CHAPTER 12

LOGISTIC REGRESSION

12.1 INTRODUCTION

In our discussion of regression analysis so far the response variable Y has been regarded as a continuous quantitative variable. The predictor variables, however, have been both quantitative, as well as qualitative. Indicator variables, which we have described earlier, fall into the second category. There are situations, however, where the response variable is qualitative. In this chapter we present methods for dealing with this situation. The methods presented in this chapter are very different from the method of least squares considered in earlier chapters.

Consider a procedure in which individuals are selected on the basis of their scores in a battery of tests. After five years the candidates are classified as "good" or "poor". We are interested in examining the ability of the tests to predict the job performance of the candidates. Here the response variable, performance, is dichotomous. We can code "good" as 1 and "poor" as 0, for example. The predictor variables are the scores in the tests.

In a study to determine the risk factors for cancer, health records of several people were studied. Data were collected on several variables, such as age, sex, smoking, diet, and the family's medical history. The response variable was, the person had cancer ($Y = 1$), or did not have cancer ($Y = 0$).

Regression Analysis by Example, Fourth Edition. By Samprit Chatterjee and Ali S. Hadi

In the financial community the "health" of a business is of primary concern. The response variable is solvency of the firm (bankrupt = 0, solvent =1), and the predictor variables are the various financial characteristics associated with the firm. Situations where the response variable is a dichotomous variable are quite common and occur extensively in statistical applications.

12.2 MODELING QUALITATIVE DATA

The qualitative data with which are dealing, the binary response variable, can always be coded as having two values, 0 or 1. Rather than predicting these two values we try to model the probabilities that the response takes one of these two values. The limitation of the previously considered standard linear regression model is obvious.

We illustrate this point by considering a simple regression problem, in which we have only one predictor. The same considerations hold for the multiple regression case. Let π denote the probability that $Y = 1$ when $X = x$. If we use the standard linear model to describe π, then our model for the probability would be

$$\pi = Pr(Y = 1|X = x) = \beta_0 + \beta_1 x + \varepsilon. \tag{12.1}$$

Since π is a probability it must lie between 0 and 1. The linear function given in (12.1) is unbounded, and hence cannot be used to model probability. There is another reason why ordinary least squares method is unsuitable. The response variable Y is a binomial random variable, consequently its variance will be a function of π, and depends on X. The assumption of equal variance (homoscedasticity) does not hold. We could use the weighted least squares, but there are problems with that approach. The values of π are not known. In order to use weighted least squares approach, we will have to start with an initial guess for the value of π, and then iterate. Instead of this complex method we will describe an alternative method for modeling probabilities.

12.3 THE LOGIT MODEL

The relationship between the probability π and X can often be represented by a *logistic response function*. It resembles a S-shaped curve, a sketch of which is given in Figure 12.1. The probability π initially increases slowly with increase in X, then the increase accelerates, finally stabilizes, but does not increase beyond 1. Intuitively this makes sense. Consider the probability of a questionnaire being returned as a function of cash reward, or the probability of passing a test as a function of the time put in studying for it.

The shape of the S-curve given in Figure 12.1 can be reproduced if we model the probabilities as follows:

$$\pi = \Pr(Y = 1|X = x) = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}}, \tag{12.2}$$

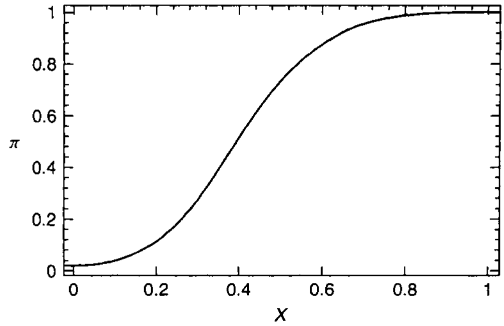

Figure 12.1 Logistic response function.

where *e* is the base of the natural logarithm. The probabilities here are modeled by the distribution function (cumulative probability function) of the logistic distribution. There are other ways of modeling the probabilities that would also produce the S-curve. The cumulative distribution of the normal curve has also been used. This gives rise to the *probit* model. We will not discuss the probit model here, as we consider the logistic model simpler and superior to the probit model.

The logistic model can be generalized directly to the situation where we have several predictor variables. The probability π is modeled as

$$\begin{aligned}\pi &= \Pr(Y = 1|X_1 = x_1, \ldots, X_p = x_p)\\ &= \frac{e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\ldots+\beta_p x_p}}{1+e^{\beta_0+\beta_1 x_1+\ldots+\beta_p x_p}}. \qquad (12.3)\end{aligned}$$

The equation in (12.3) is called the *logistic regression function*. It is nonlinear in the parameters β_0, β_1, ..., β_p. However, it can be linearized by the *logit transformation*.[1] Instead of working directly with π we work with a transformed value of π. If π is the probability of an event happening, the ratio $\pi/(1-\pi)$ is called the *odds ratio* for the event. Since

$$1-\pi = \Pr(Y = 0|X_1 = x_1, \ldots, X_p = x_p) = \frac{1}{1+e^{\beta_0+\beta_1 x_1+\ldots+\beta_p x_p}},$$

then

$$\frac{\pi}{1-\pi} = e^{\beta_0+\beta_1 x_1+\ldots+\beta_p x_p}. \qquad (12.4)$$

Taking the natural logarithm of both sides of (12.4), we obtain

$$\begin{aligned}g(x_1, \ldots, x_p) &= log\left(\frac{\pi}{1-\pi}\right)\\ &= \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p. \qquad (12.5)\end{aligned}$$

[1] See Chapter 6 for transformation of variables.

The logarithm of the odds ratio is called the *logit*. It can be seen from (12.5) that the logit transformation produces a linear function of the parameters $\beta_0, \beta_1, \ldots, \beta_p$. Note also that while the range of values of π in (12.3) is between 0 and 1, the range of values of $log(\pi/(1-\pi))$ is between $-\infty$ and $+\infty$, which makes the logits (the logarithm of the odds ratio) more appropriate for linear regression fitting.

Modeling the response probabilities by the logistic distribution and estimating the parameters of the model given in (12.3) constitutes fitting a logistic regression. In logistic regression the fitting is carried out by working with the logits. The logit transformation produces a model that is linear in the parameters. The method of estimation used is the *maximum likelihood* method. The maximum likelihood estimates are obtained numerically, using an iterative procedure. Unlike least squares fitting, no closed-form expression exists for the estimates of the parameters. We will not go into the computational aspects of the problem but refer the reader to McCullagh and Nelder (1983), Seber (1984), and Hosmer and Lemeshow (1989).

