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:

$$\begin{gathered}Y_i \text{ are independent Poisson random variables with expected} \\ \text{values } \mu_i \text{, where:} \\ \mu_i = \mu(\mathbf{X}_i, \boldsymbol{\beta})\end{gathered} \tag{14.113}$$

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:

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

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:

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

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

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

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.

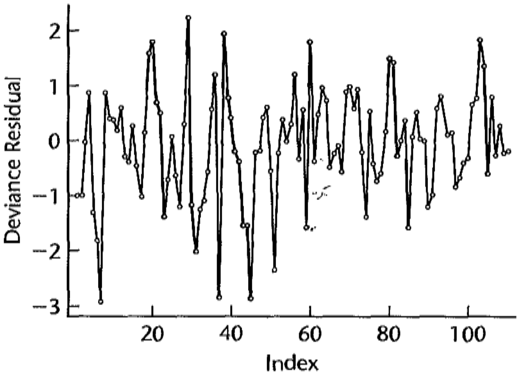
Likelihood ratio test statistics (14.60) were calculated for each of the individual regression coefficients. These G^2 test statistics are shown in Table 14.15b, together with their associated P-values, each based on the chi-square distribution with one degree of freedom. We note from the P-values that 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 from 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 to determine whether duration, coomplications, and comorbidities can be dropped from the 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 the 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 face value of the company, number of shares offered, and whether or not the company underwent a leveraged buyout. 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. 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.