \exp(-\mathbf{X}'\boldsymbol{\beta})]^{-1} \end{aligned} \tag{14.76}$$

the appropriate decision rule is:

$$\begin{aligned} &\text{If } DEV(X_0, X_1, \ldots, X_{p-1}) \leq \chi^2(1-\alpha; c-p), \text{ conclude } H_0 \\ &\text{If } DEV(X_0, X_1, \ldots, X_{p-1}) > \chi^2(1-\alpha; c-p), \text{ conclude } H_a \end{aligned} \tag{14.77}$$

Example

For the coupon effectiveness example, we use the results in Table 14.2 to calculate the deviance in (14.75) directly:

$$\begin{aligned} DEV(X_0, X_1) &= -2\left[30\log_e\left(\frac{.1736}{.150}\right) + (200-30)\log_e\left(\frac{.8264}{.850}\right)\right. \\ &\quad \left. + \cdots + 137\log_e\left(\frac{.7028}{.685}\right) + (200-137)\log_e\left(\frac{.2972}{.315}\right)\right] \\ &= 2.16 \end{aligned}$$

For $\alpha = .05$ and $c - p = 3$, we require $\chi^2(.95; 3) = 7.81$. Since $DEV(X_0, X_1) = 2.16 \leq 7.81$, we conclude H_0, that the logistic model is a satisfactory fit. The P-value of this test is approximately .54, the same as that obtained earlier for the Pearson chi-square goodness of fit test.

Comment

If $p_j = 0$ for some j in the first term in (14.75), then $Y_{\cdot j} = 0$ and:

$$Y_{\cdot j}\log_e\left(\frac{\hat{\pi}_j}{p_j}\right) = 0$$

Similarly, if $p_j = 1$ for some j in the second term in (14.75), then $Y_{\cdot j} = n_j$ and:

$$(n_j - Y_{\cdot j})\log_e\left(\frac{1-\hat{\pi}_j}{1-p_j}\right) = 0$$

■

er-Lemeshow Goodness of Fit Test

Hosmer and Lemeshow (Reference 14.4) proposed, for either unreplicated data sets or data sets with few replicates, the grouping of cases based on the values of the estimated probabilities. Suppose there are no replicates, i.e., $n_j = 1$ for all j. The procedure consists of grouping the data into classes with similar fitted values $\hat{\pi}_i$, with approximately the same

TABLE 14.8 Hosmer-Lemeshow Goodness of Fit Test for Logistic Regression Function—Disease Outbreak Example.

Class j	$\hat{\pi}_i'$ Interval	n_j	Number of Persons without Disease: Observed O_{j0}	Number of Persons without Disease: Expected E_{j0}	Number of Persons with Disease: Observed O_{j1}	Number of Persons with Disease: Expected E_{j1}
1	−2.60—under −2.08	20	19	18.196	1	1.804
2	−2.08—under −1.43	20	17	17.093	3	2.907
3	−1.43—under −.70	20	14	14.707	6	5.293
4	−.70—under .16	19	9	10.887	10	8.113
5	.16—under 1.70	19	8	6.297	11	12.703
	Total	98	67	67.180	31	30.820

number of cases in each class. The grouping may be accomplished equivalently by using the fitted logit values $\hat{\pi}_i' = \mathbf{X}_i'\mathbf{b}$ since the logit values $\hat{\pi}_i'$ are monotonically related to the fitted mean responses $\hat{\pi}_i$. We shall do the grouping according to the fitted logit values $\hat{\pi}_i'$. Use of from 5 to 10 classes is common, depending on the total number of cases. Once the groups are formed, then the Hosmer-Lemeshow goodness of fit statistic is calculated by using the Pearson chi-square test statistic (14.70) from the $c \times 2$ table of observed and expected frequencies as described earlier. Hosmer and Lemeshow showed, using an extensive simulation study, that the test statistic (14.70) is well approximated by the chi-square distribution with $c - 2$ degrees of freedom.

Example

For the disease outbreak example, we shall use five classes. Table 14.8 shows the class intervals for the logit fitted values $\hat{\pi}_i'$ and the number of cases n_j in each class. It also gives O_{j0} and O_{j1}, the number of cases with $Y_i = 0$ and $Y_i = 1$ for each class. Finally, Table 14.8 contains the estimated expected frequencies E_{j0} and E_{j1} based on logistic regression model (14.46) (calculations not shown).

Test statistic (14.70) is calculated as follows:

$$X^2 = \frac{(19 - 18.196)^2}{18.196} + \frac{(1 - 1.804)^2}{1.804} + \cdots + \frac{(8 - 6.297)^2}{6.297} + \frac{(11 - 12.703)^2}{12.703}$$
$$= 1.98$$

Since all of the n_j are approximately 20 and only two expected frequencies are less than 5 and both are greater than 1, the chi-square test is appropriate here. For $\alpha = .05$ and $c-2 = 3$, we require $\chi^2(.95; 3) = 7.81$. Since $X^2 = 1.98 \leq 7.81$, we conclude H_0, that the logistic response function is appropriate. The P-value of the test is .58.

Comment

We have noted that the Pearson chi-square and deviance goodness of fit tests are only appropriate when there are repeat observations and when the number of replicates at each X category is sufficiently large. Care must be taken in interpreting logistic regression output since some packages will provide these statistics and the associated P-values whether or not sufficient numbers of replicate observations are present. ■

14.8 Logistic Regression Diagnostics

In this section we take up the analysis of residuals and the identification of influential cases for logistic regression. We shall first introduce various residuals that have been defined for logistic regression and some associated plots. We then turn to the identification of influential observations. Throughout, we shall assume that the responses are binary; i.e., we focus on the ungrouped case.

Logistic Regression Residuals

Residual analysis for logistic regression is more difficult than for linear regression models because the responses Y_i take on only the values 0 and 1. Consequently, the ith ordinary residual, e_i will assume one of two values:

$$e_i = \begin{cases} 1 - \hat{\pi}_i & \text{if } Y_i = 1 \\ -\hat{\pi}_i & \text{if } Y_i = 0 \end{cases} \tag{14.78}$$

The ordinary residuals will not be normally distributed and, indeed, their distribution under the assumption that the fitted model is correct is unknown. Plots of ordinary residuals against fitted values or predictor variables will generally be uninformative.

Pearson Residuals. The ordinary residuals can be made more comparable by dividing them by the estimated standard error of Y_i, namely, $\sqrt{\hat{\pi}_i(1-\hat{\pi}_i)}$. The resulting *Pearson residuals* are given by:

$$r_{P_i} = \frac{Y_i - \hat{\pi}_i}{\sqrt{\hat{\pi}_i(1-\hat{\pi}_i)}} \tag{14.79}$$

The Pearson residuals are directly related to Pearson chi-square goodness of fit statistic (14.70). To see this we first expand (14.70) as follows:

$$X^2 = \sum_{j=1}^{c}\sum_{k=0}^{1}\frac{(O_{jk} - E_{jk})^2}{E_{jk}} = \sum_{j=1}^{c}\frac{(O_{j0} - E_{j0})^2}{E_{j0}} + \sum_{j=1}^{c}\frac{(O_{j1} - E_{j1})^2}{E_{j1}} \tag{14.79a}$$

For binary outcome data, we set $j = i$, $c = n$, $O_{j1} = Y_i$, $O_{j0} = 1 - Y_i$, $E_{j1} = \hat{\pi}_i$, $E_{j0} = 1 - \hat{\pi}_i$, and (14.79a) becomes:

$$\begin{aligned} X^2 &= \sum_{i=1}^{n}\frac{[(1-Y_i) - (1-\hat{\pi}_i)]^2}{1-\hat{\pi}_i} + \sum_{i=1}^{n}\frac{(Y_i - \hat{\pi}_i)^2}{\hat{\pi}_i} \\ &= \sum_{i=1}^{n}\frac{(Y_i - \hat{\pi}_i)^2}{1-\hat{\pi}_i} + \sum_{i=1}^{n}\frac{(Y_i - \hat{\pi}_i)^2}{\hat{\pi}_i} \\ &= \sum_{i=1}^{n}\frac{(Y_i - \hat{\pi}_i)^2}{\hat{\pi}_i(1-\hat{\pi}_i)} \end{aligned} \tag{14.79b}$$

Hence, we see that the sum of the squares of the Pearson residuals (14.79) is numerically equal to the Pearson chi-square test statistic (14.79a). Therefore the square of each Pearson residual measures the contribution of each binary response to the Pearson chi-square test statistic. Note that test statistic (14.79b) does not follow an approximate chi-square distribution for binary data without replicates.

Studentized Pearson Residuals. The Pearson residuals do not have unit variance since no allowance has been made for the inherent variation in the fitted value $\hat{\pi}_i$. A better procedure is to divide the ordinary residuals by their estimated standard deviation. This value is approximated by $\sqrt{\hat{\pi}_i(1-\hat{\pi}_i)(1-h_{ii})}$, where h_{ii} is the ith diagonal element of the $n \times n$ estimated hat matrix for logistic regression:

$$\mathbf{H} = \hat{\mathbf{W}}^{\frac{1}{2}}\mathbf{X}(\mathbf{X}'\hat{\mathbf{W}}\mathbf{X})^{-1}\mathbf{X}'\hat{\mathbf{W}}^{\frac{1}{2}} \tag{14.80}$$

Here, $\hat{\mathbf{W}}$ is the $n \times n$ diagonal matrix with elements $\hat{\pi}_i(1-\hat{\pi}_i)$, $\mathbf{X}$ is the usual $n \times p$ design matrix (6.18b), and $\hat{\mathbf{W}}^{\frac{1}{2}}$ is a diagonal matrix with diagonal elements equal to the square roots of those in $\hat{\mathbf{W}}$. The resulting *studentized Pearson residuals* are defined as:

$$r_{SP_i} = \frac{r_{P_i}}{\sqrt{1-h_{ii}}} \tag{14.81}$$

Recall that for multiple linear regression, the hat matrix satisfies the matrix expression $\hat{\mathbf{Y}} = \mathbf{HY}$. The hat matrix for logistic regression is developed in analogous fashion; it satisfies approximately the expression $\hat{\pi}' = \mathbf{HY}$, where $\hat{\pi}'$ is the $(n \times 1)$ vector of linear predictors.

Deviance Residuals. The model deviance (14.75) was obtained by carrying out the likelihood ratio test where the reduced model is the logistic regression model and the full model is the saturated model for grouped outcome data. For binary outcome data, we take the number of X categories to be $c = n$, $n_j = 1$, $j = i$, $Y_{\cdot j} = Y_i$, $p_j = Y_{\cdot j}/n_j = Y_i$, and (14.75) becomes:

$$\begin{aligned} G^2 &= -2\sum_{i=1}^{n}\left[Y_i \log_e\left(\frac{\hat{\pi}_i}{Y_i}\right) + (1-Y_i)\log_e\left(\frac{1-\hat{\pi}_i}{1-Y_i}\right)\right] \\ &= -2\sum_{i=1}^{n}[Y_i \log_e(\hat{\pi}_i) + (1-Y_i)\log_e(1-\hat{\pi}_i) - Y_i \log_e(Y_i) - (1-Y_i)\log_e(1-Y_i)] \\ &= -2\sum_{i=1}^{n}[Y_i \log_e(\hat{\pi}_i) + (1-Y_i)\log_e(1-\hat{\pi}_i)] \end{aligned} \tag{14.82}$$

since $Y_i \log_e(Y_i) = (1-Y_i)\log_e(1-Y_i) = 0$ for $Y_i = 0$ or $Y_i = 1$. Thus for binary data the model deviance in (14.75) is:

$$DEV(X_0, \ldots, X_{p-1}) = -2\sum_{i=1}^{n}[Y_i \log_e(\hat{\pi}_i) + (1-Y_i)\log_e(1-\hat{\pi}_i)] \tag{14.82a}$$

The deviance residual for case i, denoted by dev_i, is defined as the signed square root of the contribution of the ith case to the model deviance DEV in (14.82a):

$$dev_i = sign(Y_i - \hat{\pi}_i)\sqrt{-2[Y_i \log_e(\hat{\pi}_i) + (1-Y_i)\log_e(1-\hat{\pi}_i)]} \tag{14.83}$$

where the sign is positive when $Y_i \geq \hat{\pi}_i$ and negative when $Y_i < \hat{\pi}_i$. Thus the sum of the squared deviance residuals equals the model deviance in (14.82a):

$$\sum_{i=1}^{n}(dev_i)^2 = DEV(X_0, X_1, \ldots, X_{p-1})$$

TABLE 14.9 Logistic Regression Residuals and Hat Matrix Diagonal Elements—Disease Outbreak Example.

i	(1) Y_i	(2) $\hat{\pi}_i$	(3) e_i	(4) r_{P_i}	(5) r_{SP_i}	(6) dev_i	(7) h_{ii}
1	0	0.209	−0.209	−0.514	−0.524	−0.685	.039
2	0	0.219	−0.219	−0.529	−0.541	−0.703	.040
3	0	0.106	−0.106	−0.344	−0.350	−0.473	.033
...	...	...	...	...	...	...	...
96	0	0.114	−0.114	−0.358	−0.363	−0.491	.025
97	0	0.092	−0.092	−0.318	−0.322	−0.439	.024
98	0	0.171	−0.171	−0.455	−0.463	−0.613	.036

Therefore the square of each deviance residual measures the contribution of each binary response to the deviance goodness of fit test statistic (14.82a). Note that test statistic (14.82a) does not follow an approximate chi-square distribution for binary data without replicates.

Example

Table 14.9 lists in columns 1–7, for a portion of the disease outbreak example, the response Y_i, the predicted mean response $\hat{\pi}_i$, the ordinary residual e_i, the Pearson residual r_{P_i}, the studentized Pearson residual r_{SP_i}, the deviance residual dev_i, and the hat matrix diagonal elements h_{ii}. We illustrate the calculations needed to obtain these residuals for the first case. The ordinary residual for the first case is from (14.78):

$$e_1 = Y_1 - \hat{\pi}_1 = 0 - .209 = -.209$$

The first Pearson residual (14.79) is:

$$r_{P_1} = \frac{e_1}{\sqrt{\hat{\pi}_1(1-\hat{\pi}_1)}} = \frac{-.209}{\sqrt{.209(1-.209)}} = -.514$$

Substitution of r_{P_1} and the leverage value h_{11} from column 7 of Table 14.9 into (14.81) yields the studentized Pearson residual:

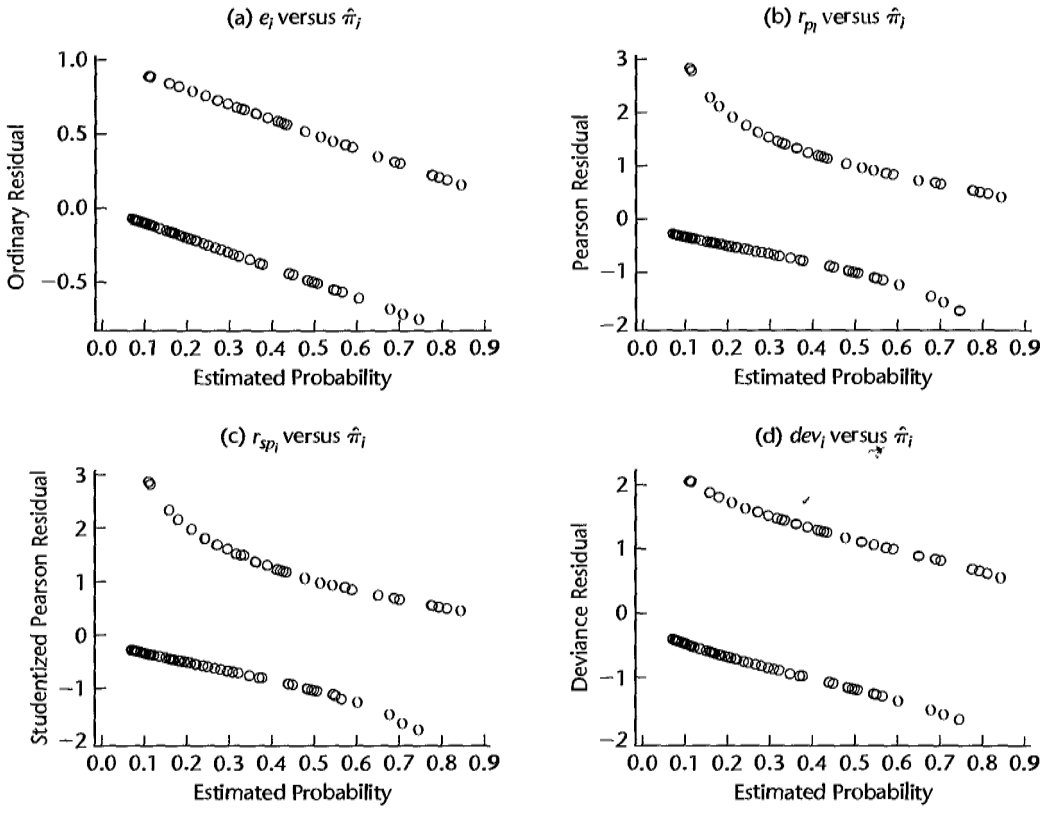
$$r_{SP_1} = \frac{r_{P_1}}{\sqrt{1-h_{11}}} = \frac{-.514}{\sqrt{1-.039}} = -.524$$

Finally, the first deviance residual is obtained from (14.83):

$$\begin{aligned} dev_1 &= sign(Y_1 - \hat{\pi}_1)\sqrt{-2[Y_1 \log_e(\hat{\pi}_1) + (1-Y_1)\log_e(1-\hat{\pi}_1)]} \\ &= sign(-.209)\sqrt{-2[0 \log_e(.209) + (1-0)\log_e(1-.209)]} \\ &= -\sqrt{-2\log_e(.791)} = -.685 \end{aligned}$$

The various residuals are plotted against the predicted mean response in Figure 14.12, although we emphasize that such plots are not particularly informative. Consider, for example, the ordinary residuals in Figure 14.12a. Here we see two trends of decreasing residuals with slope equal to −1. These two linear trends result from the fact, noted above, that the residuals take on just one of two values at a point X_i, $1 - \hat{\pi}_i$ or $0 - \hat{\pi}_i$. Plotting these values against $\hat{\pi}_i$ will always result in two linear trends with slope −1. The remaining plots lead to similar patterns.

FIGURE 14.12 Selected Residuals Plotted against Predicted Mean Response—Disease Outbreak Example.