To fit a logistic regression in practice a computer program is essential. Most regression packages have a logistic regression option. After the fitting one looks at the same set of questions that are usually considered in linear regression. Questions about the suitability of the model, the variables to be retained, and goodness of fit are all considered. Tools used are not the usual R^2, t, and F tests, the ones employed in least squares regression, but others which provide answers to these same questions. Hypothesis testing is done by different methods, since the method of estimation is maximum likelihood as opposed to least squares. Information Criteria such as AIC and BIC can be used for model selection. Instead of SSE, the logarithm of the likelihood for the fitted model is used. An explicit formula is given in Section 12.6.

12.4 EXAMPLE: ESTIMATING PROBABILITY OF BANKRUPTCIES

Detecting ailing financial and business establishments is an important function of audit and control. Systematic failure to do audit and control can lead to grave consequences, such as the savings-and-loan fiasco of the 1980s in the United States. Table 12.1 gives some of the operating financial ratios of 33 firms that went bankrupt after 2 years and 33 that remained solvent during the same period. The data can also be found in the book's Web site.[2] A multiple logistic regression model is fitted using variables X_1, X_2, and X_3. The output from fitting the model is given in Table 12.2.

Three financial ratios were available for each firm:

$$\begin{aligned} X_1 &= \frac{\text{Retained Earnings}}{\text{Total Assets}}, \\ X_2 &= \frac{\text{Earnings Before Interest and Taxes}}{\text{Total Assets}}, \\ X_3 &= \frac{\text{Sales}}{\text{Total Assets}}. \end{aligned}$$

[2] http://www.ilr.cornell.edu/˜hadi/RABE4

The response variable is defined as

$$Y = \begin{cases} 0, & \text{if bankrupt after 2 years,} \\ 1, & \text{if solvent after 2 years.} \end{cases}$$

Table 12.2 has a certain resemblance to the standard regression output. Some of the output serve similar functions. We now describe and interpret the output obtained from fitting a logistic regression. If π denotes the probability of a firm remaining solvent after 2 years, the fitted logit is given by:

$$\hat{g}(x_1, \ldots, x_p) = -10.15 + 0.33\, x_1 + 0.18\, x_2 + 5.09\, x_3. \qquad (12.6)$$

This corresponds to the fitted regression equation in standard analysis. Here instead of predicting Y we obtain a model to predict the logits, $log(\pi/(1-\pi))$. From the logits, after transformation, we can get the predicted probabilities. The constant and the coefficients are read directly from the second column in the table. The standard errors (s.e.) of the coefficients are given in the third column. The fourth column headed by Z is the ratio of the coefficient and the standard deviation. The Z is sometimes referred to as the Wald Statistic (Test). The Z corresponding to the coefficient of X_2 is obtained from dividing 0.181 by 0.107. In the standard regression this would be the t-test. This ratio for the logistic regression has a normal distribution as opposed to a t-distribution that we get in linear regression. The fifth column gives the p-value corresponding to the observed Z value, and should be interpreted like any p-value (see Chapters 2 and 3). These p-values are used to judge the significance of the coefficient. Values smaller than 0.05 would lead us to conclude that the coefficient is significantly different from 0 at the 5% significance level. From the p-values in Table 12.2, we see that none of the variables individually are significant for predicting the logits of the observations.

In the standard regression output the regression coefficients have a simple interpretation. The regression coefficient of the jth predictor variable X_j is the expected change in Y for unit change in X_j when other variables are held fixed. The coefficient of X_2 in (12.6) is the expected change in the logit for unit change in X_2 when the other variables are held fixed. The coefficients of a logistic regression fit have another interpretation that is of major practical importance. Keeping X_1 and X_3 fixed, for unit increase in X_2 the relative odds of

$$\frac{\text{Pr(Firm solvent after 2 years)}}{\text{Pr(Firm bankrupt)}}$$

is multiplied by $e^{\hat{\beta}_2} = e^{0.181}$ =1.198, that is there is an increase of 20%. These values for each of the variables is given in the sixth column headed by Odds Ratio. They represent the change in odds ratio for unit change of a particular variable while the others are held constant. The change in odds ratio for unit change in variable X_j, while the other variables are held fixed, is $e^{\hat{\beta}_j}$. If X_j was a binary variable, taking values 1 or 0, then $e^{\hat{\beta}_j}$ would be the actual value of the odds ratio rather than the change in the value of the odds ratio.

Table 12.1 Financial Ratios of Solvent and Bankrupt Firms

Row	Y	X_1	X_2	X_3	Row	Y	X_1	X_2	X_3
1	0	−62.8	−89.5	1.7	34	1	43.0	16.4	1.3
2	0	3.3	−3.5	1.1	35	1	47.0	16.0	1.9
3	0	−120.8	−103.2	2.5	36	1	−3.3	4.0	2.7
4	0	−18.1	−28.8	1.1	37	1	35.0	20.8	1.9
5	0	−3.8	−50.6	0.9	38	1	46.7	12.6	0.9
6	0	−61.2	−56.2	1.7	39	1	20.8	12.5	2.4
7	0	−20.3	−17.4	1.0	40	1	33.0	23.6	1.5
8	0	−194.5	−25.8	0.5	41	1	26.1	10.4	2.1
9	0	20.8	−4.3	1.0	42	1	68.6	13.8	1.6
10	0	−106.1	−22.9	1.5	43	1	37.3	33.4	3.5
11	0	−39.4	−35.7	1.2	44	1	59.0	23.1	5.5
12	0	−164.1	−17.7	1.3	45	1	49.6	23.8	1.9
13	0	−308.9	−65.8	0.8	46	1	12.5	7.0	1.8
14	0	7.2	−22.6	2.0	47	1	37.3	34.1	1.5
15	0	−118.3	−34.2	1.5	48	1	35.3	4.2	0.9
16	0	−185.9	−280.0	6.7	49	1	49.5	25.1	2.6
17	0	−34.6	−19.4	3.4	50	1	18.1	13.5	4.0
18	0	−27.9	6.3	1.3	51	1	31.4	15.7	1.9
19	0	−48.2	6.8	1.6	52	1	21.5	−14.4	1.0
20	0	−49.2	−17.2	0.3	53	1	8.5	5.8	1.5
21	0	−19.2	−36.7	0.8	54	1	40.6	5.8	1.8
22	0	−18.1	−6.5	0.9	55	1	34.6	26.4	1.8
23	0	−98.0	−20.8	1.7	56	1	19.9	26.7	2.3
24	0	−129.0	−14.2	1.3	57	1	17.4	12.6	1.3
25	0	−4.0	−15.8	2.1	58	1	54.7	14.6	1.7
26	0	−8.7	−36.3	2.8	59	1	53.5	20.6	1.1
27	0	−59.2	−12.8	2.1	60	1	35.9	26.4	2.0
28	0	−13.1	−17.6	0.9	61	1	39.4	30.5	1.9
29	0	−38.0	1.6	1.2	62	1	53.1	7.1	1.9
30	0	−57.9	0.7	0.8	63	1	39.8	13.8	1.2
31	0	−8.8	−9.1	0.9	64	1	59.5	7.0	2.0
32	0	−64.7	−4.0	0.1	65	1	16.3	20.4	1.0
33	0	−11.4	4.8	0.9	66	1	21.7	−7.8	1.6

Table 12.2 Output from the Logistic Regression Using X_1, X_2, and X_3

					Odds	95% C.I.	
Variable	Coeff.	s.e.	Z-test	p-value	Ratio	Lower	Upper
Constant	−10.15	10.84	−0.94	0.349			
X_1	0.33	0.30	1.10	0.27	1.39	0.77	2.51
X_2	0.18	0.11	1.69	0.09	1.20	0.97	1.48
X_3	5.09	5.08	1.00	0.32	161.98	0.01	3.43×10^6
Log-Likelihood = −2.906			G = 85.683		$d.f.$ = 3	p-value < 0.000	

The 95% confidence intervals of the odds ratios are given in the last two columns of the table. If the confidence interval does not contain the value 1 the variable has a significant effect on the odds ratio. If the interval is below 1 the variable lowers significantly the relative odds. On the other hand, if the interval lies above 1 the relative odds is significantly increased by the variable.

To see whether the variables collectively contribute in explaining the logits a test that examines whether the coefficients $\beta_1, \ldots, \beta_p$ are all zero is performed. This corresponds to the case in multiple regression analysis where we test whether all the regression coefficients can be taken to be zero. The statistic G given at the bottom of Table 12.2 performs that task. The statistic G has a chi-square distribution. The p-value is considerably smaller than .05, and indicates that the variables collectively influence the logits.

12.5 LOGISTIC REGRESSION DIAGNOSTICS

After fitting a logistic regression model certain diagnostic measures can be examined for the detection of outliers, high leverage points, influential observations, and other model deficiencies. The diagnostic measures developed in Chapter 4 for the standard linear regression model can be adapted to the logistic regression model. Regression packages with a logistic regression option usually give various diagnostic measures. These include:

1. The estimated probabilities $\hat{\pi}_i$, $i = 1, \ldots, n$.

2. One or more types of residuals, for example, the *standardized deviance residuals*, DR_i, and the *standardized Personian residuals*, PR_i, $i = 1, \ldots, n$.

3. The *weighted leverages*, p^*_{ii}, which measure the potential effects of the observations in the predictor variables on the obtained logistic regression results.

4. The scaled difference in the regression coefficients when the ith observation is deleted: DBETA_i, $i = 1, \ldots, n$.

5. The change in the chi-squared statistics G when the ith observation is deleted: DFG_i, $i = 1, \ldots, n$.

The formulas and derivations of these measures are beyond the scope of this book. The interested reader is referred to Pregibon (1981), Landwehr, Pregibon, and Shoemaker (1984), Hosmer and Lemeshow (1989) and the references therein. The above measures, however, can be used in the same way as the corresponding measures obtained from a linear fit (Chapter 4). For example, the following graphical displays can be examined:

1. The scatter plot of DR_i versus $\hat{\pi}_i$.

2. The scatter plot of PR_i versus $\hat{\pi}_i$.

3. The index plots of DR_i, DBETA_i, DG_i, and p^*_{ii}.

As an illustrative example using the Bankruptcy data, the index plots of DR_i, DBETA_i, and DG_i obtained from the fitted logistic regression model in (12.6), are shown in Figures 12.2, 12.3, and 12.4, respectively. It can easily be seen from these graphs that observations 9, 14, 52, and 53 are unusual and that they may have undue influence on the logistic regression results. We leave it as an exercise for the reader to determine if their deletion would make a significant difference in the results and the conclusion drawn from the analysis.

12.6 DETERMINATION OF VARIABLES TO RETAIN

In the analysis of the Bankruptcy data we have determined so far that the variables X_1, X_2, and X_3, collectively have explanatory power. Do we need all three variables? This is analogous to the problem of variable selection in multiple regression that was discussed in Chapter 11. Instead of looking at the reduction in the error sum of squares we look at the change in the likelihood (more precisely, the logarithm of the likelihood) for the two fitted models. The reason for this is that in logistic regression the fitting criterion is the likelihood, whereas in least squares it is the sum of squares. Let $L(p)$ denote the logarithm of the likelihood when we have a model with p variables and a constant. Similarly, let $L(p+q)$ be the logarithm of the likelihood for a model in which we have $p+q$ variables and a constant. To see whether the q additional variables contribute significantly we look at $2(L(p+q) - L(p))$. This quantity is twice the difference between the log-likelihood for the two models. This difference is distributed as a chi-square variable with q degrees of freedom (see Table A.3).

The magnitude of this quantity determines the significance of the test. A small value of chi-square would lead to the conclusion that the q variables do not add significantly to the improvement in prediction of the logits, and is therefore not necessary in the model. A large value of chi-square would call for the retention of the q variables in the model. The critical value is determined by the significance

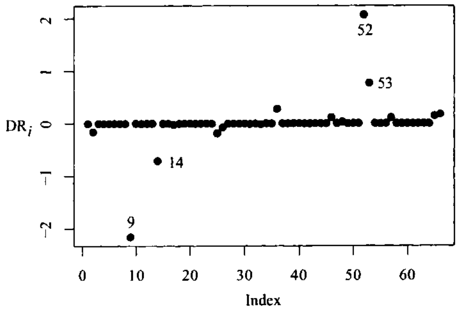

Figure 12.2 Bankruptcy data: Index plot of DR$_i$, the standardized deviance residuals.

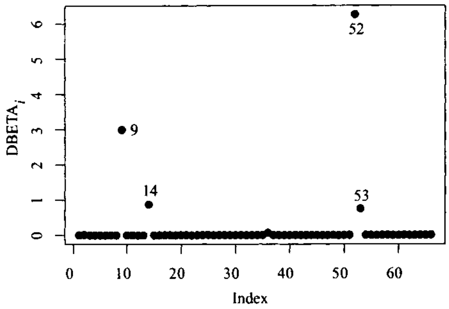

Figure 12.3 Bankruptcy data: Index plot of DBETA$_i$, the scaled difference in the regression coefficients when the ith observation is deleted.

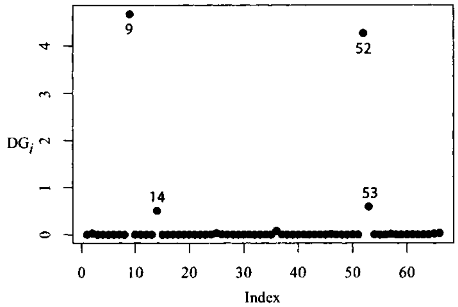

Figure 12.4 Bankruptcy data: Index plot of DG$_i$, the change in the chi-squared statistics G when the ith observation is deleted.

level of the test. This test procedure is valid when n, the number of observations available for fitting the model, is large.