Diagnostic Residual Plots

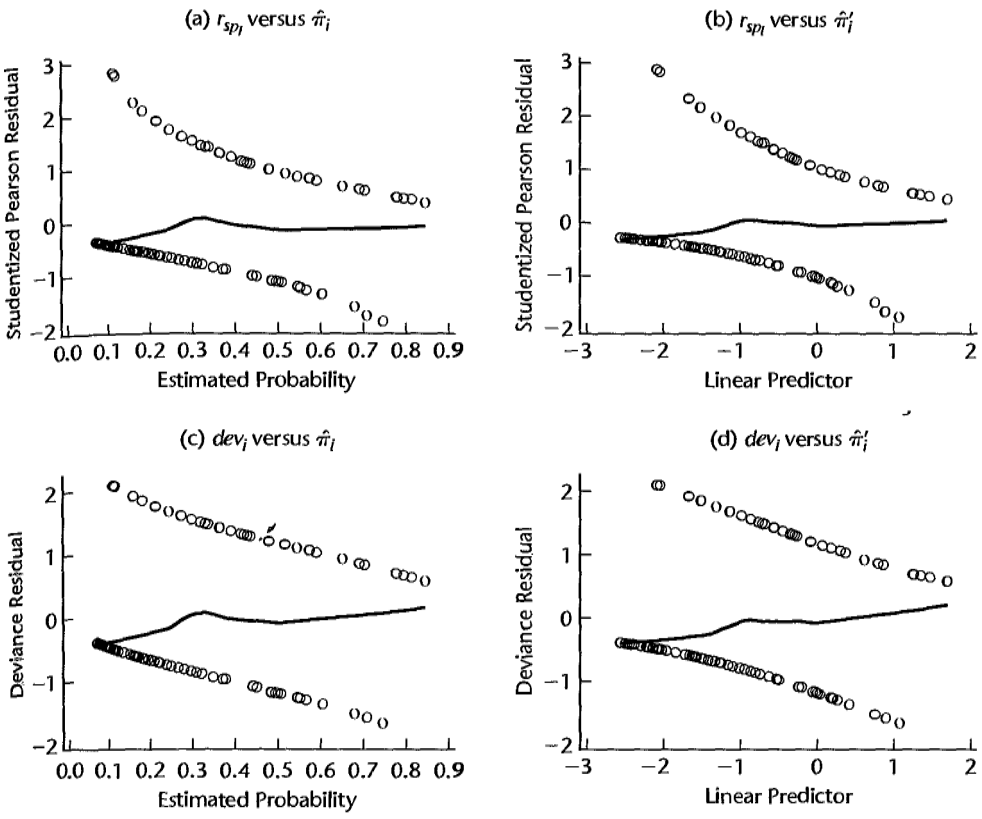
In this section we consider two useful residual plots that provide some information about the adequacy of the logistic regression fit. Recall that in ordinary regression, residual plots are useful for diagnosing model inadequacy, nonconstant variance, and the presence of response outliers. In logistic regression, we generally focus only on the detection of model inadequacy. As we discussed in Section 14.1, nonconstant variance is always present in the logistic regression setting, and the form that it takes is known. Moreover, response outliers in binary logistic regression are difficult to diagnose and may only be evident if all responses in a particular region of the X space have the same response value except one or two. Thus we focus here on model adequacy.

Residuals versus Predicted Probabilities with Lowess Smooth. If the logistic regression model is correct, then $E\{Y_i\} = \pi_i$ and it follows asymptotically that:

$$E\{Y_i - \hat{\pi}_i\} = E\{e_i\} = 0$$

This suggests that if the model is correct, a lowess smooth of the plot of the residuals against the estimated probability $\hat{\pi}_i$ (or against the linear predictor $\hat{\pi}_i'$) should result approximately in a horizontal line with zero intercept. Any significant departure from this

FIGURE 14.13 Residual Plots with Lowess Smooth—Disease Outbreak Example.



suggests that the model may be inadequate. In practice, the lowess smooth of the ordinary residuals, the Pearson residuals, or the studentized Pearson residuals can be employed. (Further details regarding the plotting of logistic regression residuals can be found in Reference 14.5.)

Example

Shown in Figures 14.13a–d are residual plots for the disease outbreak example, each with the suggested lowess smooth superimposed. (We used the MINITAB lowess option with degree of smoothing equal to .7 and number of steps equal to 0 to produce these plots.) In Figures 14.13a and 14.13b, the studentized Pearson residuals are plotted respectively against the estimated probability and the linear predictor. Figures 14.13c and 14.13d provide similar plots for the deviance residuals. In all cases, the lowess smooth approximates a line having zero slope and intercept, and we conclude that no significant model inadequacy is apparent.

Half-Normal Probability Plot with Simulated Envelope. A half-normal probability plot of the deviance residuals with a simulated envelope is useful both for examining the adequacy of the linear part of the logistic regression model and for identifying deviance residuals that are outlying. A half-normal probability plot helps to highlight outlying deviance residuals even though the residuals are not normally distributed. In a normal probability plot, the kth

ordered residual is plotted against the percentile $z[(k - .375)/(n + .25)]$ or against $\sqrt{MSE}$ times this percentile, as shown in (3.6). In a half-normal probability plot, the kth ordered *absolute* residual is plotted against:

$$z\left(\frac{k + n - 1/8}{2n + 1/2}\right) \tag{14.84}$$

Outliers will appear at the top right of a half-normal probability plot as points separated from the others. However, a half-normal plot of the absolute residuals will not necessarily give a straight line even when the fitted model is in fact correct.

To identify outlying deviance residuals, we combine a half-normal probability plot with a *simulated envelope* (Reference 14.6). This envelope constitutes a band such that the plotted residuals are all likely to fall within the band if the fitted model is correct.

A simulated envelope for a half-normal probability plot of the absolute deviance residuals is constructed in the following way:

1. For each of the n cases, generate a Bernoulli outcome (0, 1), where the Bernoulli parameter for case i is $\hat{\pi}_i$, the estimated probability of response $Y_i = 1$ according to the originally fitted model.

2. Fit the logistic regression model for the n new responses where the predictor variables keep their original values, and obtain the deviance residuals. Order the absolute deviance residuals in ascending order.

3. Repeat the first two steps 18 times.

4. Assemble the smallest absolute deviance residuals from the 19 groups and determine the minimum value, the mean, and the maximum value of these 19 residuals.

5. Repeat step 4 by assembling the group of second smallest absolute residuals, the group of third smallest absolute residuals, etc.

6. Plot the minimum, mean, and maximum values for each of the n ordered residual groups against the corresponding expected value in (14.84) on the half-normal probability plot for the original data and connect the points by straight lines.

By using 19 simulations, there is one chance in 20, or 5 percent, that the largest absolute deviance residual from the original data set lies outside the simulated envelope when the fitted model is correct. Large deviations of points from the means of the simulated values or the occurrence of points near to or outside the simulated envelope, are indications that the fitted model is not appropriate.

Example

Table 14.10a repeats a portion of the data for the disease outbreak example, as well as the fitted values for the logistic regression model. It also contains a portion of the simulated responses for the 19 simulation samples. For instance, the simulated responses for case 1 were obtained by generating Bernoulli random outcomes with probability $\hat{\pi}_1 = .209$.

Table 14.10b shows some of the ordered absolute deviance residuals for the 19 simulation samples. Finally, Table 14.10c presents the minimum, mean, and maximum for the 19 simulation samples for some of the rank order positions, the ordered absolute deviance for the original sample for these rank order positions, and corresponding z percentiles. The results in Table 14.10c are plotted in Figure 14.14. We see clearly from this figure that the largest deviance residuals (which here correspond to cases 5 and 14) are farthest to the right and are somewhat separated from the other cases. However, they fall well within the

TABLE 14.10 Results for Simulated Envelope for Half-Normal Probability Plot—Disease Outbreak Example.

(a) Simulated Bernoulli Outcomes

			Simulation Sample		
i	Y_i	$\hat{\pi}_i$	(1)	···	(19)
1	0	.209	0	···	0
2	0	.219	0	···	0
···	···	···	···	···	···
97	0	.092	0	···	0
98	0	.171	1	···	0

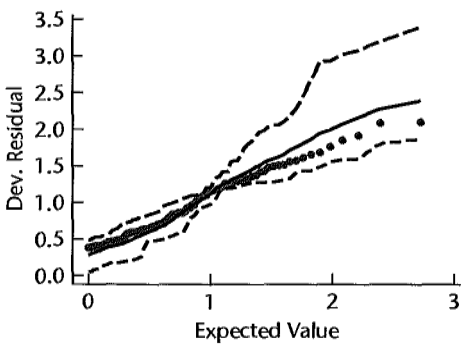
(b) Ordered Absolute Deviance Residuals for Simulation Samples

Order Position	Simulation Sample		
k	(1)	···	(19)
1	.468	···	.368
2	.468	···	.368
···	···	···	···
97	1.849	···	2.085
98	1.919	···	2.228

(c) Minimum, Mean, and Maximum of Ordered Absolute Deviance Residuals for Simulation Samples

Order Position	Simulation Samples			Original	
k	Minimum	Mean	Maximum	Data	$z\left(\frac{k+97.875}{196.5}\right)$
1	.046	.289	.491	.386	.008
2	.060	.296	.491	.386	.021
···	···	···	···	···	···
97	1.804	2.273	3.194	2.082	2.397
98	1.869	2.387	3.391	2.098	2.729

FIGURE 14.14 Half-Normal Probability Plot and Simulated Envelope—Disease Outbreak Example.



simulated envelope so that remedial measures do not appear to be required. Figure 14.10 also shows that most of the absolute deviance residuals fall near the simulation means, suggesting that the logistic regression model is appropriate here.

Detection of Influential Observations

In this section we introduce three measures that can be used to identify influential observations. We consider the influence of individual binary cases on three aspects of the analysis:

1. The Pearson chi-square statistic (14.79b).
2. The deviance statistic (14.82a).
3. The fitted linear predictor, $\hat{\pi}_i'$.

As was the case in standard regression situations, we will employ case-deletion diagnostics to assess the effect of individual cases on the results of the analysis.

Influence on Pearson Chi-Square and the Deviance Statistics. Let X^2 and DEV denote the Pearson and deviance statistics (14.79b) and (14.82a) based on the full data set, and let $X^2_{(i)}$ and $DEV_{(i)}$ denote the values of these test statistics when case i is deleted. The ith *delta chi-square statistic* is defined as the change in the Pearson statistic when the ith case is deleted:

$$\Delta X_i^2 = X^2 - X^2_{(i)}$$

Similarly, the ith *delta deviance statistic* is defined as the change in the deviance statistic when the ith case is deleted:

$$\Delta dev_i = DEV - DEV_{(i)}$$

Determination of the n delta chi-square statistics or the n delta deviance statistics requires n maximizations of the likelihood, which can be time consuming. For faster computing, the following one-step approximations have been developed:

$$\Delta X_i^2 = r^2_{SP_i} \tag{14.85}$$

$$\Delta dev_i = h_{ii} r^2_{SP_i} + dev_i^2 \tag{14.86}$$

In summary, ΔX_i^2 and Δdev_i give the change in the Pearson chi-square and deviance statistics, respectively, when the ith case is deleted. They therefore provide measures of the influence of the ith case on these summary statistics.

Interpretation of the delta chi-square and delta deviance statistics is not always a simple matter. In standard regression situations, we employ various rules of thumb for judging the magnitude of a regression diagnostic. An example of this is the Bonferroni outlier test (Section 10.2) that is used in conjunction with the studentized deleted residual (10.26). Another is the use of various percentiles of the F distribution for interpretation of Cook's distance (Section 10.4). Guidelines such as these are generally not available for logistic regression as the distribution of the delta statistics is unknown except under certain restrictive assumptions. The judgment as to whether or not a case is outlying or overly influential is typically made on the basis of a subjective visual assessment of an appropriate graphic. Usually, delta chi-square and delta deviance statistics are plotted against case number i, against

TABLE 14.11 Pearson Residuals, Studentized Pearson Residuals, Hat Diagonals, Deviance Residuals, Delta Chi-Square and Delta Deviance Statistics, and Cook's Distance—Disease Outbreak Example.

i	(1) r_{P_i}	(2) r_{SP_i}	(3) h_{ii}	(4) dev_i	(5) ΔX_i^2	(6) Δdev_i	(7) D_i
1	−0.514	−0.524	.039	−0.685	0.275	0.479	0.002
2	−0.529	−0.541	.040	−0.703	0.292	0.506	0.002
3	−0.344	−0.350	.033	−0.473	0.122	0.228	0.001
...	...	...	...	...	...	...	...
96	−0.358	−0.363	.025	−0.491	0.132	0.245	0.001
97	−0.318	−0.322	.024	−0.439	0.104	0.195	0.001
98	−0.455	−0.463	.036	−0.613	0.214	0.383	0.002

or against $\hat{\pi}_i'$. Extreme values appear as spikes when plotted against case number, or as outliers in the upper corners of the plot when plotted against $\hat{\pi}_i$ or $\hat{\pi}_i'$.

Example

Table 14.11 lists in columns 1–6 for a portion of the disease outbreak data the Pearson residuals r_{P_i}, the studentized Pearson residuals r_{SP_i}, the hat matrix diagonal elements h_{ii}, the deviance residuals, dev_i, the delta chi-square statistics ΔX_i^2, and the delta deviance residuals Δdev_i. We illustrate the calculations needed to obtain ΔX_i^2, and Δdev_i, for the first case. As noted in (14.85) the first delta chi-square statistic is given by the square of the first studentized Pearson residual:

$$\Delta X_1^2 = r_{SP_1}^2 = (-.524)^2 = .275$$

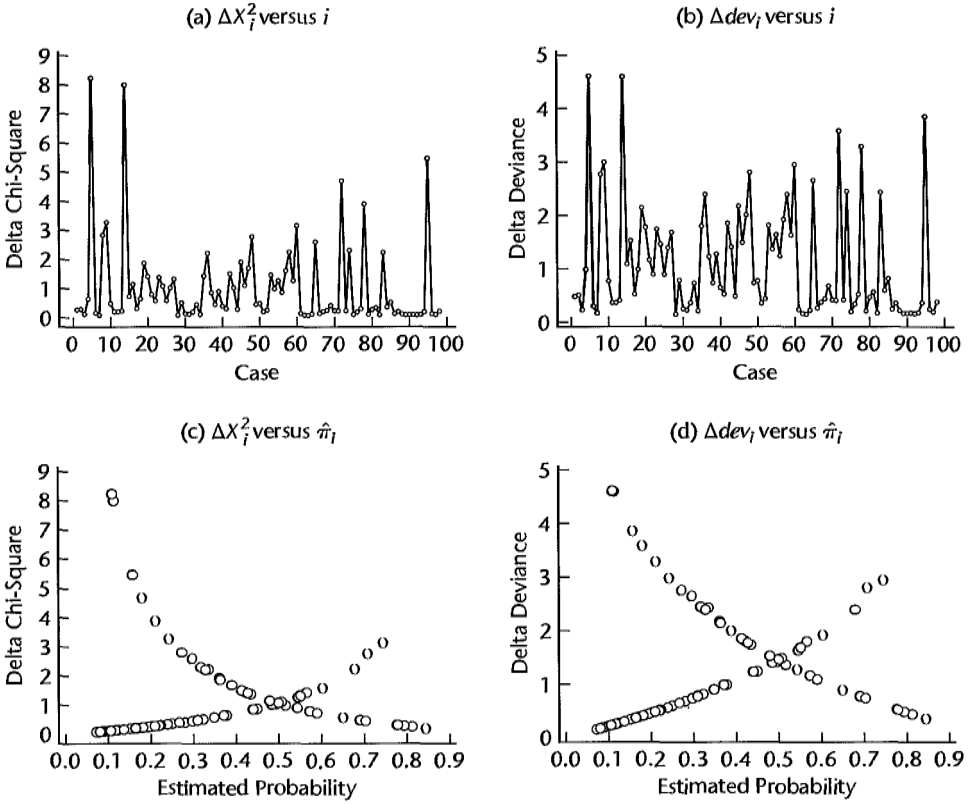
Using (14.86) with $h_{11} = .039$ and $dev_1 = -.685$ from columns 3 and 4 of Table 14.11, the first delta deviance statistic is:

$$\Delta dev_1 = h_{11} r_{SP_1}^2 + dev_1^2 = .039(-.524)^2 + (-.685)^2 = .479$$

Figures 14.15a and 14.15b provide index plots of the delta chi-square and delta deviance statistics for the disease outbreak example. The two spikes corresponding to cases 5 and 14 indicate clearly that these cases have the largest values of the delta deviance and delta chi-square statistics. Shown just below each of these in Figures 14.15c and 14.15d are plots of the delta chi-square and delta deviance statistics against the model-estimated probabilities. Note that cases 5 and 14 again stand out—this time in the upper left corner of the plot. The results suggest that cases 5 and 14 may substantively affect the conclusions. The cases were therefore flagged for potential remedial action at a later stage of the analysis.

Influence on the Fitted Linear Predictor: Cook's Distance. In Chapter 10, we introduced Cook's distance statistic, D_i, for the identification of influential observations. We noted that for the standard regression case D_i measures the standardized change in the fitted response vector $\hat{\mathbf{Y}}$ when the ith case is deleted. Similarly, Cook's distance for logistic regression measures the standardized change in the linear predictor $\hat{\pi}_i$ when the ith case is deleted. Like the delta statistics described above, obtaining these values exactly requires n maximizations of the likelihood. Instead, the following one-step approximation is used

FIGURE 14.15 Delta Chi-Square and Delta Deviance Plots—Disease Outbreak Example.



(Reference 14.5):

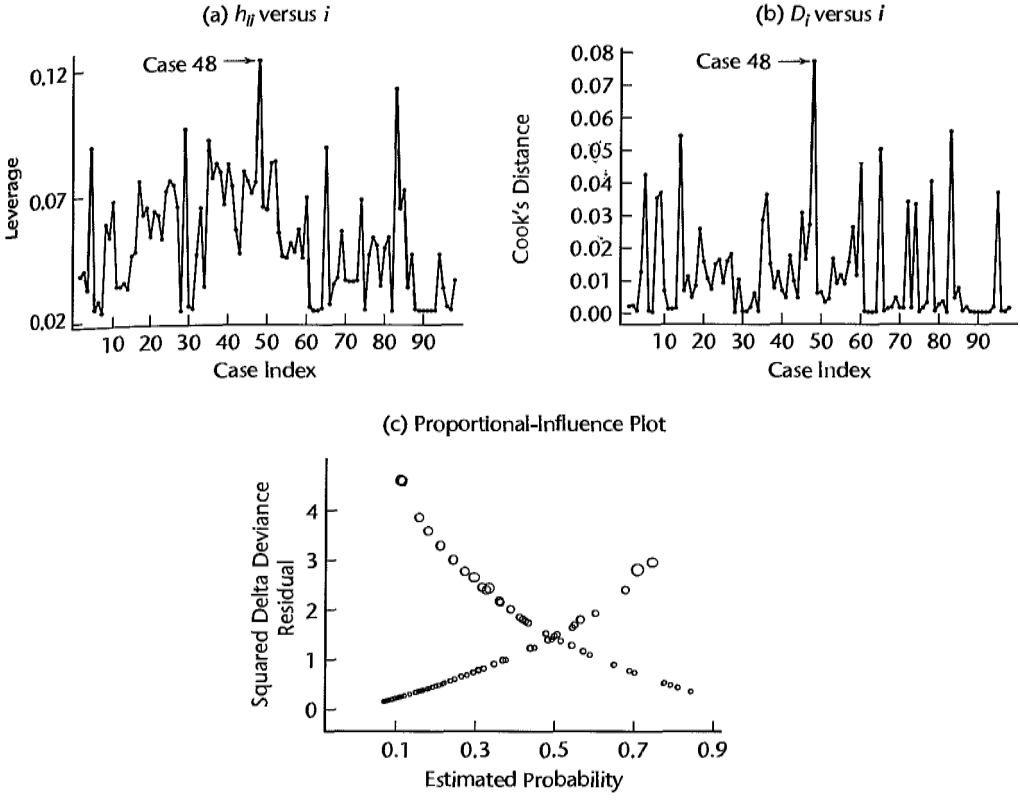
$$D_i = \frac{r_{P_i}^2 h_{ii}}{p(1 - h_{ii})^2} \tag{14.87}$$

Index plots of leverage values h_{ii} are useful for identifying outliers in the X space, and index plots of D_i can be used to identify cases that have a large effect on the fitted linear predictor. As was the case with the delta chi-square and delta deviance statistics, rules of thumb for judging the magnitudes of these diagnostics are not available, and we must rely on a visual assessment of an appropriate graphic. Note that influence on both the deviance (or Pearson chi-square) statistic and the linear predictor can be assessed simultaneously using a *proportional influence* or *bubble* plot of the delta deviance (or delta chi-square) statistics, in which the area of the plot symbol is proportional to D_i.