An idea of the predictive power of a variable for possible inclusion in the logistic model can be obtained from a simple graphical plot. Side-by-side boxplots are constructed for each of the explanatory variable. Side- by-side boxplot will indicate

Table 12.3 Output From the Logistic Regression Using X_1 and X_2

					Odds	95% C.I.	
Variable	Coefficient	s.e.	Z-test	p-value	Ratio	Lower	Upper
Constant	−0.550	0.951	−0.58	0.563			
X_1	0.157	0.075	2.10	0.036	1.17	1.01	1.36
X_2	0.195	0.122	1.59	0.112	1.21	0.96	1.54
Log-Likelihood = −4.736		G = 82.024		$d.f.$ = 2	p-value < 0.000		

Table 12.4 Output from the Logistic Regression Using X_1

					Odds	95% C.I.	
Variable	Coefficient	s.e.	Z-test	p-value	Ratio	Lower	Upper
Constant	−1.167	0.816	−1.43	0.153			
X_1	0.177	0.057	3.09	0.002	1.19	1.07	1.33
Log-Likelihood = −7.902		G = 75.692		$d.f.$ = 1	p-value < 0.000		

the variables that may be useful for this purpose. Variables with boxplots different for the two groups are likely candidates. Note that this does not take into account the correlation between the variables. The formal procedure described above takes into account the correlations. With a large number of explanatory variables the boxplots provide a quick screening procedure.

In the Bankruptcy data we are analyzing, let us see if the variable X_3 can be deleted without degrading the model. We want to answer the question: Should the variable X_3 be retained in the model? We fit a logistic regression using X_1 and X_2. The results are given in Table 12.3. The log-likelihood for the model with X_1, X_2, and X_3 is −2.906, whereas with only X_1 and X_2 it is −4.736. Here $p = 2$ and $q = 1$, and $2(L(3) - L(2)) = 3.66$. This is a chi-square variable with 1 degree of freedom. From Table A.3, we find that the 5% critical value of the chi-square distribution with 1 degree of freedom is 3.84. At the 5% level we can conclude that the variable X_3 can be deleted without affecting the effectiveness of the model.

Let us now see if we can delete X_2. The result of regressing Y on X_1 is given in Table 12.4. The resulting log-likelihood is −7.902. The test statistic, which we have described earlier, has a value of 6.332. This is distributed as a chi-square random variable with 1 degree of freedom. The 5% value, as we saw earlier, was 3.84. The analysis indicates that we should not delete X_2 from our model. The p-value for this test, as can be verified, is 0.019. To predict probabilities of bankruptcies of firms in our data we should include both X_1 and X_2 in our model.

The procedure that we have outlined above enables us to test any *nested model.* A set of models are said to be nested if they can be obtained from a larger model as

Table 12.5 The AIC and BIC Criteria for Various Logistic Regression models

Variables	AIC	BIC
$X_1X_2X_3$	13.81	22.57
X_1X_2	15.47	22.04
X_1X_3	18.12	24.69
X_2X_3	33.40	39.97
X_1	19.80	24.18
X_2	34.50	38.88
X_3	92.46	96.84
None	93.5	95.69

special cases. The methodology is similar to that used in analyzing nested models in multiple regression. The only difference is that here our test statistic is based on log of the likelihood instead of sum of squares.

The AIC and BIC criteria discussed in Section 11.5.3 can be used to judge the suitability of various logistic models, and thereby the desirability of retaining a variable in the model. In the context of p-term logistic logistic regression, AIC and BIC are:

$$\begin{aligned} \text{AIC} &= -2(\text{Log-Likelihood of the Fitted Model}) + 2p, \qquad (12.7)\\ \text{BIC} &= -2(\text{Log-Likelihood of the Fitted Model}) + p\log n, \qquad (12.8) \end{aligned}$$

where p denotes the number of variables in the model. Table 12.5 shows AIC and BIC for all possible models. The best AIC model is the one that includes all three variables (lowest AIC). While BIC picks X_1X_2 as the best model, but the one containing all three variables is equally adequate. The BIC for the two top models differ by less than 2.

12.7 JUDGING THE FIT OF A LOGISTIC REGRESSION

The overall fit of a multiple regression model is judged, for example, by the value of R^2 from the fitted model. No such simple satisfactory measure exists for logistic regression. Some ad hoc measures have been proposed which are based on the ratio of likelihoods. Most of these are functions of the ratio of the likelihood for the model and the likelihood of the data under a binomial model. These measures are not particularly informative and we will consider a different approach.

The logistic regression equation attempts to model probabilities for the two values of Y (0 or 1). To judge how well the model is doing we will determine the number of observations in the sample that the model is classifying correctly. Our approach will be to fit the logistic model to the data, and calculate the fitted logits. From the fitted logits we will calculate the fitted probabilities for each observation. If the fitted probability for an observation is greater than 0.5 we will assign it to

Group 1 ($Y = 1$), and if less than 0.5 we will classify it in Group 0 ($Y = 0$). We will then determine what proportion of the data is classified correctly. A high proportion of correct classification will indicate to us that the logistic model is working well. A low proportion of correct classification will indicate poor performance.

Different cutoff values, other than 0.5, have been suggested in the literature. In most practical situations, without any auxiliary information, such as the relative cost of misclassification or the relative frequency of the two categories in the population, 0.5 is recommended as a cutoff value.

A slightly more problematical question is how high the correct classification probability has to be before logistic regression is thought to be effective. Suppose that in sample of size n there are n_1 observations from Group 1, and n_2 from Group 2. If we classify all the observations into one group or the other, then we will get either n_1/n or n_2/n proportions of observations classified correctly. As a base level for correct classification we can take the $max(n_1/n, n_2/n)$. The proportion of observation classified correctly by the logistic regression should be much higher than the base level for the logistic model to be deemed useful.

For the Bankruptcy data that we have been analyzing logistic regression performs very well. Using variables X_1 and X_2, we find that the model misclassifies one observation from the solvent group (observation number 36), and one observation from the bankruptcy group (observation number 9). The overall correct classification rate $(64/66) = 0.97$. This is considerably higher than the base level rate of 0.5.

The concept of overall correct classification for the observed sample to judge the adequacy of the logistic model that we have discussed has been generalized. This generalization is used to produce a statistic to judge the fit of the logistic model. It is sometimes called the *Concordance Index* and is denoted by C. This statistic is calculated by considering all possible pairs formed by taking one observation from each group. Each of the pairs is then classified by using the fitted model. The Concordance Index is the percent of all possible pairs that is classified correctly. Thus, C lies between 0.5 and 1. Values of C close to 0.5 shows the logistic model performing poorly (no better than guessing). The value of C for the logistic model with $X_1X_2X_3$ is 0.99. Several currently available software computes the value of C.

The observed correct classification rate should be treated with caution. In practice, if this logistic regression was applied to a new set of observations from this population, it would be very unlikely to do as well. The classification probability has an upward bias. The bias arises due to the fact that the same data that were used to fit the model, was used to judge the performance of the model. The model fitted to a given body of data is expected to perform well on the same body of data. The true measure of the performance of the logistic regression model for classification is the probability of classifying a future observation correctly and not a sample observation. This upward bias in the estimate of correct classification probability can be reduced by using resampling methods, such as jack-knife or bootstrap. These will not be discussed here. The reader is referred to Efron (1982) and Diaconis and Efron (1983).

12.8 THE MULTINOMIAL LOGIT MODEL

In our discussion of logistic regression we have so far assumed that the qualitative response variable assumes only two values, generically, 1 for success and 0 for failure. The logistic regression model can be extended to situations where the response variable assumes more than two values. In a study of the choice of mode of transportation to work, the response variable may be private automobile, car pool, public transport, bicycle, or walking. The response falls into five categories. There is no natural ordering of the categories. We might want to analyze how the choice is related to factors such as age, sex, income, distance traveled, etc. The resulting model can be analyzed by using slightly modified methods that were used in analyzing the dichotomous outcomes. This method is called the *multinomial (polytomous)* logistic regression.

The response categories are not ordered in the example described above. There are situations where the response categories are ordered. In an opinion survey, the response categories might be, strongly agree, agree, no opinion, disagree, and strongly disagree. The response categories are naturally ordered. In a clinical trial the responses to a treatment could be classified as improved, no change, worse. For these situations a different method called the *Proportional Odds Model* is used. We discuss it Section 12.8.3.

12.8.1 Multinomial Logistic Regression

We have n independent observation with p explanatory variables. The qualitative response variable has k categories. To construct the logits in the multinomial case one of the categories is considered the base level and all the logits are constructed relative to it. Any category can be taken as the base level. We will take category k as the base level in our description of the method. Since there is no ordering, it is apparent that any category may be labeled k. Let π_j denote the multinomial probability of an observation falling in the jth category. We want to find the relationship between this probability and the p explanatory variables, $X_1, X_2, \ldots, X_p$. The multiple logistic regression model then is

$$\log\left(\frac{\pi_j(x_i)}{\pi_k(x_i)}\right) = \beta_{0j} + \beta_{1j}x_{1i} + \beta_{2j}x_{2i} + \ldots + \beta_{pj}x_{pi}; \qquad \begin{array}{l} j = 1, 2, \ldots, (k-1), \\ i = 1, 2, \ldots, n. \end{array}$$

Since all the π's add to unity, this reduces to

$$\log(\pi_j(x_i)) = \frac{\exp(\beta_{0j} + \beta_{1j}x_{1i} + \beta_{2j}x_{2i} + \ldots + \beta_{pj}x_{pi})}{1 + \sum_{j=1}^{k-1} \exp(\beta_{0j} + \beta_{1j}x_{1i} + \beta_{2j}x_{2i} + \ldots + \beta_{pj}x_{pi})},$$

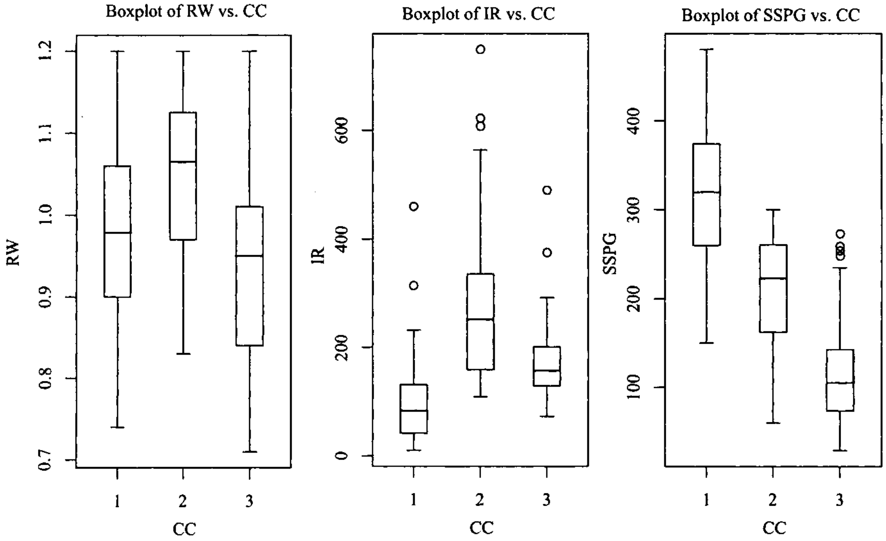

Figure 12.5 Side-by-Side Boxplots for the Diabetes Data.

for $j = 1, 2, \ldots, (k - 1)$. The model parameters are estimated by the method of maximum likelihood. Statistical software is available to do this fitting. We illustrate the method by an example.

12.8.2 Example: Determining Chemical Diabetes

To determine the treatment and management of diabetes it is necessary to determine whether the patient has chemical diabetes or overt diabetes. The data presented in Tables 12.6 and 12.7 is from a study conducted to determine the nature of chemical diabetes. The measurements were taken on 145 nonobese volunteers who were subjected to the same regimen. Many variables were measured, but we consider only three of them. These are, insulin response (IR), the steady state plasma glucose (SSPG), which measures insulin resistance, and relative weight (RW). The diabetic status of each subject was recorded. The clinical classification (CC) categories were overt diabetes (1), chemical diabetes (2), and normal (3). The dataset is found in Andrews and Herzberg (1985). More details of the study are found in Reaven and Miller (1979).

A side-by-side boxplots of the explanatory variables indicate that the distribution of IR and SSPG differ for the three categories. The distribution of RW on the other hand does not differ substantially for the three categories. The boxplots are shown in Figure 12.5. The results of fitting a multinomial logistic model using the variables IR, SSPG, and RW is given in Table 12.8. Each of the logistic models are given relative to normal patients.