Example

Cook's distances are listed in column 7 of Table 14.11 for a portion of the disease outbreak example. To illustrate the calculation of Cook's distance we again focus on the first case. We require $h_{11} = .039$, $r_{P_1} = -.514$ from columns 1 and 3 of Table 14.11. Then, we have

FIGURE 14.16 Index Plots of Leverage Values, Cook's Distances, and Proportional-Influence Plot of Delta Deviance Statistic—Disease Outbreak Example.



from (14.87) with $p = 5$:

$$D_1 = \frac{r_{Pi}^2 h_{ii}}{p(1-h_{ii})^2} = \frac{(-.514)^2(.039)}{5(1-.039)^2} = .0022$$

Figures 14.16a–c display an index plot of h_{ii}, an index plot of D_i, and a proportional-influence plot of the delta deviance statistics. The leverage plot identifies case 48 as being somewhat outlying in the X space—and therefore potentially influential—and the plot of Cook's distances indicates that case 48 is indeed the most influential in terms of effect on the linear predictor. Note that cases 5 and 14—previously identified as most influential in terms of their effect on the Pearson chi-square and deviance statistics—have relatively less influence on the linear predictor. This is shown also by the proportional-influence plot in Figure 14.16c. These two cases, which have the largest delta deviance values, are located in the upper left region of the plot. The plot symbols for these cases are not overly large, indicating that these cases are not particularly influential in terms of the fitted linear predictor values. Case 48 was temporarily deleted and the logistic regression fit was obtained (not shown). The results were not appreciably different from those obtained from the full data set, and the case was retained.

Frequently, estimation of the probability π for one or several different sets of values of the predictor variables is required. In the disease outbreak example, for instance, there may be interest in the probability of 10-year-old persons of lower socioeconomic status living in city sector 1 having contracted the disease.

Point Estimator

As usual, we denote the vector of the levels of the X variables for which π is to be estimated by $\mathbf{X}_h$:

$$\underset{p\times 1}{\mathbf{X}_h} = \begin{bmatrix} 1 \\ X_{h1} \\ X_{h2} \\ \vdots \\ X_{h,p-1} \end{bmatrix} \tag{14.88}$$

and the mean response of interest by π_h:

$$\pi_h = [1 + \exp(-\mathbf{X}_h'\beta)]^{-1} \tag{14.89}$$

The point estimator of π_h will be denoted by $\hat{\pi}_h$ and is as follows:

$$\hat{\pi}_h = [1 + \exp(-\mathbf{X}_h'\mathbf{b})]^{-1} \tag{14.90}$$

where $\mathbf{b}$ is the vector of estimated regression coefficients in (14.43).

Interval Estimation

We obtain a confidence interval for π_h in two stages. First, we calculate confidence limits for the logit mean response π_h'. Then we use the relation (14.38a) to obtain confidence limits for the mean response π_h. To see this clearly, we consider (14.38a) for $\mathbf{X} = \mathbf{X}_h$:

$$E\{Y_h\} = [1 + \exp(-\mathbf{X}_h'\beta)]^{-1}$$

and restate the expression by using the fact that $E\{Y_h\} = \pi_h$ and $\mathbf{X}_h'\beta = \pi_h'$:

$$\pi_h = [1 + \exp(-\pi_h')]^{-1} \tag{14.91}$$

It is this relation in (14.91) that we utilize to convert confidence limits for π_h' into confidence limits for π_h.

The point estimator of the logit mean response $\pi_h' = \mathbf{X}_h'\beta$ is $\hat{\pi}_h' = \mathbf{X}_h'\mathbf{b}$. The estimated approximate variance of $\hat{\pi}_h' = \mathbf{X}_h'\mathbf{b}$ according to (5.46) is:

$$s^2\{\hat{\pi}_h'\} = s^2\{\mathbf{X}_h'\mathbf{b}\} = \mathbf{X}_h'\mathbf{s}^2\{\mathbf{b}\}\mathbf{X}_h \tag{14.92}$$

where $\mathbf{s}^2\{\mathbf{b}\}$ is the estimated approximate variance-covariance matrix of the regression coefficients in (14.51) when n is large.

Approximate $1-\alpha$ large-sample confidence limits for the logit mean response π_h' are then obtained in the usual fashion:

$$L = \hat{\pi}_h' - z(1-\alpha/2)s\{\hat{\pi}_h'\} \tag{14.93a}$$

$$U = \hat{\pi}_h' + z(1-\alpha/2)s\{\hat{\pi}_h'\} \tag{14.93b}$$

Here, L and U are, respectively, the lower and upper confidence limits for π_h'.

Finally, we use the monotonic relation between π_h and π_h' in (14.91) to convert the confidence limits L and U for π_h' into approximate $1-\alpha$ confidence limits L^* and U^* for the mean response π_h:

$$L^* = [1 + \exp(-L)]^{-1} \quad \textbf{(14.94a)}$$

$$U^* = [1 + \exp(-U)]^{-1} \quad \textbf{(14.94b)}$$

Simultaneous Confidence Intervals for Several Mean Responses

When it is desired to estimate several mean responses π_h corresponding to different $\mathbf{X}_h$ vectors with family confidence coefficient $1-\alpha$, Bonferroni simultaneous confidence intervals may be used. The procedure for g confidence intervals is the same as that for a single confidence interval except that $z(1-\alpha/2)$ in (14.93) is replaced by $z(1-\alpha/2g)$.

Example

In the disease outbreak example of Table 14.3, it is desired to find an approximate 95 percent confidence interval for the probability π_h that persons 10 years old who are of lower socioeconomic status and live in sector 1 have contracted the disease. The vector $\mathbf{X}_h$ in (14.88) here is:

$$\mathbf{X}_h = \begin{bmatrix} 1 \\ 10 \\ 0 \\ 1 \\ 0 \end{bmatrix}$$

Using the results in Table 14.4a, we obtain the point estimate of the logit mean response:

$$\begin{aligned} \hat{\pi}_h' &= \mathbf{X}_h'\mathbf{b} = -2.3129(1) + .02975(10) + .4088(0) - .30525(1) + 1.5747(0) \\ &= -2.32065 \end{aligned}$$

The estimated variance of $\hat{\pi}_h'$ is obtained by using (14.92) (calculations not shown):

$$s^2\{\hat{\pi}_h'\} = .2945$$

so that $s\{\hat{\pi}_h'\} = .54268$. For $1-\alpha = .95$, we require $z(.975) = 1.960$. Hence, the confidence limits for the logit mean response π_h' are according to (14.93):

$$L = -2.32065 - 1.960(.54268) = -3.38430$$

$$U = -2.32065 + 1.960(.54268) = -1.25700$$

Finally, we use (14.94) to obtain the confidence limits for the mean response π_h:

$$L^* = [1 + \exp(3.38430)]^{-1} = .033$$

$$U^* = [1 + \exp(1.25700)]^{-1} = .22$$

Thus, the approximate 95 percent confidence interval for the mean response π_h is:

$$.033 \leq \pi_h \leq .22$$

We therefore find, with approximate 95 percent confidence, that the probability is between .033 and .22 that 10-year-old persons of lower socioeconomic status who live in sector 1 have contracted the disease. This confidence interval is useful for indicating that persons with the specified characteristics are not subject to a very high probability of having contracted the disease, but the confidence interval is quite wide and thus not precise.

Comment

The confidence limits for π_h in (14.94) are not symmetric around the point estimate. In the disease outbreak example, for instance, the point estimate is:

$$\hat{\pi}_h = [1 + \exp(2.32065)]^{-1} = .089$$

while the confidence limits are .033 and .22. The reason for the asymmetry is that $\hat{\pi}_h$ is not a linear function of $\hat{\pi}'_h$. ■

14.10 Prediction of a New Observation

Multiple logistic regression is frequently employed for making predictions for new observations. In one application, for example, health personnel wished to predict whether a certain surgical procedure will ameliorate a new patient's condition, given the patient's age, gender, and various symptoms. In another application, marketing officials of a computer firm wished to predict whether a retail chain will purchase a new computer, on the basis of the age of the company's current computer, the company's current workload, and other factors.

Choice of Prediction Rule

Forecasting a binary outcome for given levels $\mathbf{X}_h$ of the X variables is simple in the sense that the outcome 1 will be predicted if the estimated value $\hat{\pi}_h$ is large, and the outcome 0 will be predicted if $\hat{\pi}_h$ is small. The difficulty in making predictions of a binary outcome is in determining the cutoff point, below which the outcome 0 is predicted and above which the outcome 1 is predicted. A variety of approaches are possible to determine where this cutoff point is to be located. We consider three approaches.

1. *Use .5 as the cutoff.* With this approach, the prediction rule is:

   If $\hat{\pi}_h$ exceeds .5, predict 1; otherwise predict 0.

This approach is reasonable when (a) it is equally likely in the population of interest that outcomes 0 and 1 will occur; and (b) the costs of incorrectly predicting 0 and 1 are approximately the same.

2. *Find the best cutoff for the data set on which the multiple logistic regression model is based.* This approach involves evaluating different cutoffs. For each cutoff, the rule is employed on the n cases in the model-building data set and the proportion of cases incorrectly predicted is ascertained. The cutoff for which the proportion of incorrect predictions is lowest is the one to be employed.

This approach is reasonable when (a) the data set is a random sample from the relevant population, and thus reflects the proper proportions of 0s and 1s in the population, and (b) the costs of incorrectly predicting 0 and 1 are approximately the same. The proportion of incorrect predictions observed for the optimal cutoff is likely to be an overstatement of the ability of the cutoff to correctly predict new observations, especially if the model-building data set is not large. The reason is that the cutoff is chosen with reference to the same data set from which the logistic model was fitted and thus is best for these data only. Consequently, as we explained in Chapter 9, it is important that a validation data set be employed to indicate whether the observed predictive ability for a fitted regression model is a valid indicator for predicting new observations.

3. *Use prior probabilities and costs of incorrect predictions in determining the cutoff.* When prior information is available about the likelihood of 1s and 0s in the population and the data set is not a random sample from the population, the prior information can be used in finding an optimal cutoff. In addition, when the cost of incorrectly predicting outcome 1 differs substantially from the cost of incorrectly predicting outcome 0, these costs of incorrect consequences can be incorporated into the determination of the cutoff so that the expected cost of incorrect predictions will be minimized. Specialized references, such as Reference 14.7, discuss the use of prior information and costs of incorrect predictions for determining the optimal cutoff.

Example

We shall use the disease outbreak example of Table 14.3 to illustrate how to obtain the cutoff point for predicting a new observation, even though the main purpose of that study was to determine whether age, socioeconomic status, and city sector are important risk factors. We assume that the cost of incorrectly predicting that a person has contracted the disease is about the same as the cost of incorrectly predicting that a person has not contracted the disease. The estimated logistic response function is given in (14.46).

Since a random sample of individuals was selected in the two city sectors, the 98 cases in the study constitute a cross section of the relevant population. Consequently, information is provided in the sample about the proportion of persons who have contracted the disease in the population. Of the 98 persons in the study, 31 had contracted the disease (see the disease outbreak data set in Appendix C.10); hence the estimated proportion of persons who had contracted the disease is $31/98 = .316$. This proportion can be used as the starting point in the search for the best cutoff in the prediction rule.

Thus, the first rule investigated was:

$$\text{Predict 1 if } \hat{\pi}_h \geq .316\text{; predict 0 if } \hat{\pi}_h < .316 \qquad \textbf{(14.95)}$$

Note from Table 14.3, column 6, that $\hat{\pi}_1 = .209$ for case 1; hence prediction rule (14.95) calls for a prediction that the person has not contracted the disease. This would be a correct prediction. Similarly, prediction rule (14.95) would correctly predict cases 2 and 3 not to have contracted the disease. However, the prediction with rule (14.95) for case 4 (person has contracted the disease because $\hat{\pi}_4 = .371 \geq .316$) would be incorrect. Similarly, the prediction for case 5 (person has not contracted the disease because $\hat{\pi}_5 = .111 < .316$) would be incorrect. Table 14.12a provides a summary of the number of correct and incorrect classifications based on prediction rule (14.95). Of the 67 persons without the disease, 20 would be incorrectly predicted to have contracted the disease, or an error rate of 29.9 percent.

TABLE 14.12 **Classification Based on Logistic Response Function (14.46) and Prediction Rules (14.95) and (14.96)—Disease Outbreak Example.**

True Classification	(a) Rule (14.95)			(b) Rule (14.96)		
	$\hat{Y} = 0$	$\hat{Y} = 1$	Total	$\hat{Y} = 0$	$\hat{Y} = 1$	Total
$Y = 0$	47	20	67	50	17	67
$Y = 1$	8	23	31	9	22	31
Total	55	43	98	59	39	98

Of the 31 persons with the disease, eight would be incorrectly predicted with rule (14.95) not to have contracted the disease, or 25.8 percent. Altogether, $20 + 8 = 28$ of the 98 predictions would be incorrect, so that the prediction error rate for rule (14.95) is $28/98 = .286$ or 28.6 percent.

Similar analyses were made for other cutoff points and it appears that among the cutoffs considered, use of the following rule may be best:

$$\text{Predict 1 if } \hat{\pi}_h \geq .325\text{; predict 0 if } \hat{\pi}_h < .325 \tag{14.96}$$

Table 14.12b provides a summary of the correct and incorrect classifications based on prediction rule (14.96). The prediction error rate for this rule is $(9 + 17)/98 = .265$ or 26.5 percent. Note also that for this rule, the error rates for persons with and without the disease (9/31 and 17/67) are quite close to each other. Thus, the risks of incorrect predictions for the two groups are fairly balanced, which is often desirable. Note also that the error rates for persons with and without the disease are much less balanced as the cutoff is shifted further away from the optimal one in either direction.

An effective way to display this information graphically is through the *receiver operating characteristic* (ROC) *curve*, which plots $P(\hat{Y} = 1|Y = 1)$ (also called *sensitivity*) as a function of $1 - P(\hat{Y} = 0|Y = 0)$ (also called *1−specificity*) for the possible cutpoints $\hat{\pi}_h$. Figure 14.17 exhibits the ROC curve for model (14.46) for all possible cutpoints between 0 and 1. (See A.7a for the definition of conditional probability.)

To see how a single point on the ROC curve in Figure 14.17 is determined, we consider rule (14.95), for which the cutoff is .316. From Table 14.12a, the *sensitivity* is:

$$P(\hat{Y} = 1|Y = 1) = \frac{23}{31} = .74$$

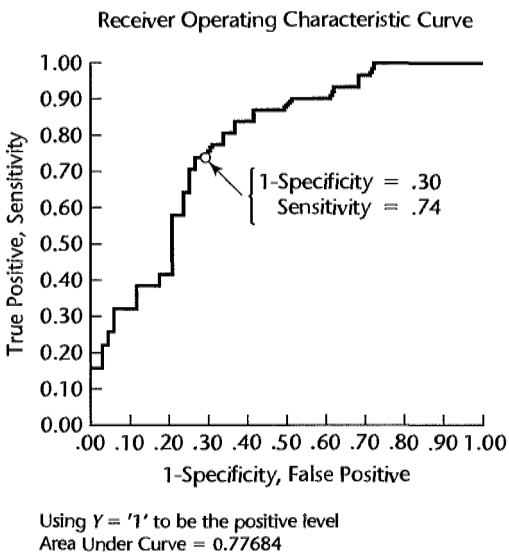


FIGURE 14.17 JMP ROC Curve—Disease Outbreak Example.

Also, 1−*specificity* here is:

$$1 - P(\hat{Y} = 0|Y = 0) = 1 - \frac{47}{67} = .30$$

This point is highlighted on the ROC curve in Figure 14.17.

The area under the ROC curve is a useful summary measure of the model's predictive power and is identical to the *concordance index*. Consider any pair of observations (i, j) such that $Y_i = 1$ and $Y_j = 0$. Since $Y_i > Y_j$, this pair is said to be concordant if $\hat{\pi}_i > \hat{\pi}_j$. The concordance index estimates the probability that the predictions and the outcomes are concordant (Reference 14.2). A value of 0.5 means that the predictions were no better than random guessing. For the disease outbreak model (14.96), the ROC area is 0.777.

A validation study will now be required to determine whether the observed prediction error rate for the optimal cutoff properly indicates the risks of incorrect predictions for new observations, or whether it seriously understates them. In any case, it appears already that fitted logistic regression model (14.96) may not be too useful as a predictive model because of the relatively high risks of making incorrect predictions.

Comment

A limitation of the prediction rule approach is that it dichotomizes a continuous predictor $\hat{\pi}$ where the choice of cutpoint $\hat{\pi}_h$ is arbitrary and is highly dependent upon the relative frequencies of 1s and 0s observed in the sample. ■

ation of Prediction Error Rate

The reliability of the prediction error rate observed in the model-building data set is examined by applying the chosen prediction rule to a validation data set. If the new prediction error rate is about the same as that for the model-building data set, then the latter gives a reliable indication of the predictive ability of the fitted logistic regression model and the chosen prediction rule. If the new data lead to a considerably higher prediction error rate, then the fitted logistic regression model and the chosen prediction rule do not predict new observations as well as originally indicated.

ple

In the disease outbreak example, the fitted logistic regression function (14.46) based on the model-building data set:

$$\hat{\pi} = [1 + \exp(-3.8877 - .02975X_1 - .4088X_2 + .30525X_3 - 1.5747X_4)]^{-1}$$

was used to calculate estimated probabilities $\hat{\pi}_h$ for cases 99-196 in the disease outbreak data set in Appendix C.10. These cases constitute the validation data set. The chosen prediction rule (14.96):

$$\text{Predict 1 if } \hat{\pi}_h \geq .325\text{; predict 0 if } \hat{\pi}_h < .325$$

was then applied to these estimated probabilities. The percent prediction error rates were as follows:

Disease Status		
With Disease	Without Disease	Total
46.2	38.9	40.8

Note that the total prediction error rate of 40.8 percent is considerably higher than the 26.5 percent error rate based on the model-building data set. The latter therefore is not a reliable indicator of the predictive capability of the fitted logistic regression model and the chosen prediction rule.

We should mention again that making predictions was not the primary objective in the disease outbreak study. Rather, the main purpose was to identify key explanatory variables. Still, the prediction error rate for the validation data set shows that there must be other key explanatory variables affecting whether a person has contracted the disease that have not yet been identified for inclusion in the logistic regression model.

Comment

An alternative to multiple logistic regression for predicting a binary response variable when the predictor variables are continuous is *discriminant analysis*. This approach assumes that the predictor variables follow a joint multivariate normal distribution. Discriminant analysis can also be used when this condition is not met, but the approach is not optimal then and logistic regression frequently is preferable. The reader is referred to Reference 14.8 for an in-depth discussion of discriminant analysis. ■

14.11 Polytomous Logistic Regression for Nominal Response

Logistic regression is most frequently used to model the relationship between a dichotomous response variable and a set of predictor variables. On occasion, however, the response variable may have more than two levels. Logistic regression can still be employed by means of a *polytomous*—or *multicategory*—logistic regression model. Polytomous logistic regression models are used in many fields. In business, for instance, a market researcher may wish to relate a consumer's choice of product (product A, product B, product C) to the consumer's age, gender, geographic location, and several other potential explanatory variables. This is an example of *nominal* polytomous regression, because the response categories are purely qualitative and not ordered in any way. *Ordinal* response categories can also be modeled using polytomous regression. For example, the relation between severity of disease measured on an ordinal scale (mild, moderate, severe) and age of patient, gender of patient, and some other explanatory variables may be of interest. We consider ordinal polytomous logistic regression in detail in Section 14.12.

In this section we discuss the use of polytomous logistic regression for nominal multicategory responses. Throughout, we will use the pregnancy duration example, introduced in Section 14.2 in the context of binary logistic regression, to illustrate concepts. This time, however, the response will have more than two categories.

Pregnancy Duration Data with Polytomous Response

A study was undertaken to determine the strength of association between several risk factors and the duration of pregnancies. The risk factors considered were mother's age, nutritional status, history of tobacco use, and history of alcohol use. The response of interest, pregnancy duration, is a three-category variable that was coded as follows:

Y_i	Pregnancy Duration Category
1	Preterm (less than 36 weeks)
2	Intermediate term (36 to 37 weeks)
3	Full term (38 weeks or greater)

Relevant data for 102 women who had recently given birth at a large metropolitan hospital were obtained. A portion of these data is displayed in Table 14.13. The polytomous response, pregancy duration (Y), is shown in column 1. Nutritional status (X_1), shown in column 5, is an index of nutritional status (higher score denotes better nutritional status). The predictor variable age was categorized into three groups: less than 20 years of age (coded 1), from 21 to 30 years of age (coded 2), and greater than 30 years of age (coded 3). It is represented by two indicator variables (X_2 and X_3), shown in columns 6 and 7 of Table 14.13, as follows:

Class	X_2	X_3
Less than or equal to 20 years of age	1	0
21 to 30 years of age	0	0
Greater than 30 years of age	0	1

(The researchers chose the middle category—21 to 30 years of age—as the referent category for this qualitative predictor because mothers in this age group tend to have the lowest risk of preterm deliveries. This leads to positive regression coefficients for these predictors, and a slightly simpler interpretation.) Alcohol and smoking history were also qualitative predictors; the categories were "Yes" (coded 1) and "No" (coded 0). Alcohol use history (X_4), and smoking history (X_5) are listed in columns 8 and 9 of Table 14.13.

TABLE 14.13 Data—Pregnancy Duration Example with Polytomous Response.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	Duration	Response Category			Nutritional Status	Age-Category		Alcohol Use History	Smoking History
	Y_i	Y_{i1}	Y_{i2}	Y_{i3}	X_{i1}	X_{i2}	X_{i3}	X_{i4}	X_{i5}
	1	1	0	0	150	0	0	0	1
	1	1	0	0	124	1	0	0	0
	1	1	0	0	128	0	0	0	1
	...	...	...	...	...	...	...	...	...
	3	0	0	1	117	0	0	1	1
	3	0	0	1	165	0	0	1	1
	3	0	0	1	134	0	0	1	1

Because pregnancy duration is a qualitative variable with three categories, we will create three binary response variables, one for each response category as follows:

$$Y_{i1} = \begin{cases} 1 & \text{if case } i \text{ response is category 1} \\ 0 & \text{otherwise} \end{cases}$$

$$Y_{i2} = \begin{cases} 1 & \text{if case } i \text{ response is category 2} \\ 0 & \text{otherwise} \end{cases}$$

$$Y_{i3} = \begin{cases} 1 & \text{if case } i \text{ response is category 3} \\ 0 & \text{otherwise} \end{cases}$$

These three coded variables are also included in Table 14.13 in columns 2, 3, and 4. Note that because $Y_{i1} + Y_{i2} + Y_{i3} = 1$, the value of any one of these three binary variables can be determined from the other two. For example, $Y_{i3} = 1 - Y_{i1} - Y_{i2}$.

We first treat pregnancy duration as a nominal response, ignoring the time-based ordering of the categories; later we will show how a more parsimonious model results when we treat pregnancy duration as an ordinal response.

$J-1$ Baseline-Category Logits for Nominal Response

In general, we will assume there are J response categories. Then for the ith observation, there will be J binary response variables, $Y_{i1}, \ldots, Y_{iJ}$, where:

$$Y_{ij} = \begin{cases} 1 & \text{if case } i \text{ response is category } j \\ 0 & \text{otherwise} \end{cases}$$

Since only one category can be selected for response i, we have:

$$\sum_{j=1}^{J} Y_{ij} = 1$$

We will require some additional notation for the multicategory case. First, let π_{ij} denote the probability that category j is selected for the ith response. Then:

$$\pi_{ij} = P(Y_{ij} = 1)$$

In the binary case, $J = 2$. Suppose that we code $Y_i = 1$ if the ith response is category 1, and we code $Y_i = 0$ if the ith response is category 2. Then:

$$\pi_i = \pi_{i1} \quad \text{and} \quad 1 - \pi_i = \pi_{i2}$$

For binary logistic regression, we model the logit of π_i using the linear predictor. Since there are only two categories in binary logistic regression, the logit in fact compares the probability of a category-1 response to the probability of a category-2 response:

$$\pi_i' = \log_e\left[\frac{\pi_i}{1 - \pi_i}\right] = \log_e\left[\frac{\pi_{i1}}{\pi_{i2}}\right] = \pi_{i12}' = \mathbf{X}_i'\boldsymbol{\beta}_{12}$$

Note that we have used π_{i12}' and $\boldsymbol{\beta}_{12}$ to emphasize that the linear predictor is modeling the logarithm of the ratio of the probabilities for categories 1 and 2.

Now for the J polytomous categories, there are $J(J-1)/2$ pairs of categories, and therefore $J(J-1)/2$ linear predictors. For example, for the pregnancy duration data,

$J = 3$ and we have $3(3-1)/2 = 3$ comparisons:

$$\pi'_{i12} = \log_e\left[\frac{\pi_{i1}}{\pi_{i2}}\right] = \mathbf{X}'_i\boldsymbol{\beta}_{12}$$

$$\pi'_{i13} = \log_e\left[\frac{\pi_{i1}}{\pi_{i3}}\right] = \mathbf{X}'_i\boldsymbol{\beta}_{13}$$

$$\pi'_{i23} = \log_e\left[\frac{\pi_{i2}}{\pi_{i3}}\right] = \mathbf{X}'_i\boldsymbol{\beta}_{23}$$

Fortunately, it is not necessary to develop all $J(J-1)/2$ logistic regression models. One category will be chosen as the *baseline* or *referent* category, and then all other categories will be compared to it. The choice of baseline or referent category is arbitrary. Frequently the last category is chosen and, indeed, this is usually the default choice for statistical software programs. One exception to this may be found in epidemiological studies, where the category having the lowest risk is often used as the referent category.

Using category J to denote the baseline category, we need consider only the $J-1$ comparisons to this referent category. The logit for the jth such comparison is:

$$\pi'_{ijJ} = \log_e\left[\frac{\pi_{ij}}{\pi_{iJ}}\right] = \mathbf{X}'_i\boldsymbol{\beta}_{jJ} \qquad j = 1, 2, \ldots, J-1 \tag{14.97a}$$

Since it is understood that comparisons are always made to category J, we let $\pi'_{ij} = \pi'_{ijJ}$ and $\boldsymbol{\beta}_j = \boldsymbol{\beta}_{jJ}$ in (14.97a), giving:

$$\pi'_{ij} = \log_e\left[\frac{\pi_{ij}}{\pi_{iJ}}\right] = \mathbf{X}'_i\boldsymbol{\beta}_j \qquad j = 1, 2, \ldots, J-1 \tag{14.97b}$$

The reason that we need to consider only these $J-1$ logits is that the logits for any other comparisons can be obtained from them. To see this, suppose $J = 4$, and we wish to compare categories 1 and 2. Then:

$$\begin{aligned}\log_e\left[\frac{\pi_{i1}}{\pi_{i2}}\right] &= \log_e\left[\frac{\pi_{i1}}{\pi_{i4}} \times \frac{\pi_{i4}}{\pi_{i2}}\right] \\ &= \log_e\left[\frac{\pi_{i1}}{\pi_{i4}}\right] - \log_e\left[\frac{\pi_{i2}}{\pi_{i4}}\right] \\ &= \mathbf{X}'_i\boldsymbol{\beta}_1 - \mathbf{X}'_i\boldsymbol{\beta}_2\end{aligned}$$

In general, to compare categories k and l, we have:

$$\log_e\left[\frac{\pi_{ik}}{\pi_{il}}\right] = \mathbf{X}'_i(\boldsymbol{\beta}_k - \boldsymbol{\beta}_l) \tag{14.98}$$

Given the $J-1$ logit expressions in (14.98) it is possible (algebra not shown) to obtain the $J-1$ direct expressions for the category probabilities in terms of the $J-1$ linear predictors, $\mathbf{X}'\boldsymbol{\beta}_j$. The resulting expressions are:

$$\pi_{ij} = \frac{\exp(\mathbf{X}'_i\boldsymbol{\beta}_j)}{1 + \sum_{k=1}^{J-1}\exp(\mathbf{X}'_i\boldsymbol{\beta}_k)} \qquad j = 1, 2, \ldots, J-1 \tag{14.99}$$

We next consider methods for obtaining estimates of the $J-1$ parameter vectors $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_{J-1}$.

Maximum Likelihood Estimation

There are two approaches commonly used for obtaining estimates of the parameter vectors, $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_{J-1}$; both employ maximum likelihood estimation. With the first approach, separate binary logistic regressions are carried out for each of the $J-1$ comparisons to the baseline category. For example, to estimate $\boldsymbol{\beta}_1$, we drop from the data set all cases except those for which either $Y_{i1} = 1$ or $Y_{iJ} = 1$. Since only two categories are then present, we can apply binary logistic regression directly. This approach is particularly useful when statistical software is not available for multicategory logistic regression (Reference 14.9).

A more effective approach from a statistical viewpoint is to obtain estimates of the $J-1$ logits simultaneously. To do so, we require the likelihood for the full data set. To fix ideas, suppose that there are $J = 4$ categories and that the third category is selected for the ith response. That is, for case i we have:

$$Y_{i1} = 0 \qquad Y_{i2} = 0 \qquad Y_{i3} = 1 \qquad Y_{i4} = 0$$

The probability of this response is:

$$\begin{aligned} P(Y_i = 3) &= \pi_{i3} \\ &= [\pi_{i1}]^0 \times [\pi_{i2}]^0 \times [\pi_{i3}]^1 \times [\pi_{i4}]^0 \\ &= \prod_{j=1}^{4} [\pi_{ij}]^{Y_{ij}} \end{aligned}$$

For n independent observations and J categories, it is easily seen that the likelihood is:

$$P(Y_1, \ldots, Y_n) = \prod_{i=1}^{n} P(Y_i) = \prod_{i=1}^{n} \left[\prod_{j=1}^{J} [\pi_{ij}]^{Y_{ij}} \right] \tag{14.100}$$

It can be shown that the log likelihood is given by:

$$\log_e[P(Y_1, \ldots, Y_n)] = \sum_{i=1}^{n} \left(\sum_{j=1}^{J-1} (Y_{ij} \mathbf{X}_i' \boldsymbol{\beta}_j) - \log_e \left[1 + \sum_{j=1}^{J-1} \exp(\mathbf{X}_i' \boldsymbol{\beta}_j) \right] \right) \tag{14.101}$$

The maximum likelihood estimates of $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_{J-1}$ are those values, $\mathbf{b}_1, \ldots, \mathbf{b}_{J-1}$, that maximize (14.101). As usual, we will rely on standard statistical software programs to obtain these estimates.

As was the case for binary logistic regression, the $J-1$ fitted response functions may be obtained by substituting the maximum likelihood estimates of the $J-1$ parameter vectors into the expression in (14.99):

$$\hat{\pi}_{ij} = \frac{\exp(\mathbf{X}_i' \mathbf{b}_j)}{1 + \sum_{k=1}^{J-1} \exp(\mathbf{X}_i' \mathbf{b}_k)} \tag{14.102}$$

We turn now to an example to illustrate the analysis and interpretation of a nominal-level polytomous logistic regression model.

Example

For the pregnancy duration data in Table 14.13, a set of $J-1=2$ first-order linear predictors was initially proposed:

$$\log_e\left[\frac{\pi_{ij}}{\pi_{i3}}\right] = \mathbf{X}_i'\boldsymbol{\beta}_j \qquad \text{for } j = 1, 2$$

MINITAB's nominal logistic regression output is displayed in Figure 14.18. It first indicates that the response had three levels, 1, 2, and 3, and that the referent response event is $Y_i = 3$. Following this summary is the logistic regression table, which contains the estimated regression coefficients, estimated approximate standard errors, the Wald test statistics and P-values, the estimated odds ratios for the two estimated linear predictors, and the 95 percent confidence intervals for the odds ratios. The maximum likelihood estimates of $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ are:

$$\mathbf{b}_1 = \begin{bmatrix} 3.958 \\ -0.0464 \\ 2.9135 \\ 1.8875 \\ 1.0670 \\ 2.2305 \end{bmatrix} \qquad \mathbf{b}_2 = \begin{bmatrix} 5.475 \\ -0.0654 \\ 2.9570 \\ 2.0597 \\ 2.0429 \\ 2.4524 \end{bmatrix}$$

Before using the fitted model to make inferences, various regression diagnostics similar to those already discussed for binary logistic regression should be examined. In polytomous logistic regression, the multiple outcome categories make this a more difficult problem

FIGURE 14.18
MINITAB Nominal Logistic Regression Outpnt—Pregnancy Duration Example.

Polytomous Nominal MTB Output

Response Information

Variable	Value	Count	
preterm	3	41	(Reference Event)
	2	35	
	1	26	
	Total	102	

Logistic Regression Table

Predictor	Coef	SE Coef	Z	P	Odds Ratio	95% CI Lower	95% CI Upper
Logit 1: (2/3)							
Constant	3.958	1.941	2.04	0.041			
nutritio	-0.04645	0.01489	-3.12	0.002	0.95	0.93	0.98
agecat1	2.9135	0.8575	3.40	0.001	18.42	3.43	98.91
agecat3	1.8875	0.8088	2.33	0.020	6.60	1.35	32.23
alcohol	1.0670	0.6495	1.64	0.100	2.91	0.81	10.38
smoking	2.2305	0.6682	3.34	0.001	9.30	2.51	34.47
Logit 2: (1/3)							
Constant	5.475	2.272	2.41	0.016			
nutritio	-0.06542	0.01824	-3.59	0.000	0.94	0.90	0.97
agecat1	2.9570	0.9645	3.07	0.002	19.24	2.91	127.41
agecat3	2.0597	0.8947	2.30	0.021	7.84	1.36	45.30
alcohol	2.0429	0.7097	2.88	0.004	7.71	1.92	31.00
smoking	2.4524	0.7315	3.35	0.001	11.62	2.77	48.72

Log-likelihood = -84.338
Test that all slopes are zero: G = 52.011, DF = 10, P-Value = 0.000

than was the case for binary logistic regression. We thus recommend assessing the fit and monitoring logistic regression diagnostics using the $J - 1$ individual binary logistic regressions, as described in the first paragraph on page 612. Hence, we would assess the fit of the two logistic regression models separately, and then make a statement about the fit of the polytomous logistic model descriptively. Diagnostics, including the Hosmer-Lemeshow test for goodness of fit, simulated envelopes for deviance residuals, and plots of influence statistics were examined for the pregnancy duration data, and no serious departures were found (results not shown). We turn now to model interpretation and inference.

As indicated in Figure 14.18, all Wald test P-values are less than .05—with the exception of alcohol in the first linear predictor—indicating that all of the predictors should be retained. In all cases, the direction of the association between the predictors and the estimated logits, as indicated by the signs of the estimated regression coefficients, were as expected.

For teenagers, the estimated odds of delivering preterm compared to full term are 18.42 times the estimated odds for women 20–30 years of age; the 95% confidence interval for this odds ratio has a lower limit of 3.43 and an upper limit of 98.91. Thus while the age effect is estimated to be very large, there is considerable uncertainty in the estimate. Similarly, the estimated odds for teenagers of delivering intermediate term compared to full term are 19.24; the lower 95% confidence limit is 2.91 and the upper limit is 127.41. History of smoking, history of alcohol use, and being in the 30-and-over age category also increase the estimated odds of delivering preterm or intermediate term compared to full term, though less dramatically. The negative estimated coefficients for nutritional status indicate that a lower nutritional status is associated with increased odds of delivering preterm or intermediate term compared to full term.