Table 12.6 Diabetes Data: Blood Glucose, Insulin Levels, Relative Weight, Clinical Classification (Patients 1 to 90)

Patient	RW	IR	SSPG	CC	Patient	RW	IR	SSPG	CC
1	0.81	124	55	3	46	0.91	106	56	3
2	0.95	117	76	3	47	0.95	118	122	3
3	0.94	143	105	3	48	0.95	112	73	3
4	1.04	199	108	3	49	1.03	157	122	3
5	1.00	240	143	3	50	0.87	292	128	3
6	0.76	157	165	3	51	0.87	200	233	3
7	0.91	221	119	3	52	1.17	220	132	3
8	1.10	186	105	3	53	0.83	144	138	3
9	0.99	142	98	3	54	0.82	109	83	3
10	0.78	131	94	3	55	0.86	151	109	3
11	0.90	221	53	3	56	1.01	158	96	3
12	0.73	178	66	3	57	0.88	73	52	3
13	0.96	136	142	3	58	0.75	81	42	3
14	0.84	200	93	3	59	0.99	151	122	2
15	0.74	208	68	3	60	1.12	122	176	3
16	0.98	202	102	3	61	1.09	117	118	3
17	1.10	152	76	3	62	1.02	208	244	2
18	0.85	185	37	3	63	1.19	201	194	2
19	0.83	116	60	3	64	1.06	131	136	3
20	0.93	123	50	3	65	1.20	162	257	2
21	0.95	136	47	3	66	1.05	148	167	2
22	0.74	134	50	3	67	1.18	130	153	3
23	0.95	184	91	3	68	1.01	137	248	3
24	0.97	192	124	3	69	0.91	375	273	3
25	0.72	279	74	3	70	0.81	146	80	3
26	1.11	228	235	3	71	1.10	344	270	2
27	1.20	145	158	3	72	1.03	192	180	3
28	1.13	172	140	3	73	0.97	115	85	3
29	1.00	179	145	3	74	0.96	195	106	3
30	0.78	222	99	3	75	1.10	267	254	3
31	1.00	134	90	3	76	1.07	281	119	3
32	1.00	143	105	3	77	1.08	213	177	2
33	0.71	169	32	3	78	0.95	156	159	3
34	0.76	263	165	3	79	0.74	221	103	3
35	0.89	174	78	3	80	0.84	199	59	3
36	0.88	134	80	3	81	0.89	76	108	3
37	1.17	182	54	3	82	1.11	490	259	3
38	0.85	241	175	3	83	1.19	143	204	2
39	0.97	128	80	3	84	1.18	73	220	3
40	1.00	222	186	3	85	1.06	237	111	2
41	1.00	165	117	3	86	0.95	748	122	2
42	0.89	282	160	3	87	1.06	320	253	2
43	0.98	94	71	3	88	0.98	188	211	2
44	0.78	121	29	3	89	1.16	607	271	2
45	0.74	73	42	3	90	1.18	297	220	2

Table 12.7 Diabetes Data: Blood Glucose, Insulin Levels, Relative Weight, Clinical Classification (Patients 91 to 145)

Patient	RW	IR	SSPG	CC
91	1.20	232	276	2
92	1.08	480	233	2
93	0.91	622	264	2
94	1.03	287	231	2
95	1.09	266	268	2
96	1.05	124	60	2
97	1.20	297	272	2
98	1.05	326	235	2
99	1.10	564	206	2
100	1.12	408	300	2
101	0.96	325	286	2
102	1.13	433	226	2
103	1.07	180	239	2
104	1.10	392	242	2
105	0.94	109	157	2
106	1.12	313	267	2
107	0.88	132	155	2
108	0.93	285	194	2
109	1.16	139	198	2
110	0.94	212	156	2
111	0.91	155	100	2
112	0.83	120	135	2
113	0.92	28	455	1
114	0.86	23	327	1
115	0.85	232	279	1
116	0.83	54	382	1
117	0.85	81	378	1
118	1.06	87	374	1
119	1.06	76	260	1
120	0.92	42	346	1
121	1.20	102	319	1
122	1.04	138	351	1
123	1.16	160	357	1
124	1.08	131	248	1
125	0.95	145	324	1
126	0.86	45	300	1
127	0.90	118	300	1
128	0.97	159	310	1
129	1.16	73	458	1
130	1.12	103	339	1
131	1.07	460	320	1
132	0.93	42	297	1
133	0.85	13	303	1
134	0.81	130	152	1
135	0.98	44	167	1
136	1.01	314	220	1
137	1.19	219	209	1
138	1.04	100	351	1
139	1.06	10	450	1
140	1.03	83	413	1
141	1.05	41	480	1
142	0.91	77	150	1
143	0.90	29	209	1
144	1.11	124	442	1
145	0.74	15	253	1

We see that RW has insignificant values in each of the logit models. This is consistent with what we observed in the side-by-side boxplots. We now fit the multinomial logistic model with two variables, SSPG and IR. The results are given in Table 12.9.

Looking at Logit (1/3), we see that higher values of SSPG increases the odds of overt diabetes, while a decrease in IR reduces the same odds when compared to normal subjects. Looking at Logit (2/3), we see that the higher values SSPG increases the odds of chemical diabetes when compared to the normal subjects. The IR value does not significantly affect the odds. This indicates the difference between chemical and overt diabetes and has implications for the treatment of the two conditions.

Table 12.8 Multinomial Logistic Regression Output with RW, SSPG, and IR (Base Level =3)

					Odds	95% C.I.	
Variable	Coefficient	s.e.	Z-test	p-value	Ratio	Lower	Upper
Logit 1: (2/3)							
Constant	-7.615	2.336	-3.26	0.001			
RW	3.473	2.446	1.42	0.156	32.23	0.27	3894.21
SSPG	0.016	0.005	3.29	0.001	1.02	1.01	1.03
IR	0.004	0.002	1.53	0.127	1.00	1.00	1.01
Logit 2: (1/3)							
Constant	-1.845	3.463	-0.53	0.594			
RW	-5.868	3.867	-1.52	0.129	0.00	0.00	5.53
SSPG	0.046	0.009	4.92	0.000	1.05	1.03	1.07
IR	-0.0134	0.005	-2.66	0.008	0.99	0.98	1.00
Log-Likelihood = −68.415	G = 159.369		$d.f.$ = 6		p-value < 0.000		

Table 12.9 Multinomial Logistic Regression Output with SSPG and IR (Base Level = 3)

					Odds	95% C.I.	
Variable	Coefficient	s.e.	Z-test	p-value	Ratio	Lower	Upper
Logit 1: (2/3)							
Constant	-4.549	0.771	-5.90	0.000			
SSPG	0.020	0.004	4.38	0.000	1.02	1.01	1.03
IR	0.003	0.002	1.42	0.155	1.00	1.00	1.01
Logit 2: (1/3)							
Constant	-7.111	1.688	-4.21	0.000			
SSPG	0.0426	0.008	5.34	0.000	1.04	1.03	1.06
IR	-0.013	0.005	-2.89	0.004	0.99	0.98	1.00
Log-Likelihood = −72.029	G = 152.141		$d.f.$ = 4		p-value < 0.000		

Table 12.10 Classification Table of Diabetes Data Using Multinomial Logistic Regression

	Predict			
CC	1	2	3	All
1	27	3	3	33
2	1	22	13	36
3	2	5	69	76
All	30	30	85	145

Although we have taken 3 as the base level, from our computation we can derive other comparisons. We can get Logit (1/2) from the relation

$$\text{Logit}(1/2) = \text{Logit}(1/3) - \text{Logit}(2/3). \tag{12.9}$$

We can judge how well the multinomial logistic regression classifies the observations into different categories. The methodology is similar to binary logistic regression. An observation is classified to that category for which it has the highest estimated probability. The classification table for the multinomial logistic regression is given in Table 12.10.