Comment

To derive expression (14.101) for the log likelihood, we first obtain the logarithm of (14.100) and let $\pi_{iJ} = 1 - \sum_{j=1}^{J-1} \pi_{ij}$ and $Y_{iJ} = 1 - \sum_{j=1}^{J-1} Y_{ij}$. It follows that:

$$\begin{aligned}
\log_e P(Y_1, \ldots, Y_n) &= \sum_{i=1}^{n} \left(\sum_{j=1}^{J-1} Y_{ij} \log_e[\pi_{ij}] + \left(1 - \sum_{j=1}^{J-1} Y_{ij}\right) \log_e \left[1 - \sum_{j=1}^{J-1} \pi_{ij}\right] \right) \\
&= \sum_{i=1}^{n} \left(\sum_{j=1}^{J-1} Y_{ij} \log_e[\pi_{ij}] + \log_e \left[1 - \sum_{j=1}^{J-1} \pi_{ij}\right] - \sum_{j=1}^{J-1} Y_{ij} \log_e \left[1 - \sum_{j=1}^{J-1} \pi_{ij}\right] \right) \\
&= \sum_{i=1}^{n} \left(\sum_{j=1}^{J-1} Y_{ij} \log_e \left[\frac{\pi_{ij}}{\pi_{iJ}}\right] + \log_e \left[1 - \sum_{j=1}^{J-1} \pi_{ij}\right] \right)
\end{aligned}$$

Substitution of the expressions in (14.97b) for $\log_e[\pi_{ij}/\pi_{iJ}]$ and in (14.99) for π_{ij} in the second term leads to the desired log likelihood in (14.101). ■

14.12 Polytomous Logistic Regression for Ordinal Response

Up to this point, we have considered polytomous logistic regression models for unordered categories. Categories, however, are frequently ordered. Consider the following response variables:

1. A food product is rated by consumers on a 1–10 hedonic scale.

2. In an economic study, persons are classified as either not employed, employed part time, or employed full time.
3. The quality of sheet metal produced is rated on a 1–5 scale, depending on the clarity and reflectivity of the surface.
4. Employees are asked to rate working conditions using a 7-point scale (unacceptable, poor, fair, acceptable, good, excellent, outstanding).
5. The severity of cancer is rated by stages on a 1–4 basis.

Such responses can be analyzed by using the techniques for nominal logistic regression described in Section 14.11, but a more effective strategy, yielding a more parsimonious and more easily interpreted model, results if the ordering of the categories is taken into account explicitly. The model that is usually employed is called the *proportional odds model.*

To motivate this model, we revisit the pregnancy duration example. We will assume that pregnancy duration is a continuous response denoted by Y_i^c. For ease of exposition, we will also assume that there is just one (quantitiative) predictor, nutrition index, X_{i1}. Assume that Y_i^c can be represented by the simple linear regression model:

$$Y_i^c = \beta_0^* + \beta_1^* X_{i1} + k\varepsilon_L$$

where ε_L follows the standard logistic distribution (14.14) with mean zero and standard deviation $\pi/\sqrt{3}$, and k is a constant that satisfies:

$$\sigma\{Y_i^c\} = k\sigma\{\varepsilon_L\} = k\frac{\pi}{\sqrt{3}}$$

Researchers were interested in specific categories of pregnancy delivery time and therefore discretized pregnancy duration Y_i^c using the following upperbounds or cutpoints for each category:

Y_i	Category	Y_i^c	Cutpoint T
1	Preterm	$0 \le Y_i^c < 36$ weeks	$T_1 = 36$ weeks
2	Intermediate term	36 weeks $\le Y_i^c <$ 38 weeks	$T_2 = 38$ weeks
3	Full term	38 weeks $\le Y_i^c < \infty$	$T_3 = \infty$

The proportional odds model for ordinal logistic regression models the cumulative probabilities $P(Y_i \le j)$ rather than the specific category probabilities $P(Y_i = j)$ as was the case for nominal logistic regression. We now develop the required expressions for the cumulative probabilities.

For $j = 1$ we have:

$$\begin{aligned}
P(Y_i \le 1) &= P\left(Y_i^c \le T_1\right) && \textbf{(14.103a)}\\
&= P(\beta_0^* + \beta_1^* X_i + k\varepsilon_L \le T_1) && \textbf{(14.103b)}\\
&= P(k\varepsilon_L \le T_1 - \beta_0^* - \beta_1^* X_i) && \textbf{(14.103c)}\\
&= P\left(\varepsilon_L \le \frac{T_1 - \beta_0^*}{k} - \frac{\beta_1^*}{k} X_i\right) && \textbf{(14.103d)}\\
&= P(\varepsilon_L \le \alpha_1 + \beta_1 X_i) && \textbf{(14.103e)}
\end{aligned}$$

where $\alpha_1 = (T_1 - \beta_0^*)/k$ and $\beta_1 = -\beta_1^*/k$. Since ε_L follows a standard logistic distribution, the cumulative probability in (14.103e) is obtained by using the cumulative distribution function (14.14b):

$$P(Y_i \leq 1) = \pi_{i1} = \frac{\exp(\alpha_1 + \beta_1 X_i)}{1 + \exp(\alpha_1 + \beta_1 X_i)} \tag{14.103f}$$

For $j = 2$, following the development in (14.103), we have:

$$\begin{aligned} P(Y_i \leq 2) &= P(Y_i^c \leq T_2) && \text{(14.104a)} \\ &= P(\beta_0^* + \beta_1^* X_i + k\varepsilon_L \leq T_2) && \text{(14.104b)} \\ &= P(k\varepsilon_L \leq T_2 - \beta_0^* - \beta_1^* X_i) && \text{(14.104c)} \\ &= P\left(\varepsilon_L \leq \frac{T_2 - \beta_0^*}{k} - \frac{\beta_1^*}{k} X_i\right) && \text{(14.104d)} \\ &= P(\varepsilon_L \leq \alpha_2 + \beta_1 X_i) && \text{(14.104e)} \\ &= \frac{\exp(\alpha_2 + \beta_1 X_i)}{1 + \exp(\alpha_2 + \beta_1 X_i)} && \text{(14.104f)} \end{aligned}$$

Notice that the only difference between (14.103f) and (14.104f) involves the intercept terms α_1 and α_2. The slopes β_1 are the same in both expressions. For the multiple regression case involving J ordered categories, we let:

$$\mathbf{X}_i = \begin{bmatrix} X_{i1} \\ X_{i2} \\ \cdots \\ X_{i,p-1} \end{bmatrix} \qquad \boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \cdots \\ \beta_{p-1} \end{bmatrix}$$

Equations (14.103f) and (14.104f) become for category j:

$$P(Y_i \leq j) = \frac{\exp(\alpha_j + \mathbf{X}_i'\boldsymbol{\beta})}{1 + \exp(\alpha_j + \mathbf{X}_i'\boldsymbol{\beta})} \qquad \text{for } j = 1, 2, \ldots, J-1 \tag{14.105}$$

Model (14.105) is often referred to as the *proportional odds model*. Taking the logit transformation of both sides yields the $J - 1$ *cumulative logits*:

$$\log_e\left[\frac{P(Y_i \leq j)}{1 - P(Y_i \leq j)}\right] = \alpha_j + \mathbf{X}_i'\boldsymbol{\beta} \qquad \text{for } j = 1, \ldots, J-1 \tag{14.106}$$

The difference between the ordinal logits in (14.106) and the nominal logits in (14.97b) should now be clear. In the nominal case, each of the $J - 1$ parameter vectors $\boldsymbol{\beta}_j$ is unique. For ordinal responses, the slope coefficient vectors $\boldsymbol{\beta}$ are identical for each of the $J - 1$ cumulative logits, but the intercepts differ.

As in the binary logistic regression case, each slope parameter can again be interpreted as the change in the logarithm of an odds ratio—this time the cumulative odds ratio—for a unit change in its associated predictor. In general, (14.106) satisfies, for $j = 1, \ldots, J-1$:

$$\log_e\left[\frac{P(Y_i \leq k)}{P(Y_i > k)} \div \frac{P(Y_j \leq k)}{P(Y_j > k)}\right] = (\mathbf{X}_i - \mathbf{X}_j)'\boldsymbol{\beta} \tag{14.107}$$

We now briefly discuss estimation methods before returning to the pregnancy duration example.

Maximum Likelihood Estimation. As was the case for nominal logistic regression, separate binary logistic regressions can be used to obtain estimates of the $J-1$ linear predictors in (14.106). For $j=1,\ldots,J-1$, we construct the binary outcome variable:

$$Y_i^{(j)} = \begin{cases} 1 & \text{if } Y_i \le j \\ 0 & \text{if } Y_i > j \end{cases}$$

and carry out a logistic regression analysis based on $Y_i^{(j)}$. Note that this approach leads to $J-1$ separate estimates of the slope parameter vector $\boldsymbol{\beta}$.

A better approach, if the required software is available, is to estimate $\alpha_1,\ldots,\alpha_{J-1}$ and $\boldsymbol{\beta}$ simultaneously using maximum likelihood estimation. From (14.100), the likelihood is given by:

$$\begin{aligned} P(Y_1,\ldots,Y_n) &= \prod_{i=1}^{n}\left(\prod_{j=1}^{J}[\pi_{ij}]^{Y_{ij}}\right) \\ &= \prod_{i=1}^{n}\left(\prod_{j=1}^{J}[P(Y_i \le j) - P(Y_i \le j-1)]^{Y_{ij}}\right) \end{aligned} \qquad \textbf{(14.108)}$$

Substitution of $P(Y_i \le J)=1$, $P(Y_i \le 0)=0$, and the expression for $P(Y_i \le j)$, $j=1,\ldots,J-1$, in (14.105) yields the required expression for the likelihood in terms of $\alpha_1,\ldots,\alpha_{J-1}$, and $\boldsymbol{\beta}$. The maximum likelihood estimates are those values of $\alpha_1,\ldots,\alpha_{J-1}$ and $\boldsymbol{\beta}$, namely, $a_1,\ldots,a_{J-1}$ and $\mathbf{b}$ that maximize (14.108). As always, we shall rely on standard statistical software to carry out the maximization. We now return to the pregnancy duration example.

Example

We continue the analysis of the pregnancy duration data, this time under the assumption that the response is ordinal, rather than nominal. Recall that $Y_i=1$ indicates preterm delivery, $Y_i=2$ indicates intermediate-term delivery, and $Y_i=3$ indicates full-term delivery. MINITAB ordinal logistic regression output is shown in Figure 14.19. As required with $J=3$, the program provides estimates for two intercepts, $a_1=2.930$ and $a_2=5.025$, and $p-1=5$ slope coefficients, $b_1=-.04887$, $b_2=1.9760$, $b_3=1.3635$, $b_4=1.5915$, and $b_5=1.6699$. The Wald P-values indicate that all of the regression coefficients are statistically significant at the .05 level.

As noted above, the coefficients can be interpreted as the change in the cumulative odds ratio for a unit change in the predictor. For example, the results indicate that the logarithm of the odds of a pre- or intermediate-term delivery ($Y_i \le 2$) for smokers ($X_5=1$) is estimated to be $b_4=1.5915$ times the logarithm of the odds for nonsmokers ($X_5=0$). The estimated cumulative odds ratio is given by $\exp(1.519)=4.91$ and a 95% confidence interval for the true cumulative odds ratio has a lower limit of 2.02 and an upper limit of 11.92. The remaining slope parameters can be interpreted in a similar fashion.

Notice again that the interpretation of the ordinal logistic regression model is much simpler than that for the nominal logistic regression model, because only a single slope vector $\boldsymbol{\beta}$ is estimated.

FIGURE 14.19 MINITAB Ordinal Logistic Regression Output—Pregnancy Duration Example.

```
Link Function: Logit

Response Information

Variable   Value      Count
preterm    1             26
           2             35
           3             41
           Total        102

Logistic Regression Table

                                                          Odds          95% CI
Predictor        Coef      SE Coef        Z      P      Ratio     Lower     Upper
Const(1)        2.930        1.465     2.00  0.045
Const(2)        5.025        1.521     3.30  0.001
nutritio     -0.04887      0.01168    -4.18  0.000     0.95      0.93      0.97
agecat1        1.9760       0.5875     3.36  0.001     7.21      2.28     22.82
agecat3        1.3635       0.5547     2.46  0.014     3.91      1.32     11.60
smoking        1.5915       0.4525     3.52  0.000     4.91      2.02     11.92
alcohol        1.6699       0.4727     3.53  0.000     5.31      2.10     13.42

Log-likelihood = -86.756
Test that all slopes are zero: G = 47.174, DF = 5, P-Value = 0.000
```

Comment

Our development of the proportional odds model assumed that the ordinal response Y_i was obtained from an explicit discretization of an observed continuous response Y_i^c, but this is not required. This model often works well for ordinal responses that do not arise from such a discretization. ■

14.13 Poisson Regression

We consider now another nonlinear regression model where the response outcomes are discrete. Poisson regression is useful when the outcome is a count, with large-count outcomes being rare events. For instance, the number of times a household shops at a particular supermarket in a week is a count, with a large number of shopping trips to the store during the week being a rare event. A researcher may wish to study the relation between a family's number of shopping trips to the store during a particular week and the family's income, number of children, distance from the store, and some other explanatory variables. As another example, the relation between the number of hospitalizations of a member of a health maintenance organization during the past year and the member's age, income, and previous health status may be of interest.

Poisson Distribution

The Poisson distribution can be utilized for outcomes that are counts ($Y_i = 0, 1, 2, \ldots$), with a large count or frequency being a rare event. The Poisson probability distribution is

as follows:

$$f(Y) = \frac{\mu^Y \exp(-\mu)}{Y!} \qquad Y = 0, 1, 2, \ldots \tag{14.109}$$

where $f(Y)$ denotes the probability that the outcome is Y and $Y! = Y(Y-1)\cdots 3 \cdot 2 \cdot 1$.

The mean and variance of the Poisson probability distribution are:

$$E\{Y\} = \mu \tag{14.110a}$$

$$\sigma^2\{Y\} = \mu \tag{14.110b}$$

Note that the variance is the same as the mean. Hence, if the number of store trips follows the Poisson distribution and the mean number of store trips for a family with three children is larger than the mean number of trips for a family with no children, the variances of the distributions of outcomes for the two families will also differ.

Comment

At times, the count responses Y will pertain to different units of time or space. For instance, in a survey intended to obtain the total number of store trips during a particular month, some of the counts pertained only to the last week of the month. In such cases, let μ denote the mean response for Y for a unit of time or space (e.g., one month), and let t denote the number of units of time or space to which Y corresponds. For instance, $t = 7/30$ if Y is the number of store trips during one week where the unit time is one month; $t = 1$ if Y is the number of store trips during the month. The Poisson probability distribution is then expressed as follows:

$$f(Y) = \frac{(t\mu)^Y \exp(-t\mu)}{Y!} \qquad Y = 0, 1, 2, \ldots \tag{14.111}$$

Our discussion throughout this section assumes that all responses Y_i pertain to the same unit of time or space. ■

Poisson Regression Model

The Poisson regression model, like any nonlinear regression model, can be stated as follows:

$$Y_i = E\{Y_i\} + \varepsilon_i \qquad i = 1, 2, \ldots, n$$

The mean response for the ith case, to be denoted now by μ_i for simplicity, is assumed as always to be a function of the set of predictor variables, $X_1, \ldots, X_{p-1}$. We use the notation $\mu(\mathbf{X}_i, \boldsymbol{\beta})$ to denote the function that relates the mean response μ_i to $\mathbf{X}_i$, the values of the predictor variables for case i, and $\boldsymbol{\beta}$, the values of the regression coefficients. Some commonly used functions for Poisson regression are:

$$\mu_i = \mu(\mathbf{X}_i, \boldsymbol{\beta}) = \mathbf{X}_i'\boldsymbol{\beta} \tag{14.112a}$$

$$\mu_i = \mu(\mathbf{X}_i, \boldsymbol{\beta}) = \exp(\mathbf{X}_i'\boldsymbol{\beta}) \tag{14.112b}$$

$$\mu_i = \mu(\mathbf{X}_i, \boldsymbol{\beta}) = \log_e(\mathbf{X}_i'\boldsymbol{\beta}) \tag{14.112c}$$

In all three cases, the mean responses μ_i must be nonnegative.

Since the distribution of the error terms ε_i for Poisson regression is a function of the distribution of the response Y_i, which is Poisson, it is easiest to state the Poisson regression

model in the following form:

Y_i are independent Poisson random variables with expected values μ_i, where:

$$\mu_i = \mu(\mathbf{X}_i, \boldsymbol{\beta}) \tag{14.113}$$

The most commonly used response function is $\mu_i = \exp(\mathbf{X}'\boldsymbol{\beta})$.

Maximum Likelihood Estimation

For Poisson regression model (14.113), the likelihood function is as follows:

$$\begin{aligned} L(\boldsymbol{\beta}) &= \prod_{i=1}^{n} f_i(Y_i) = \prod_{i=1}^{n} \frac{[\mu(\mathbf{X}_i, \boldsymbol{\beta})]^{Y_i} \exp[-\mu(\mathbf{X}_i, \boldsymbol{\beta})]}{Y_i!} \\ &= \frac{\{\prod_{i=1}^{n} [\mu(\mathbf{X}_i, \boldsymbol{\beta})]^{Y_i}\} \exp[-\sum_{i=1}^{n} \mu(\mathbf{X}_i, \boldsymbol{\beta})]}{\prod_{i=1}^{n} Y_i!} \end{aligned} \tag{14.114}$$

Once the functional form of $\mu(\mathbf{X}_i, \boldsymbol{\beta})$ is chosen, the maximization of (14.114) produces the maximum likelihood estimates of the regression coefficients $\boldsymbol{\beta}$. As before, it is easier to work with the logarithm of the likelihood function:

$$\log_e L(\boldsymbol{\beta}) = \sum_{i=1}^{n} Y_i \log_e[\mu(\mathbf{X}_i, \boldsymbol{\beta})] - \sum_{i=1}^{n} \mu(\mathbf{X}_i, \boldsymbol{\beta}) - \sum_{i=1}^{n} \log_e(Y_i!) \tag{14.115}$$

Numerical search procedures are used to find the maximum likelihood estimates $b_0, b_1, \ldots, b_{p-1}$. Iteratively reweighted least squares can again be used to obtain these estimates. We shall rely on standard statistical software packages specifically designed to handle Poisson regression to obtain the maximum likelihood estimates.

After the maximum likelihood estimates have been found, we can obtain the fitted response function and the fitted values:

$$\hat{\mu} = \mu(\mathbf{X}, \mathbf{b}) \tag{14.116a}$$

$$\hat{\mu}_i = \mu(\mathbf{X}_i, \mathbf{b}) \tag{14.116b}$$

For the three functions in (14.112), the fitted response functions and fitted values are:

$$\mu = \mathbf{X}'\boldsymbol{\beta}: \quad \hat{\mu} = \mathbf{X}'\mathbf{b} \quad \hat{\mu}_i = \mathbf{X}_i'\mathbf{b} \tag{14.116c}$$

$$\mu = \exp(\mathbf{X}'\boldsymbol{\beta}): \quad \hat{\mu} = \exp(\mathbf{X}'\mathbf{b}) \quad \hat{\mu}_i = \exp(\mathbf{X}_i'\mathbf{b}) \tag{14.116d}$$

$$\mu = \log_e(\mathbf{X}'\boldsymbol{\beta}): \quad \hat{\mu} = \log_e(\mathbf{X}'\mathbf{b}) \quad \hat{\mu}_i = \log_e(\mathbf{X}_i'\mathbf{b}) \tag{14.116e}$$

Model Development

Model development for a Poisson regression model is carried out in a similar fashion to that for logistic regression, conducting tests for individual coefficients or groups of coefficients based on the likelihood ratio test statistic G^2 in (14.60). For Poisson regression

model (14.113), the model deviance is as follows:

$$DEV(X_0, X_1, \ldots, X_{p-1}) = -2\left[\sum_{i=1}^{n} Y_i \log_e\left(\frac{\hat{\mu}_i}{Y_i}\right) + \sum_{i=1}^{n}(Y_i - \hat{\mu}_i)\right] \qquad \textbf{(14.117)}$$

where $\hat{\mu}_i$ is the fitted value for the ith case according to (14.116b). The deviance residual for the ith case is:

$$dev_i = \pm\left[-2Y_i \log_e\left(\frac{\hat{\mu}_i}{Y_i}\right) - 2(Y_i - \hat{\mu}_i)\right]^{1/2} \qquad \textbf{(14.118)}$$

The sign of the deviance residual is selected according to whether $Y_i - \hat{\mu}_i$ is positive or negative. Index plots of the deviance residuals and half-normal probability plots with simulated envelopes are useful for identifying outliers and checking the model fit.

Comment

If $Y_i = 0$, the term $[Y_i \log_e(\hat{\mu}_i / Y_i)]$ in (14.117) and (14.118) equals 0. ■

Inferences

Inferences for a Poisson regression model are carried out in the same way as for logistic regression. For instance, there is often interest in estimating the mean response for predictor variables $\mathbf{X}_h$. This estimate is obtained by substituting $\mathbf{X}_h$ into (14.116).

In Poisson regression analysis, there is sometimes also interest in estimating probabilities of certain outcomes for given levels of the predictor variables, for instance, $P(Y = 0 \mid \mathbf{X}_h)$. Such an estimated probability can be obtained readily by substituting $\hat{\mu}_h$ into (14.109).

Interval estimation of individual regression coefficients can be carried out by use of the large-sample estimated standard deviations furnished by regression programs with Poisson regression capabilities.

Example

The Miller Lumber Company is a large retailer of lumber and paint, as well as of plumbing, electrical, and other household supplies. During a representative two-week period, in-store surveys were conducted and addresses of customers were obtained. The addresses were then used to identify the metropolitan area census tracts in which the customers reside. At the end of the survey period, the total number of customers who visited the store from each census tract within a 10-mile radius was determined and relevant demographic information for each tract (average income, number of housing units, etc.) was obtained. Several other variables expected to be related to customer counts were constructed from maps, including distance from census tract to nearest competitor and distance to store.

Initial screening of the potential predictor variables was conducted which led to the retention of five predictor variables:

X_1: Number of housing units
X_2: Average income, in dollars
X_3: Average housing unit age, in years
X_4: Distance to nearest competitor, in miles
X_5: Distance to store, in miles
Y_i: Number of customers who visited store from census tract

TABLE 14.14 Data—Miller Lumber Company Example.

Census Tract i	Housing Units X_1	Average Income X_2	Average Age X_3	Competitor Distance X_4	Store Distance X_5	Number of Customers Y
1	606	41,393	3	3.04	6.32	9
2	641	23,635	18	1.95	8.89	6
3	505	55,475	27	6.54	2.05	28
...	...	...	...	...	...	...
108	817	54,429	47	1.90	9.90	6
109	268	34,022	54	1.20	9.51	4
110	519	52,850	43	2.92	8.62	6

TABLE 14.15 Fitted Poisson Response Function and Related Results—Miller Lumber Company Example.

(a) Fitted Poisson Response Function

$$\hat{\mu} = \exp[2.942 + .000606X_1 - .0000117X_2 - .00373X_3 + .168X_4 - .129X_5]$$

$$DEV(X_0, X_1, X_2, X_3, X_4, X_5) = 114.985$$

(b) Estimated Coefficients, Standard Deviations, and G^2 Test Statistics

Regression Coefficient	Estimated Regression Coefficient	Estimated Standard Deviation	G^2	P-value
β_0	2.9424	.207		
β_1	.0006058	.00014	18.21	.000
β_2	−.00001169	.0000021	31.80	.000
β_3	−.003726	.0018	4.38	.036
β_4	.1684	.026	41.66	.000
β_5	−.1288	.016	67.50	.000

Data for a portion of the $n = 110$ census tracts are shown in Table 14.14.

Poisson regression model (14.113) with response function:

$$\mu(\mathbf{X}, \boldsymbol{\beta}) = \exp(\mathbf{X}'\boldsymbol{\beta})$$

was fitted to the data, using LISP-STAT (Reference 14.10). Some principal results are presented in Table 14.15. Note that the deviance for this model is 114.985.

Likelihood ratio test statistics (14.60) were calculated for each of the individual regression coefficients. These G^2 test statistics are shown in Table 14.15b, together with their associated P-values, each based on the chi-square distribution with one degree of freedom. We note from the P-values that each predictor variable makes a marginal contribution to the fit of the regression model and consequently should be retained in the model.

A portion of the deviance residuals dev_i is shown in Table 14.16, together with the responses Y_i and the fitted values $\hat{\mu}_i$. Analysis of the deviance residuals did not disclose any major problems. Figure 14.20 contains an index plot of the deviance residuals. We note a few large negative deviance residuals; these are for census tracts where $Y = 0$; i.e.,

TABLE 14.16 Responses, Fitted Values, and Deviance Residuals—Miller Lumber Company Example.

Census Tract i	Y_i	$\hat{\mu}_i$	dev_i
1	9	12.3	−.999
2	6	8.8	−.992
3	28	28.1	−.024
...	...	...	...
108	6	5.3	.289
109	4	4.4	−.197
110	6	6.4	−.171

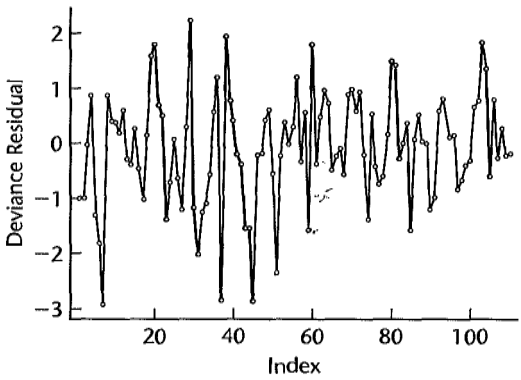


FIGURE 14.20 Index Plot of Deviance Residuals—Miller Lumber Company Example.

there were no customers from these areas. These may be difficult cases to fit with a Poisson regression model.

14.14 Generalized Linear Models

We conclude this chapter and the regression portion of this book by noting that all of the regression models considered, linear and nonlinear, belong to a family of models called *generalized linear models*. This family was first introduced by Nelder and Wedderburn (Reference 14.11) and encompasses normal error linear regression models and the nonlinear exponential, logistic, and Poisson regression models, as well as many other models, such as log-linear models for categorical data.

The class of generalized linear models can be described as follows:

1. $Y_1, \ldots, Y_n$ are n independent responses that follow a probability distribution belonging to the *exponential family* of probability distributions, with expected value $E\{Y_i\} = \mu_i$.

2. A *linear predictor* based on the predictor variables $X_{i1}, \ldots, X_{i,p-1}$ is utilized, denoted by $\mathbf{X}_i'\boldsymbol{\beta}$:

$$\mathbf{X}_i'\boldsymbol{\beta} = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_{p-1} X_{i,p-1}$$

3. The *link function* g relates the linear predictor to the mean response:

$$\mathbf{X}_i'\boldsymbol{\beta} = g(\mu_i)$$

Generalized linear models may have nonconstant variances σ_i^2 for the responses Y_i, but the variance σ_i^2 must be a function of the predictor variables through the mean response μ_i.

To illustrate the concept of the link function, consider first logistic regression model (14.41). There, the logit transformation $F_L^{-1}(\pi_i)$ in (14.18a) serves to link the linear predictor $\mathbf{X}_i'\boldsymbol{\beta}$ to the mean response $\mu_i = \pi_i$:

$$g(\mu_i) = g(\pi_i) = \log_e\left(\frac{\pi_i}{1-\pi_i}\right) = \mathbf{X}_i'\boldsymbol{\beta}$$

As a second example, consider Poisson regression model (14.113). There we considered several response functions in (14.112). For the response function $\mu_i = \exp(\mathbf{X}_i'\boldsymbol{\beta})$ in (14.112b), the linking relation is:

$$g(\mu_i) = \log_e(\mu_i) = \mathbf{X}_i'\boldsymbol{\beta}$$

We see from the Poisson regression models that there may be many different possible link functions that can be employed. They need only be monotonic and differentiable.

Finally, we consider the normal error regression model in (6.7). There the link function is simply:

$$g(\mu_i) = \mu_i$$

since the linking relation is:

$$\mathbf{X}_i'\boldsymbol{\beta} = \mu_i$$

The link function $g(\mu_i)$ for the normal error case is called the identity or unity link function.

Any regression model that belongs to the family of generalized linear models can be analyzed in a unified fashion. The maximum likelihood estimates of the regression parameters can be obtained by iteratively reweighted least squares [by ordinary least squares for normal error linear regression models (6.7)]. Tests for model development to determine whether some predictor variables may be dropped from the model can be conducted using likelihood ratio tests. Reference 14.12 provides further details about generalized linear models and their analysis.

Cited References

14.1. Kennedy, W. J., Jr., and J. E. Gentle. *Statistical Computing*. New York: Marcel Dekker, 1980.

14.2. Agresti, A. *Categorical Data Analysis*. 2nd ed. New York: John Wiley & Sons, 2002.

14.3. *LogXact 5*. Cytel Software Corporation. Cambridge, Massachusetts, 2003.

14.4. Hosmer, D. W., and S. Lemeshow. *Applied Logistic Regression*. 2nd ed. New York: John Wiley & Sons, 2000.

14.5. Cook, R. D., and S. Weisberg. *Applied Regression Including Computing and Graphics*. New York: John Wiley & Sons, 1999.

14.6. Atkinson, A. C. "Two Graphical Displays for Outlying and Influential Observations in Regression," *Biometrika* 68 (1981), pp. 13–20.

14.7. Johnson, R. A., and D. W. Wichern. *Applied Multivariate Statistical Analysis*. 5th ed. Englewood Cliffs, N.J.: Prentice Hall, 2001.

14.8. Lachenbruch, P. A. *Discriminant Analysis*. New York: Hafner Press. 1975.

14.9. Begg, C. B., and R. Gray. "Calculation of Polytomous Logistic Regression Parameters Using Individualized Regressions," *Biometrika* 71 (1984), pp. 11–18.

14.10. Tierney, L. *LISP-STAT: An Object-Oriented Environment for Statistical Computing and Dynamic Graphics*. New York: John Wiley & Sons, 1990.

14.11. Nelder, J. A., and R. W. M. Wedderburn. "Generalized Linear Models," *Journal of the Royal Statistical Society A* 135 (1972), pp. 370–84.

14.12. McCullagh, P., and J. A. Nelder. *Generalized Linear Models*. 2nd ed. London: Chapman and Hall, 1999.

Problems

14.1. A student stated: "I fail to see why the response function needs to be constrained between 0 and 1 when the response variable is binary and has a Bernoulli distribution. The fit to 0, 1 data will take care of this problem for any response function." Comment.

14.2. Since the logit transformation (14.18) linearizes the logistic response function, why can't this transformation be used on the individual responses Y_i and a linear response function then fitted? Explain.

14.3. If the true response function is J-shaped when the response variable is binary, would the use of the logistic response function be appropriate? Explain.

14.4. a. Plot the logistic mean response function (14.16) when $\beta_0 = -25$ and $\beta_1 = .2$.

b. For what value of X is the mean response equal to .5?

c. Find the odds when $X = 150$, when $X = 151$, and the ratio of the odds when $X = 151$ to the odds when $X = 150$. Is this odds ratio equal to $\exp(\beta_1)$ as it should be?

*14.5. a. Plot the logistic mean response function (14.16) when $\beta_0 = 20$ and $\beta_1 = -.2$.

b. For what value of X is the mean response equal to .5?

c. Find the odds when $X = 125$, when $X = 126$, and the ratio of the odds when $X = 126$ to the odds when $X = 125$. Is the odds ratio equal to $\exp(\beta_1)$ as it should be?

14.6. a. Plot the probit mean response function (14.12) for $\beta_0^* = -25$ and $\beta_1^* = .2$. How does this function compare to the logistic mean response function in part (a) of Problem 14.4?

b. For what value of X is the mean response equal to .5?

*14.7. **Annual dues.** The board of directors of a professional association conducted a random sample survey of 30 members to assess the effects of several possible amounts of dues increase. The sample results follow. X denotes the dollar increase in annual dues posited in the survey interview, and $Y = 1$ if the interviewee indicated that the membership will not be renewed at that amount of dues increase and 0 if the membership will be renewed.

i:	1	2	3	...	28	29	30
X_i:	30	30	30	...	49	50	50
Y_i:	0	1	0	...	0	1	1

Logistic regression model (14.20) is assumed to be appropriate.

a. Find the maximum likelihood estimates of β_0 and β_1. State the fitted response function.

b. Obtain a scatter plot of the data with both the fitted logistic response function from part (a) and a lowess smooth superimposed. Does the fitted logistic response function appear to fit well?

c. Obtain $\exp(b_1)$ and interpret this number.

d. What is the estimated probability that association members will not renew their membership if the dues are increased by \$40?

e. Estimate the amount of dues increase for which 75 percent of the members are expected not to renew their association membership.

14.8. Refer to **Annual dues** Problem 14.7.

a. Fit a probit mean response function (14.12) to the data. Qualitatively compare the fit here with the logistic fit obtained in part (a) of Problem 14.7. What do you conclude?

b. Fit a complimentary log-log mean response function (14.19) to the data. Qualitatively compare the fit here with the logistic fit obtained in part (a) of Problem 14.7. What do you conclude?

14.9. **Performance ability.** A psychologist conducted a study to examine the nature of the relation, if any, between an employee's emotional stability (X) and the employee's ability to perform in a task group (Y). Emotional stability was measured by a written test for which the higher the score, the greater is the emotional stability. Ability to perform in a task group ($Y = 1$ if able, $Y = 0$ if unable) was evaluated by the supervisor. The results for 27 employees were:

i:	1	2	3	...	25	26	27
X_i:	474	432	453	...	562	506	600
Y_i:	0	0	0		1	0	1

Logistic regression model (14.20) is assumed to be appropriate.

a. Find the maximum likelihood estimates of β_0 and β_1. State the fitted response function.

b. Obtain a scatter plot of the data with both the fitted logistic response function from part (a) and a lowess smooth superimposed. Does the fitted logistic response function appear to fit well?

c. Obtain $\exp(b_1)$ and interpret this number.

d. What is the estimated probability that employees with an emotional stability test score of 550 will be able to perform in a task group?

e. Estimate the emotional stability test score for which 70 percent of the employees with this test score are expected to be able to perform in a task group.

14.10. Refer to **Performance ability** Problem 14.9.

a. Fit a probit mean response function (14.12) to the data. Qualitatively compare the fit here with the logistic fit obtained in part (a) of Problem 14.9. What do you conclude?

b. Fit a complementary log-log mean response function (14.19) to the data. Qualitatively compare the fit here with the logistic fit obtained in part (a) of Problem 14.9. What do you conclude?

*14.11. **Bottle return.** A carefully controlled experiment was conducted to study the effect of the size of the deposit level on the likelihood that a returnable one-liter soft-drink bottle will be returned. A bottle return was scored 1, and no return was scored 0. The data to follow show the number of bottles that were returned ($Y_{.j}$) out of 500 sold (n_j) at each of six deposit levels (X_j, in cents):

j:	1	2	3	4	5	6
Deposit level X_j:	2	5	10	20	25	30
Number sold n_j:	500	500	500	500	500	500
Number returned $Y_{.j}$:	72	103	170	296	406	449

An analyst believes that logistic regression model (14.20) is appropriate for studying the relation between size of deposit and the probability a bottle will be returned.

a. Plot the estimated proportions $p_j = Y_{.j}/n_j$ against X_j. Does the plot support the analyst's belief that the logistic response function is appropriate?

b. Find the maximum likelihood estimates of β_0 and β_1. State the fitted response function.

c. Obtain a scatter plot of the data with the estimated proportions from part (a), and superimpose the fitted logistic response function from part (b). Does the fitted logistic response function appear to fit well?

d. Obtain $\exp(b_1)$ and interpret this number.

e. What is the estimated probability that a bottle will be returned when the deposit is 15 cents?

f. Estimate the amount of deposit for which 75 percent of the bottles are expected to be returned.

14.12. **Toxicity experiment.** In an experiment testing the effect of a toxic substance, 1,500 experimental insects were divided at random into six groups of 250 each. The insects in each group were exposed to a fixed dose of the toxic substance. A day later, each insect was observed. Death from exposure was scored 1, and survival was scored 0. The results are shown below; X_j denotes the dose level (on a logarithmic scale) administered to the insects in group j and $Y_{.j}$ denotes the number of insects that died out of the 250 (n_j) in the group.

j:	1	2	3	4	5	6
X_j:	1	2	3	4	5	6
n_j:	250	250	250	250	250	250
$Y_{.j}$:	28	53	93	126	172	197

Logistic regression model (14.20) is assumed to be appropriate.

a. Plot the estimated proportions $p_j = Y_{.j}/n_j$ against X_j. Does the plot support the analyst's belief that the logistic response function is appropriate?

b. Find the maximum likelihood estimates of β_0 and β_1. State the fitted response function.

c. Obtain a scatter plot of the data with the estimated proportions from part (a), and superimpose the fitted logistic response function from part (b). Does the fitted logistic response function appear to fit well?

d. Obtain $\exp(b_1)$ and interpret this number.

e. What is the estimated probability that an insect dies when the dose level is $X = 3.5$?

f. What is the estimated median lethal dose—that is, the dose for which 50 percent of the experimental insects are expected to die?

14.13. **Car purchase.** A marketing research firm was engaged by an automobile manufacturer to conduct a pilot study to examine the feasibility of using logistic regression for ascertaining the likelihood that a family will purchase a new car during the next year. A random sample of 33 suburban families was selected. Data on annual family income (X_1, in thousand dollars) and the current age of the oldest family automobile (X_2, in years) were obtained. A follow-up interview conducted 12 months later was used to determine whether the family actually purchased a new car ($Y = 1$) or did not purchase a new car ($Y = 0$) during the year.

i:	1	2	3	...	31	32	33
X_{i1}:	32	45	60	...	21	32	17
X_{i2}:	3	2	2	...	3	5	1
Y_i:	0	0	1	...	0	1	0

Multiple logistic regression model (14.41) with two predictor variables in first-order terms is assumed to be appropriate.

a. Find the maximum likelihood estimates of β_0, β_1, and β_2. State the fitted response function.

b. Obtain $\exp(b_1)$ and $\exp(b_2)$ and interpret these numbers.

c. What is the estimated probability that a family with annual income of \$50 thousand and an oldest car of 3 years will purchase a new car next year?