One can see that 118 out of 145 subjects studied are classified correctly by this procedure. 81% of the observations are correctly classified which is considerably higher than the maximum correct rate 59% (85/145), which would have been obtained if all the observations were put in one category. Multinomial logistic regression has performed well on this data. It is a powerful technique that should be used more extensively.

12.8.3 Ordered Response Category: Ordinal Logistic Regression

The response variable in many studies, as has been pointed out earlier, can be qualitative and fall in more than two categories. The categories may sometimes be ordered. In a consumer satisfaction study, the responses might be, highly satisfied, satisfied, dissatisfied, and highly dissatisfied. An analyst may want to study the socioeconomic and demographic factors that influence the response. The logistic model, slightly modified can be used for this analysis. The logits here are based on the cumulative probabilities. Several logistic models can be based on the cumulative logits. We describe one of these, the *proportional odds model.*

Again, we have n independent observations with p predictors. The response variable falls into k categories $(1, 2, \ldots, k)$. The k categories are ordered. Let Y denote the response variable. The cumulative distribution for Y is

$$F_j(x_i) = \Pr(Y \leq j | X_i = x_{i1}, \ldots, X_p = x_{ip},); \quad j = 1, 2, \ldots, (k-1).$$

Table 12.11 Ordinal Logistic Regression Model (Proportional Odds) Using SSPG and IR

					Odds	95% C.I.	
Variable	Coefficient	s.e.	Z-test	p-value	Ratio	Lower	Upper
Constant 1	-6.794	0.872	-7.79	0.000			
Constant 2	-4.189	0.665	-6.30	0.000			
IR	-0.004	0.002	-2.30	0.021	1.00	0.99	1.00
SSPG	0.028	0.004	7.73	0.000	1.03	1.02	1.04
Log-Likelihood = −81.749	G = 132.700		$d.f.$ = 2	p-value < 0.000			

The proportional odds model is given by,

$$L_j(x_i) = \log\left(\frac{F_j(x_i)}{1 - F_j(x_i)}\right) == \beta_{0j} + \beta_{1j}x_{1i} + \beta_{2j}x_{2i} + \ldots + \beta_{pj}x_{pi},$$

for $j = 1, 2, \ldots, (k-1)$. The comulative logit has a simple interpretation. It can be interpreted as the logit for a binary response in which the categories from 1 to j is one category, and the remaining categories from $(j+1)$ to k is the second category. The model is fitted by the maximum likelihood method. Several statistical software packages will carry out this procedure. Increase in the value of a response variable with a positive β will increase the probability of being in a lower numbered category, all other variables remaining the same. The number of parameters estimated to describe the data is fewer in the ordinal than in the nominal model. For a more detailed discussion the reader is referred to Agresti (2002) and Simonoff (2003).

12.8.4 Example: Determining Chemical Diabetes Revisited

We will use the data on chemical diabetes considered in Section 12.8.2 to illustrate ordinal logistic regression. The clinical classifications in the previous categories are ordered but we did not take it into consideration in our analysis. The progression of diabetes goes from normal (3), chemical (2), to overt diabetes (1). The classification states have a natural order and we will use them in our analysis. We will fit the proportional odds logit model. The result of the fit is given in Table 12.11.

The fit for the model is good. Both variables have significant relationship to the group membership. The coefficient of SSPG is positive. This indicates that higher values of SSPG increase the probability of being in a lower numbered category other factors being the same. The coefficient of IR is negative, indicating that higher values of this variable increase the probability of being in a higher numbered category, other factors remaining the same. The coefficient of concordance is high (0.90) showing the ability of the model to classify the group membership is high. In Table 12.12, we give the classification table for the ordinal logistic regression.

Table 12.12 Classification Table of Diabetes Data Using Multinomial Logistic Regression

	Predict			
CC	1	2	3	All
1	26	5	2	33
2	3	20	13	36
3	0	8	68	76
All	29	33	83	145

Of the 145 subjects ordinal logit regression classifies 114 subjects to their correct group. This gives the correct classification rate as 79%. This is comparable to the rate achieved by the multinomial logit model. It is generally expected that the ordinal model will do better than the multinomial model because of the additional information provided by the ordering of the categories. It should be also noted the ordinal logit model uses fewer parameters than the multinomial model. In our example the ordinal model uses 4 parameters, while the nominal version uses 6. For a more detailed discussion the reader is referred to Agresti (2002) and Simonoff (2003).

12.9 CLASSIFICATION PROBLEM: ANOTHER APPROACH

The method of logistic regression has been used to model the probability that an observation belongs to one group given the measurements on several characteristics. We have described how the fitted logits could then be used for classifying an observation into one of two categories. A different statistical methodology is available if our primary interest is *classification*. When the sole interest is to predict the group membership of each observation a statistical method called *discriminant analysis* is commonly used. Without discussing discriminant analysis here, we indicate a simple regression method that will accomplish the same task. The reader can find a discussion of discriminant analysis in McLachlan (1992), Rencher (1995), and Johnson (1998).

The essential idea in discriminant analysis is to find a linear combination of the predictor variables $X_1, \ldots, X_p$, such that the scores given by this linear combination separates the observations from the two groups as far as possible. One way that this separation can be accomplished is by fitting a multiple regression model to the data. The response variable is Y, taking values 0 and 1, and the predictors are $X_1, \ldots, X_p$. As has been pointed out earlier, some of the fitted values will be outside the range of 0 and 1. This does not matter here, as we are not trying to model probabilities, but only to predict group membership. We calculate the average of the predicted values of all the observations. If the predicted value for a given observation is

Table 12.13 Results from the OLS Regression of Y on X_1, X_2, X_3

Variable	Coefficient	s.e.	t-test	p-value
Constant	0.322	0.087	3.68	0.0005
X_1	0.003	0.001	3.76	0.0004
x_2	0.004	0.001	2.96	0.0044
x_3	0.149	0.045	3.28	0.0017
$n = 66$	$R^2 = 0.57$	$R_a^2 = 0.55$	$\hat{\sigma} = 0.3383$	$d.f. = 62$

greater than the average predicted value we assign that observation to the group which has $Y = 1$; if the predicted value is smaller than the average predicted value we assign it to the group with $Y = 0$. From this assignment we determine the number of observations classified correctly in the sample. The variables used in this classification procedure are determined exactly by the same methods as those used for variable selection in multiple regression.

We illustrate this method by applying it to the Bankruptcy data that we have used earlier to illustrate least squares regression. Table 12.13 gives the OLS regression results using the three predictor variables X_1, X_2, and X_3. All three variables have significant regression coefficients and should be retained for classification equation.