*14.14. **Flu shots.** A local health clinic sent fliers to its clients to encourage everyone, but especially older persons at high risk of complications, to get a flu shot in time for protection against an expected flu epidemic. In a pilot follow-up study, 159 clients were randomly selected and asked whether they actually received a flu shot. A client who received a flu shot was coded $Y = 1$, and a client who did not receive a flu shot was coded $Y = 0$. In addition, data were collected on their age (X_1) and their health awareness. The latter data were combined into a health awareness index (X_2), for which higher values indicate greater awareness. Also included in the data was client gender, where males were coded $X_3 = 1$ and females were coded $X_3 = 0$.

i:	1	2	3	...	157	158	159
X_{i1}:	59	61	82	...	76	68	73
X_{i2}:	52	55	51		22	32	56
X_{i3}:	0	1	0	...	1	0	1
Y_i:	0	0	1		1	1	1

Multiple logistic regression model (14.41) with three predictor variables in first-order terms is assumed to be appropriate.

a. Find the maximum likelihood estimates of β_0, β_1, β_2, and β_3. State the fitted response function.

b. Obtain $\exp(b_1)$, $\exp(b_2)$, and $\exp(b_3)$. Interpret these numbers.

c. What is the estimated probability that male clients aged 55 with a health awareness index of 60 will receive a flu shot?

*14.15. Refer to **Annual dues** Problem 14.7. Assume that the fitted model is appropriate and that large-sample inferences are applicable.

a. Obtain an approximate 90 percent confidence interval for $\exp(\beta_1)$. Interpret your interval.

b. Conduct a Wald test to determine whether dollar increase in dues (X) is related to the probability of membership renewal; use $\alpha = .10$. State the alternatives, decision rule, and conclusion. What is the approximate P-value of the test?

c. Conduct a likelihood ratio test to determine whether dollar increase in dues (X) is related to the probability of membership renewal; use $\alpha = .10$. State the full and reduced models, decision rule, and conclusion. What is the approximate P-value of the test? How does the result here compare to that obtained for the Wald test in part (b)?

14.16. Refer to **Performance ability** Problem 14.9. Assume that the fitted model is appropriate and that large-sample inferences are applicable.

a. Obtain an approximate 95 percent confidence interval for $\exp(\beta_1)$. Interpret your interval.

b. Conduct a Wald test to determine whether employee's emotional stability (X) is related to the probability that the employee will be able to perform in a task group; use $\alpha = .05$. State the alternatives, decision rule, and conclusion. What is the approximate P-value of the test?

c. Conduct a likelihood ratio test to determine whether employee's emotional stability (X) is related to the probability that the employee will be able to perform in a task group; use $\alpha = .05$. State the full and reduced models, decision rule, and conclusion. What is the approximate P-value of the test? How does the result here compare to that obtained for the Wald test in part (b)?

*14.17. Refer to **Bottle return** Problem 14.11. Assume that the fitted model is appropriate and that large-sample inferences are applicable.

a. Obtain an approximate 95 percent confidence interval for β_1. Convert this confidence interval into one for the odds ratio. Interpret this latter interval.

b. Conduct a Wald test to determine whether deposit level (X) is related to the probability that a bottle is returned; use $\alpha = .05$. State the alternatives, decision rule, and conclusion. What is the approximate P-value of the test?

c. Conduct a likelihood ratio test to determine whether deposit level (X) is related to the probability that a bottle is returned; use $\alpha = .05$. State the full and reduced models, decision rule, and conclusion. What is the approximate P-value of the test? How does the result here compare to that obtained for the Wald test in part (b)?

14.18. Refer to **Toxicity experiment** Problem 14.12. Assume that the fitted model is appropriate and that large-sample inferences are applicable.

a. Obtain an approximate 99 percent confidence interval for β_1. Convert this confidence interval into one for the odds ratio. Interpret this latter interval.

b. Conduct a Wald test to determine whether dose level (X) is related to the probability that an insect dies; use $\alpha = .01$. State the alternatives, decision rule, and conclusion. What is the approximate P-value of the test?

c. Conduct a likelihood ratio test to determine whether dose level (X) is related to the probability that an insect dies; use $\alpha = .01$. State the full and reduced models, decision rule, and conclusion. What is the approximate P-value of the test? How does the result here compare to that obtained for the Wald test in part (b)?

14.19. Refer to **Car purchase** Problem 14.13. Assume that the fitted model is appropriate and that large-sample inferences are applicable.

a. Obtain joint confidence intervals for the family income odds ratio $\exp(20\beta_1)$ for families whose incomes differ by 20 thousand dollars and for the age of the oldest family automobile odds ratio $\exp(2\beta_2)$ for families whose oldest automobiles differ in age by 2 years, with family confidence coefficient of approximately .90. Interpret your intervals.

b. Use the Wald test to determine whether X_2, age of oldest family automobile, can be dropped from the regression model; use $\alpha = .05$. State the alternatives, decision rule, and conclusion. What is the approximate P-value of the test?

c. Use the likelihood ratio test to determine whether X_2, age of oldest family automobile, can be dropped from the regression model; use $\alpha = .05$. State the full and reduced models, decision rule, and conclusion. What is the approximate P-value of the test? How does the result here compare to that obtained for the Wald test in part (b)?

d. Use the likelihood ratio test to determine whether the following three second-order terms, the square of annual family income, the square of age of oldest automobile, and the two-factor interaction effect between annual family income and age of oldest automobile, should be added simultaneously to the regression model containing family income and age of oldest automobile as first-order terms; use $\alpha = .05$. State the full and reduced models, decision rule, and conclusion. What is the approximate P-value of the test?

*14.20. Refer to **Flu shots** Problem 14.14.

a. Obtain joint confidence intervals for the age odds ratio $\exp(30\beta_1)$ for male clients whose ages differ by 30 years and for the health awareness index odds ratio $\exp(25\beta_2)$ for male clients whose health awareness index differs by 25, with family confidence coefficient of approximately .90. Interpret your intervals.

b. Use the Wald test to determine whether X_3, client gender, can be dropped from the regression model; use $\alpha = .05$. State the alternatives, decision rule, and conclusion. What is the approximate P-value of the test?

c. Use the likelihood ratio test to determine whether X_3, client gender, can be dropped from the regression model; use $\alpha = .05$. State the full and reduced models, decision rule, and

conclusion. What is the approximate P-value of the test? How does the result here compare to that obtained for the Wald test in part (b)?

d. Use the likelihood ratio test to determine whether the following three second-order terms, the square of age, the square of health awareness index, and the two-factor interaction effect between age and health awareness index, should be added simultaneously to the regression model containing age and health awareness index as first-order terms; use $\alpha = .05$. State the alternatives, full and reduced models, decision rule, and conclusion. What is the approximate P-value of the test?

14.21. Refer to **Car purchase** Problem 14.13 where the pool of predictors consists of all first-order terms and all second-order terms in annual family income and age of oldest family automobile.

a. Use forward selection to decide which predictor variables enter into the regression model. Control the α risk at .10 at each stage. Which variables are entered into the regression model?

b. Use backward elimination to decide which predictor variables can be dropped from the regression model. Control the α risk at .10 at each stage. Which variables are retained? How does this compare to your results in part (a)?

c. Find the best model according to the AIC_p criterion. How does this compare to your results in parts (a) and (b)?

d. Find the best model according to the SBC_p criterion. How does this compare to your results in parts (a), (b) and (c)?

*14.22. Refer to **Flu shots** Problem 14.14 where the pool of predictors consists of all first-order terms and all second-order terms in age and health awareness index.

a. Use forward selection to decide which predictor variables enter into the regression model. Control the α risk at .10 at each stage. Which variables are entered into the regression model?

b. Use backward elimination to decide which predictor variables can be dropped from the regression model. Control the α risk at .10 at each stage. Which variables are retained? How does this compare to your results in part (a)?

c. Find the best model according to the AIC_p criterion. How does this compare to your results in parts (a) and (b)?

d. Find the best model according to the SBC_p criterion. How does this compare to your results in parts (a), (b) and (c)?

*14.23. Refer to **Bottle return** Problem 14.11. Use the groups given there to conduct a chi-square goodness of fit test of the appropriateness of logistic regression model (14.20). Control the risk of a Type I error at .01. State the alternatives, decision rule, and conclusion.

14.24. Refer to **Toxicity experiment** Problem 14.12. Use the groups given there to conduct a deviance goodness of fit test of the appropriateness of logistic regression model (14.20). Control the risk of a Type I error at .01. State the alternatives, decision rule, and conclusion.

*14.25. Refer to **Annual dues** Problem 14.7.

a. To assess the appropriateness of the logistic regression function, form three groups of 10 cases each according to their fitted logit values $\hat{\pi}'$. Plot the estimated proportions p_j against the midpoints of the $\hat{\pi}'$ intervals. Is the plot consistent with a response function of monotonic sigmoidal shape? Explain.

b. Obtain the studentized Pearson residuals (14.81) and plot them against the estimated model probabilities with a lowess smooth superimposed. What does the plot suggest about the adequacy of the fit of the logistic regression model?

14.26. Refer to **Performance ability** Problem 14.9.

a. To assess the appropriateness of the logistic regression function, form three groups of nine cases each according to their fitted logit values $\hat{\pi}'$. Plot the estimated proportions p_j

against the midpoints of the $\hat{\pi}'$ intervals. Is the plot consistent with a response function of monotonic sigmoidal shape? Explain.

b. Obtain the deviance residuals (14.83) and plot them against the estimated model probabilities with a lowess smooth superimposed. What does the plot suggest about the adequacy of the fit of the logistic regression model?

14.27. Refer to **Car purchase** Problems 14.13 and 14.21.

a. To assess the appropriateness of the logistic regression model obtained in part (d) of Problem 14.21, form three groups of 11 cases each according to their fitted logit values $\hat{\pi}'$. Plot the estimated proportions p_j against the midpoints of the $\hat{\pi}'$ intervals. Is the plot consistent with a response function of monotonic sigmoidal shape? Explain.

b. Obtain the studentized Pearson residuals (14.81) and plot them against the estimated model probabilities with a lowess smooth superimposed. What does the plot suggest about the adequacy of the fit of the logistic regression model?

*14.28. Refer to **Flu shots** Problems 14.14 and 14.22.

a. To assess the appropriateness of the logistic regression model obtained in part (d) of Problem 14.22, form 8 groups of approximately 20 cases each according to their fitted logit values $\hat{\pi}'$. Plot the estimated proportions p_j against the midpoints of the $\hat{\pi}'$ intervals. Is the plot consistent with a response function of monotonic sigmoidal shape? Explain.

b. Using the groups formed in part (a), conduct a Hosmer-Lemeshow goodness of fit test for the appropriateness of the logistic regression function; use $\alpha = .05$. State the alternatives, decision rule, and conclusions. What is the P-value of the test?

c. Obtain the deviance residuals (14.83) and plot them against the estimated model probabilities with a lowess smooth superimposed. What does the plot suggest about the adequacy of the fit of the logistic regression model?

*14.29. Refer to **Annual dues** Problem 14.7.

a. For the logistic regression model fit in Problem 14.7a, prepare an index plot of the diagonal elements of the estimated hat matrix (14.80). Use the plot to identify any outlying X observations.

b. To assess the influence of individual observations, obtain the delta chi-square statistic (14.85), the delta deviance statistic (14.86), and Cook's distance (14.87) for each observation. Plot each of these in separate index plots and identify any influential observations. Summarize your findings.

14.30. Refer to **Performance ability** Problem 14.9.

a. For the logistic regression fit in Problem 14.9a, prepare an index plot of the diagonal elements of the estimated hat matrix (14.80). Use the plot to identify any outlying X observations.

b. To assess the influence of individual observations, obtain the delta chi-square statistic (14.85), the delta deviance statistic (14.86), and Cook's distance (14.87) for each observation. Plot each of these in separate index plots and identify any influential observations. Summarize your findings.

14.31. Refer to **Car Purchase** Problems 14.13 and 14.21.

a. For the logistic regression model obtained in part (d) of Problem 14.21, prepare an index plot of the diagonal elements of the estimated hat matrix (14.80). Use the plot to identify any outlying X observations.

b. To assess the influence of individual observations, obtain the delta chi-square statistic (14.85), the delta deviance statistic (14.86), and Cook's distance (14.87) for each

observation. Plot each of these in separate index plots and identify any influential observations. Summarize your findings.

*14.32. Refer to **Flu shots** Problem 14.14.

a. For the logistic regression fit in Problem 14.14a, prepare an index plot of the diagonal elements of the estimated hat matrix (14.80). Use the plot to identify any outlying X observations.

b. To assess the influence of individual observations, obtain the delta chi-square statistic (14.85), the delta deviance statistic (14.86), and Cook's distance (14.87) for each observation. Plot each of these in separate index plots and identify any influential observations. Summarize your findings.

*14.33. Refer to **Annual dues** Problem 14.7.

a. Based on the fitted regression function in Problem 14.7a, obtain an approximate 90 percent confidence interval for the mean response π_h for a dues increase of $X_h = \$40$.

b. A prediction rule is to be developed, based on the fitted regression function in Problem 14.7a. Based on the sample cases, find the total error rate, the error rate for renewers, and the error rate for nonrenewers for the following cutoffs: .40, .45, .50, .55, .60.

c. Based on your results in part (b), which cutoff minimizes the total error rate? Are the error rates for renewers and nonrenewers fairly balanced at this cutoff? Obtain the area under the ROC curve to assess the model's predictive power here. What do you conclude?

d. How can you establish whether the observed total error rate for the best cutoff in part (b) is a reliable indicator of the predictive ability of the fitted regression function and the chosen cutoff?

14.34. Refer to **Performance ability** Problem 14.9.

a. Using the fitted regression function in Problem 14.9a, obtain joint confidence intervals for the mean response π_h for persons with emotional stability test scores $X_h = 550$ and 625, respectively, with an approximate 90 percent family confidence coefficient. Interpret your intervals.

b. A prediction rule, based on the fitted regression function in Problem 14.9a, is to be developed. For the sample cases, find the total error rate, the error rate for employees able to perform in a task group, and the error rate for employees not able to perform for the following cutoffs: .325, .425, .525, .625.

c. On the basis of your results in part (b), which cutoff minimizes the total error rate? Are the error rates for employees able to perform in a task group and for employees not able to perform fairly balanced at this cutoff? Obtain the area under the ROC curve to assess the model's predictive power here. What do you conclude?

d. How can you establish whether the observed total error rate for the best cutoff in part (c) is a reliable indicator of the predictive ability of the fitted regression function and the chosen cutoff?

14.35. Refer to **Bottle return** Problem 14.11.

a. For the fitted regression function in Problem 14.11a, obtain an approximate 95 percent confidence interval for the probability of a purchase for deposit $X_h = 15$ cents. Interpret your interval.

b. A prediction rule is to be developed, based on the fitted regression function in Problem 14.11a. For the sample cases, find the total error rate, the error rate for purchasers, and the error rate for nonpurchasers for the following cutoffs: .150, .300, .450, .600, .750.

c. According to your results in part (b), which cutoff minimizes the total error rate? Are the error rates for purchasers and nonpurchasers fairly balanced at this cutoff? Obtain the area under the ROC curve to assess the model's predictive power here. What do you conclude?

d. How can you establish whether the observed total error rate for the best cutoff in part (c) is a reliable indicator of the predictive ability of the fitted regression function and the chosen cutoff?

*14.36. Refer to **Flu shots** Problem 14.14.

a. On the basis of the fitted regression function in Problem 14.14a, obtain a confidence interval for the mean response π_h for a female whose age is 65 and whose health awareness index is 50, with an approximate 90 percent family confidence coefficient. Interpret your intervals.

b. A prediction rule is to be based on the fitted regression function in Problem 14.14a. For the sample cases, find the total error rate, the error rate for clients receiving the flu shot, and the error rate for clients not receiving the flu shot for the following cutoffs: .05, .10, .15, .20.

c. Based on your results in part (b), which cutoff minimizes the total error rate? Are the error rates for clients receiving the flu shot and for clients not receiving the flu shot fairly balanced at this cutoff? Obtain the area under the ROC curve to assess the model's predictive power here. What do you conclude?

d. How can you establish whether the observed total error rate for the best cutoff in part (c) is a reliable indicator of the predictive ability of the fitted regression function and the chosen cutoff?

14.37. Polytomous logistic regression extends the binary response outcome to a multicategory response outcome for either nominal level or ordinal level data. Discuss the advantages and disadvantages of treating multicategory ordinal level outcomes as a series of binary logistic regression models, as a nominal level polytomous regression model, or as a proportional odds model.

*14.38. Refer to **Airfreight breakage** Problem 1.21.

a. Fit the Poisson regression model (14.113) with the response function $\mu(\mathbf{X}, \boldsymbol{\beta}) = \exp(\beta_0 + \beta_1 X)$. State the estimated regression coefficients, their estimated standard deviations, and the estimated response function.

b. Obtain the deviance residuals and present them in an index plot. Do there appear to be any outlying cases?

c. Estimate the mean number of ampules broken when $X = 0, 1, 2, 3$. Compare these estimates with those obtained by means of the fitted linear regression function in Problem 1.21a.

d. Plot the Poisson and linear regression functions, together with the data. Which regression function appears to be a better fit here? Discuss.

e. Management wishes to estimate the probability that 10 or fewer ampules are broken when there is no transfer of the shipment. Use the fitted Poisson regression function to obtain this estimate.

f. Obtain an approximate 95 percent confidence interval for β_1. Interpret your interval estimate.

14.39. **Geriatric study.** A researcher in geriatrics designed a prospective study to investigate the effects of two interventions on the frequency of falls. One hundred subjects were randomly assigned to one of the two interventions: education only ($X_1 = 0$) and education plus aerobic exercise training ($X_1 = 1$). Subjects were at least 65 years of age and in reasonably good health.

Three variables considered to be important as control variables were gender (X_2; 0 = female; 1 = male), a balance index (X_3), and a strength index (X_4). The higher the balance index, the more stable is the subject; and the higher the strength index, the stronger is the subject. Each subject kept a diary recording the number of falls (Y) during the six months of the study. The data follow:

Subject i	Number of Falls Y_i	Intervention X_{i1}	Gender X_{i2}	Balance Index X_{i3}	Strength Index X_{i4}
1	1	1	0	45	70
2	1	1	0	62	66
3	2	1	1	43	64
	...	...	...	..	...
98	4	0	0	69	48
99	4	0	1	50	52
100	2	0	0	37	56

a. Fit the Poisson regression model (14.113) with the response function $\mu(\mathbf{X}, \boldsymbol{\beta}) = \exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4)$. State the estimated regression coefficients, their estimated standard deviations, and the estimated response function.

b. Obtain the deviance residuals and present them in an index plot. Do there appear to be any outlying cases?

c. Assuming that the fitted model is appropriate, use the likelihood ratio test to determine whether gender (X_2) can be dropped from the model; control α at .05. State the full and reduced models, decision rule, and conclusion. What is the P-value of the test.

d. For the fitted model containing only X_1, X_3, and X_4 in first-order terms, obtain an approximate 95 percent confidence interval for β_1. Interpret your confidence interval. Does aerobic exercise reduce the frequency of falls when controlling for balance and strength?

Exercises

14.40. Show the equivalence of (14.16) and (14.17).

14.41. Derive (14.34) from (14.26).

14.42. Derive (14.18a), using (14.16) and (14.18).

14.43. (Calculus needed.) Maximum likelihood estimation theory states that the estimated large-sample variance-covariance matrix for maximum likelihood estimators is given by the inverse of the information matrix, the elements of which are the negatives of the expected values of the second-order partial derivatives of the logarithm of the likelihood function evaluated at $\boldsymbol{\beta} = \mathbf{b}$:

$$\left[-E\left\{ \frac{\partial^2 \log_e L(\boldsymbol{\beta})}{\partial \beta_i \partial \beta_j} \right\}_{\boldsymbol{\beta}=\mathbf{b}} \right]^{-1}$$

Show that this matrix simplifies to (14.51) for logistic regression. Consider the case where $p - 1 = 1$.

14.44. (Calculus needed.) Estimate the approximate variance-covariance matrix of the estimated regression coefficients for the programming task example in Table 14.1a, using (14.51), and verify the estimated standard deviations in Table 14.1b.

14.45. Show that the logistic response function (13.10) reduces to the response function in (14.20) when the Y_i are independent Bernoulli random variables with $E\{Y_i\} = \pi_i$.

14.46. Consider the multiple logistic regression model with $\mathbf{X}'\boldsymbol{\beta} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2$. Derive an expression for the odds ratio for X_1. Does $\exp(\beta_1)$ have the same meaning here as for a regression model containing no interaction term?

14.47. A Bernoulli response Y_i has expected value:

$$E\{Y_i\} = \pi_i = 1 - \exp\left[-\exp\left(\frac{X_i - \gamma_0}{\gamma_1}\right)\right]$$

Show that the link function here is the complementary log-log transformation of π_i, namely, $\log_e[-\log_e(1 - \pi_i)]$.

Projects

14.48. Refer to the **Disease outbreak** data set in Appendix C.10. Savings account status is the response variable and age, socioeconomic status, and city sector are the predictor variables. Cases 1–98 are to be utilized for developing the logistic regression model.

a. Fit logistic regression model (14.41) containing the predictor variables in first-order terms and interaction terms for all pairs of predictor variables. State the fitted response function.

b. Use the likelihood ratio test to determine whether all interaction terms can be dropped from the regression model; use $\alpha = .01$. State the alternatives, full and reduced models, decision rule, and conclusion. What is the approximate P-value of the test?

c. For logistic regression model in part (a), use backward elimination to decide which predictor variables can be dropped from the regression model. Control the α risk at .05 at each stage. Which variables are retained in the regression model?

14.49. Refer to the **Disease outbreak** data set in Appendix C.10 and Project 14.48. Logistic regression model (14.41) with predictor variables age and socioeconomic status in first-order terms is to be further evaluated.

a. Conduct the Hosmer-Lemeshow goodness of fit test for the appropriateness of the logistic regression function by forming five groups of approximately 20 cases each; use $\alpha = .05$. State the alternatives, decision rule, and conclusion. What is the approximate P-value of the test?

b. Obtain the deviance residuals and plot them against the estimated probabilities with a lowess smooth superimposed. What does the plot suggest about the adequacy of the fit of the logistic regression model?

c. Prepare an index plot of the diagonal elements of the estimated hat matrix (14.80). Use the plot to identify any outlying X observations.

d. To assess the influence of individual observations, obtain the delta chi-square statistic (14.85), the delta deviance statistic (14.86), and Cook's distance (14.87) for each observation. Plot each of these in separate index plots and identify any influential observations. Summarize your findings.

e. Construct a half-normal probability plot of the absolute deviance residuals and superimpose a simulated envelope. Are any cases outlying? Does the logistic model appear to be a good fit? Discuss.

f. To predict savings account status, you must identify the optimal cutoff. On the basis of the sample cases, find the total error rate, the error rate for persons with a savings account, and the error rate for persons with no savings account for the following cutoffs: .45, .50, .55, .60. Which of the cutoffs minimizes the total error rate? Are the two error rates for persons with and without savings accounts fairly balanced at this cutoff? Obtain the area under the ROC curve to assess the model's predictive power here. What do you conclude?

14.50. Refer to the **Disease outbreak** data set in Appendix C.10 and Project 14.49. The regression model identified in Project 14.49 is to be validated using cases 99–196.

a. Use the rule obtained in Project 14.49f to make a prediction for each of the holdout validation cases. What are the total and the two component prediction error rates for the validation data set? How do these error rates compare with those for the model-building data set in Project 14.49f?

b. Combine the model-building and validation data sets and fit the model identified in Project 14.49 to the combined data. Are the estimated coefficients and their estimated standard deviations similar to those obtained for the model-building data set? Should they be? Comment.

c. Based on the fitted regression model in part (b), obtain joint 90 percent confidence intervals for the odds ratios for age and socioeconomic status. Interpret your intervals.

14.51. Refer to the **SENIC** data set in Appendix C.1. Medical school affiliation is the response variable, to be coded $Y = 1$ if medical school affiliation and $Y = 0$ if no medical school affiliation. The pool of potential predictor variables includes age, routine chest X-ray ratio, average daily census, and number of nurses. All 113 cases are to be used in developing the logistic regression model.

a. Fit logistic regression model (14.41) containing all predictor variables in the pool in first-order terms and interaction terms for all pairs of predictor variables. State the fitted response function.

b. Test whether all interaction terms can be dropped from the regression model; use $\alpha = .05$. State the full and reduced models, decision rule, and conclusion. What is the approximate P-value of the test?

c. For logistic regression model (14.41) containing the predictor variables in first-order terms only, use forward stepwise regression to decide which predictor variables can be retained in the regression model. Control the α risk at .10 at each stage. Which variables should be retained in the regression model?

d. For logistic regression model (14.41) containing the predictor variables in first-order terms only, identify the best subset models using the AIC_p criterion and the SBC_p criterion. Does the use of these two criteria lead to the same model? Are either of the models identified the same as that found in part (c)?

14.52. Refer to the **SENIC** data set in Appendix C.1 and Project 14.51. Logistic regression model (14.41) with predictor variables age and average daily census in first-order terms is to be further evaluated.

a. Conduct Hosmer-Lemshow goodness of fit test for the appropriateness of the logistic regression function by forming five groups of approximately 23 cases each; use $\alpha = .05$. State the alternatives, decision rule, and conclusion. What is the approximate P-value of the test?

b. Obtain the deviance residuals and plot them against the estimated probabilities with a lowess smooth superimposed. What does the plot suggest about the adequacy of the fit of the logistic regression model?

c. Construct a half-normal probability plot of the absolute deviance residuals and superimpose a simulated envelope. Are any cases outlying? Does the logistic model appear to be a good fit? Discuss.

d. Prepare an index plot of the diagonal elements of the estimated hat matrix (14.80). Use the plot to identify any outlying X observations.

e. To assess the influence of individual observations, obtain the delta chi-square statistic (14.85), the delta deviance statistic (14.86), and Cook's distance (14.87) for each observation. Plot each of these in separate index plots and identify any influential observations. Summarize your findings.

f. To predict medical school affiliation, you must identify the optimal cutoff. For the sample cases, find the total error rate, the error rate for hospitals with medical school affiliation, and the error rate for hospitals without medical school affiliation for the following cutoffs: .30, .40, .50, .60. Which of the cutoffs minimizes the total error rate? Are the two error rates for hospitals with and without medical school affiliation fairly balanced at this cutoff? Obtain the area under the ROC curve to assess the model's predictive power here. What do you conclude?

g. Estimate by means of an approximate 90 percent confidence interval the odds of a hospital having medical school affiliation for hospitals with average age of patients of 55 years and average daily census of 500 patients.

14.53. Refer to **Annual dues** Problem 14.7. Obtain a simulated envelope and superimpose it on the half-normal probability plot of the absolute deviance residuals. Are there any indications that the fitted model is not appropriate? Are there any outlying cases? Discuss.

14.54. Refer to **Annual dues** Problem 14.7. In order to assess the appropriateness of large-sample inferences here, employ the following parametric bootstrap procedure: For each of the 30 cases, generate a Bernoulli outcome (0, 1), using the estimated probability $\hat{\pi}_i$ for the original X_i level according to the fitted model. Fit the logistic regression model to the bootstrap sample and obtain the bootstrap estimates b_0^* and b_1^*. Repeat this procedure 500 times. Compute the mean and standard deviation of the 500 bootstrap estimates b_0^*, and do the same for b_1^*. Plot separate histograms of the bootstrap distributions of b_0^* and b_1^*. Are these distributions approximately normal? Compare the point estimates b_0 and b_1 and their estimated standard deviations obtained in the original fit to the means and standard deviations of the bootstrap distributions. What do you conclude about the appropriateness of large-sample inferences here? Discuss.

14.55. Refer to **Car purchase** Problem 14.13. Obtain a simulated envelope and superimpose it on the half-normal probability plot of the absolute deviance residuals. Are there any indications that the fitted model is not appropriate? Are there any outlying cases? Discuss.

14.56. Refer to **Car purchase** Problem 14.13. In order to assess the appropriateness of large-sample inferences here, employ the following parametric bootstrapping procedure: For each of the 33 cases, generate a Bernoulli outcome (0, 1), using the estimated probability $\hat{\pi}_i$ for the original levels of the predictor variables according to the fitted model. Fit the logistic regression model to the bootstrap sample. Repeat this procedure 500 times. Compute the mean and standard deviation of the 500 bootstrap estimates b_1^*, and do the same for b_2^*. Plot separate histograms of the bootstrap distributions of b_1^* and b_2^*. Are these distributions approximately normal? Compare the point estimates b_1 and b_2 and their estimated standard deviations obtained in the original fit to the means and standard deviations of the bootstrap distributions. What do you conclude about the appropriateness of large-sample inferences here? Discuss.

14.57. Refer to the **SENIC** data set in Appendix C.1. Region is the nominal level response variable coded 1 = NE, 2 = NC, 3 = S, and 4 = W. The pool of potential predictor variables includes age, routine chest X-ray ratio, number of beds, medical school affiliation, average daily census, number of nurses, and available facilities and services. All 113 hospitals are to be used in developing the polytomous logistic regression model.

a. Fit polytomous regression model (14.99) using response variable region with 1 = NE as the referent category. Which predictors appear to be most important? Interpret the results.

b. Conduct a likelihood ratio test to determine if the three parameters corresponding to age can be dropped from the nominal logistic regression model. Control α at .05. State the full and reduced models, decision rule, and conclusion. What is the approximate P-value of the test?

c. Conduct a likelihood ratio test to determine if all parameters corresponding to age and available facilities and services can be dropped from the nominal logistic regression model. Control α at .05. State the full and reduced models, decision rule, and conclusion. What is the approximate P-value of the test?

d. For the full model in part (a), carry out separate binary logistic regressions for each of the three comparisons with the referent category, as described at the top of page 612. How do the slope coefficients compare to those obtained in part (a).

e. For each of the separate binary logistic regressions carried out in part (d), obtain the deviance residuals and plot them against the estimated probabilities with a lowess smooth superimposed. What do the plots suggest about the adequacy of the fit of the binary logistic regression models?

f. For each of the separate binary logistic regressions carried out in part (d), obtain the delta chi-square statistic (14.85), the delta deviance statistic (14.86), and Cook's distance (14.87) for each observation. Plot each of these in separate index plots and identify any influential observations. Summarize your findings.

14.58. Refer to the **CDI** data set in Appendix C.2. Region is the nominal level response variable coded 1 = NE, 2 = NC, 3 = S, and 4 = W. The pool of potential predictor variables includes population density (total population/land area), percent of population aged 18–34, percent of population aged 65 or older, serious crimes per capita (total serious crimes/total population), percent high school graduates, percent bachelor's degrees, percent below poverty level, percent unemployment, and per capita income. The even-numbered cases are to be used in developing the polytomous logistic regression model.

a. Fit polytomous regression model (14.99) using response variable region with 1 = NE as the referent category. Which predictors appear to be most important? Interpret the results.

b. Conduct a series of likelihood ratio tests to determine which predictors, if any, can be dropped from the nominal logistic regression model. Control α at .01 for each test. State the alternatives, decision rules, and conclusions.

c. For the full model in part (a), carry out separate binary logistic regressions for each of the three comparisons with the referent category, as described at the top of page 612. How do the slope coefficients compare to those obtained in part (a).

d. For each of the separate binary logistic regressions carried out in part (c), obtain the deviance residuals and plot them against the estimated probabilities with a lowess smooth superimposed. What do the plots suggest about the adequacy of the fit of the binary logistic regression models?

e. For each of the separate binary logistic regressions carried out in part (d), obtain the delta chi-square statistic (14.85), the delta deviance statistic (14.86), and Cook's distance (14.87) for each observation. Plot each of these in separate index plots and identify any influential observations. Summarize your findings.

14.59. Refer to the **Prostate cancer** data set in Appendix C.5. Gleason score (variable 9) is the ordinal level response variable, and the pool of potential predictor variables includes PSA level, cancer volume, weight, age, benign prostatic hyperplasia, seminal vesicle invasion, and capsular penetration (variables 2 through 8).

a. Fit the proportional odds model (14.105). Which predictors appear to be most important? Interpret the results.

b. Conduct a series of Wald tests to determine which predictors, if any, can be dropped from the nominal logistic regression model. Control α at .05 for each test. State the alternatives, decision rule, and conclusion. What is the approximate P-value of the test?

c. Starting with the full model of part (a), use backward elimination to decide which predictor variables can be dropped from the ordinal regression model. Control the α risk at .05 at each stage. Which variables should be retained?

d. For the model in part (c), carry out separate binary logistic regressions for each of the two binary variables $Y_i^{(1)}$ and $Y_i^{(2)}$, as described at the top of page 617. How do the estimated coefficients compare to those obtained in part (c)?

e. For each of the separate binary logistic regressions carried out in part (d), obtain the deviance residuals and plot them against the estimated probabilities with a lowess smooth superimposed. What do the plots suggest about the adequacy of the fit of the binary logistic regression models?

f. For each of the separate binary logistic regressions carried out in part (d), obtain the delta chi-square statistic (14.85), the delta deviance statistic (14.86), and Cook's distance (14.87) for each observation. Plot each of these in separate index plots and identify any influential observations. Summarize your findings.

14.60. Refer to the **Real estate sales** data set in Appendix C.7. Quality of construction (variable 10) is the ordinal level response variable, and the pool of potential predictor variables includes sales price, finished square feet, number of bedrooms, number of bathrooms, air conditioning, garage size, pool, year built, lot size, and adjacent to highway (variables 2 through 9 and 12 through 13).

a. Fit the proportional odds model (14.105). Which predictors appear to be most important? Interpret the results.

b. Conduct a series of Wald tests to determine which predictors, if any, can be dropped from the nominal logistic regression model. Control α at .01 for each test. State the alternatives, decision rules, and conclusions. Which predictors should be retained?

c. Starting with the full model of part (a), use backward elimination to decide which predictor variables can be dropped from the ordinal regression model. Control the α risk at .05 at each stage. Which variables should be retained?

d. For the model obtained in part (c), carry out separate binary logistic regressions for each of the two binary variables $Y_i^{(1)}$ and $Y_i^{(2)}$, as described at the top of page 617. How do the estimated coefficients compare to those obtained in part (a)?

e. For each of the separate binary logistic regressions carried out in part (d), obtain the deviance residuals and plot them against the estimated probabilities with a lowess smooth superimposed. What do the plots suggest about the adequacy of the fit of the binary logistic regression models?

f. For each of the separate binary logistic regressions carried out in part (d), obtain the delta chi-square statistic (14.85), the delta deviance statistic (14.86), and Cook's distance (14.87) for each observation. Plot each of these in separate index plots and identify any influential observations. Summarize your findings.

14.61. Refer to the **Ischemic heart disease** data set in Appendix C.9. The response is the number of emergency room visits (variable 7) and the pool of potential predictor variables includes total cost, age, gender, number of interventions, number of drugs, number of complications, number of comorbidities, and duration (variables 2 through 6 and 8 through 10).

a. Obtain the fitted the Poisson regression model (14.113) with the response function $\mu(\mathbf{X}, \boldsymbol{\beta}) = \exp(\mathbf{X}'\boldsymbol{\beta})$. State the estimated regression coefficients, their estimated standard deviations, and the estimated response function.

b. Obtain the deviance residuals (14.118) and plot them against the estimated model probabilities with a lowess smooth superimposed. What does the plot suggest about the adequacy of the fit of the Poisson regression model?

c. Conduct a series of Wald tests to determine which predictors, if any, can be dropped fro the nominal logistic regression model. Control α at .01 for each test. State the alternatives decision rules, and conclusions.

d. Assuming that the fitted model in part (a) is appropriate, use the likelihood ratio test t determine whether duration, coomplications, and comorbidities can be dropped from th model; control α at .05. State the full and reduced models, decision rule, and conclusion.

e. Use backward elimination to decide which predictor variables can be dropped from th regression model. Control the α risk at .10 at each stage. Which variables are retained?

Case Studies

14.62. Refer to the **IPO** data set in Appendix C.11. Carry out a complete analysis of this data set, where the response of interest is venture capital funding, and the pool of predictors includes fa value of the company, number of shares offered, and whether or not the company underwen a leveraged buyout. The analysis should consider transformations of predictors, inclusion o second-order predictors, analysis of residuals and influential observations, model selection, goodness of fit evaluation, and the development of an ROC curve. Model validation should also be employed. Document the steps taken in your analysis, and assess the strengths and weaknesses of your final model.

14.63. Refer to the **Real estate sales** data set in Appendix C.7. Create a new binary response variable Y, called high quality construction, by letting $Y = 1$ if quality (variable 10) equals 1, and $Y = 0$ otherwise (i.e., if quality equals 2 or 3). Carry out a complete logistic regression analysis, where the response of interest is high quality construction (Y), and the pool of predictors includes sales price, finished square feet, number of bedrooms, number of bathrooms, air conditioning, garage size, pool, year built, style, lot size, and adjacent to highway (variables 2 through 9 and 11 through 13). The analysis should consider transformations of predictors, inclusion of second-order predictors, analysis of residuals and influential observations, model selection, goodness of fit evaluation, and the development of an ROC curve. Develop a prediction rule for determining whether the quality of construction is predicted to be of high quality or not. Model validation should also be employed. Document the steps taken in your analysis, and assess the strengths and weaknesses of your final model.

14.64. Refer to the **Prostate cancer** data set in Appendix C.5. Create a new binary response variable Y, called high-grade cancer, by letting $Y = 1$ if Gleason score (variable 9) equals 8, and $Y = 0$ otherwise (i.e., if Gleason score equals 6 or 7). Carry out a complete logistic regression analysis, where the response of interest is high-grade cancer (Y), and the pool of predictors includes PSA level, cancer volume, weight, age, benign prostatic hyperplasia, seminal vesicle invasion, and capsular penetration (variables 2 through 8). The analysis should consider transformations of predictors, inclusion of second-order predictors, analysis of residuals and influential observations, model selection, goodness of fit evaluation, and the development of an ROC curve. Develop a prediction rule for determining whether the grade of disease is predicted to be high grade or not. Model validation should also be employed. Document the steps taken in your analysis, and assess the strengths and weaknesses of your final model.