Table 12.14 displays the observed Y, the predicted Y, and the assigned group for the Bankruptcy data. The average value of the predicted Y is 0.5. All observations with predicted value less than 0.5 is assigned to $Y = 0$, and those with predicted value greater than 0.5 is assigned to the group with $Y = 1$. The wrongly classified observations are marked by *. It is seen that 5 bankrupt firms are classified as solvent, and one solvent firm is classified as bankrupt. The logistic regression, it should be noted, classified only two observations wrongly. One solvent firm and one bankrupt firm were misclassified. For the Bankruptcy data presented in Table 12.2, the logistic regression performs better than the multiple regression in classifying the sample data. In general this is true. The logistic regression does not have to make the restrictive assumption of multivariate normality for the predictor variables. For classification problems we recommend the use of logistic regression. If a logistic regression package is not available, then the multiple regression approach may be tried.

EXERCISES

12.1 The diagnostic plots in Figures 12.2, 12.3, and 12.4 show three unusual observations in the Bankruptcy data. Fit a logistic regression model to the 63 observations without these three observations and compare your results with the results obtained in Section 12.5. Does the deletion of the three points cause a substantial change in the logistic regression results?

Table 12.14 Classification of Observations by Fitted Values

Row	Y	Fitted	Assigned	Row	Y	Fitted	Assigned
1	0	−0.00	0	34	1	0.72	1
2	0	0.48	0	35	1	0.82	1
3	0	−0.12	0	36	1	0.73	1
4	0	0.31	0	37	1	0.80	1
5	0	0.23	0	38	1	0.65	1
6	0	0.14	0	39	1	0.80	1
7	0	0.33	0	40	1	0.75	1
8	0	−0.32	0	41	1	0.76	1
9	0	0.52	1*	42	1	0.83	1
10	0	0.12	0	43	1	1.10	1
11	0	0.23	0	44	1	1.42	1
12	0	−0.07	0	45	1	0.86	1
13	0	−0.80	0	46	1	0.66	1
14	0	0.55	1*	47	1	0.81	1
15	0	0.03	0	48	1	0.58	1
16	0	−0.45	0	49	1	0.97	1
17	0	0.64	1*	50	1	1.03	1
18	0	0.45	0	51	1	0.77	1
19	0	0.44	0	52	1	0.48	0*
20	0	0.14	0	53	1	0.60	1
21	0	0.22	0	54	1	0.74	1
22	0	0.37	0	55	1	0.81	1
23	0	0.18	0	56	1	0.84	1
24	0	0.05	0	57	1	0.62	1
25	0	0.55	1*	58	1	0.81	1
26	0	0.56	1*	59	1	0.74	1
27	0	0.39	0	60	1	0.84	1
28	0	0.34	0	61	1	0.86	1
29	0	0.39	0	62	1	0.80	1
30	0	0.26	0	63	1	0.68	1
31	0	0.39	0	64	1	0.83	1
32	0	0.12	0	65	1	0.61	1
33	0	0.44	0	66	1	0.59	1

* Wrongly classified observations

12.2 Examine the various logistic regression diagnostics obtained from fitting the logistic regression Y on X_1 and X_2 (Table 12.3) and determine if the data contain unusual observations.

Table 12.15 Number of O-rings Damaged and the Temperature (Degrees Fahrenheit) at the Time of Launch for 23 Flights of the Space Shuttle *Challenger*

Flight	Damaged	Temperature
1	2	53
2	1	57
3	1	58
4	1	63
5	0	66
6	0	67
7	0	67
8	0	67
9	0	68
10	0	69
11	0	70
12	0	70
13	1	70
14	1	70
15	0	72
16	0	73
17	0	75
18	2	75
19	0	76
20	0	78
21	0	79
22	0	81
23	0	76

12.3 The *O-rings* in the booster rockets used in space launching play an important part in preventing rockets from exploding. Probabilities of O-ring failures are thought to be related to temperature. A detailed discussion of the background of the problem is found in The Flight of the Space Shuttle Challenger (pp. 33–35) in Chatterjee, Handcock, and Simonoff (1995). Each flight has six O-rings that could be potentially damaged in a particular flight. The data from 23 flights are given in Table 12.15 and can also be found in the the book's Web site.[3] For each flight we have the number of O-rings damaged and the temperature of the launch.

(a) Fit a logistic regression connecting the probability of an O-ring failure with temperature. Interpret the coefficients.

(b) The data for Flight 18 that was launched when the launch temperature was 75 was thought to be problematic, and was deleted. Fit a logistic regression to the reduced data set. Interpret the coefficients.

(c) From the fitted model, find the probability of an O-ring failure when the temperature at launch was 31 degrees. This was the temperature forecast for the day of the launching of the fatal *Challenger* flight on January 20, 1986.

(d) Would you have advised the launching on that particular day?

12.4 Field-goal-kicking data for the entire American Football League (AFL) and National Football League (NFL) for the 1969 season are given in Table 12.16 and can also be found in the the book's Web site. Let $\pi(X)$ denote the probability of kicking a field goal from a distance of X yards.

[3] http://www.ilr.cornell.edu/~hadi/RABE4

Table 12.16 Field-Goal-Kicking Performances of the American Football League (AFL) and National Football League (NFL) for the 1969 Season. The Variable Z Is an Indicator Variable Representing League

League	Distance	Success	Attempts	Z
NFL	14.5	68	77	0
NFL	24.5	74	95	0
NFL	34.5	61	113	0
NFL	44.5	38	138	0
NFL	52.0	2	38	0
AFL	14.5	62	67	1
AFL	24.5	49	70	1
AFL	34.5	43	79	1
AFL	44.5	25	82	1
AFL	52.0	7	24	1

Source: Morris and Rolph (1981), p. 200.

(a) For each of the leagues, fit the model

$$\pi(X) = \frac{e^{\beta_0+\beta_1\, X+\beta_2\, X^2}}{1+e^{\beta_0+\beta_1\, X+\beta_2\, X^2}}\ .$$

(b) Let Z be an indicator variable representing the league, that is,

$$Z = \begin{cases} 1, & \text{for the AFL,} \\ 0, & \text{for the NFL.} \end{cases}$$

Fit a single model combining the data from both leagues by extending the model to include the indicator variable Z; that is, fit

$$\pi(X, Z) = \frac{e^{\beta_0+\beta_1\, X+\beta_2\, X^2+\beta_3\, Z}}{1+e^{\beta_0+\beta_1\, X+\beta_2 X^2+\beta_3\, Z}}\ .$$

(c) Does the quadratic term contribute significantly to the model?

(d) Are the probabilities of scoring field goals from a given distance the same for each league?

12.5 Using the data on diabetes analyzed in Tables 12.6 and 12.7, show that inclusion of the variable RW does not result in a substantial improvement in the classification rate from the multinomial logistic model using IR and SSPG.

12.6 Using the diabetes data in Tables 12.6 and 12.7, fit an ordinal logistic model using RW, IR, and SSPG to explain CC. Show that there is no substantial improvement in fit, and the correct classification rate from a model using only IR and SSPG.