14 Logistic and Proportional Hazards Regression

Chapter Outline
14.1 Simple Logistic Regression
14.2 Multiple Logistic Regression
14.3 Ordered Logistic Regression
14.4 Conditional Logistic Regression
14.5 Introduction to Proportional Hazard Regression

In this chapter we present logistic regression, a method for examining the relationship between a dependent variable with two levels and one or more independent variables. Logistic regression represents another application of the linear model idea used in the two previous chapters. We also provide an introduction to proportional hazards regression (or Cox's regression). Proportional hazards regression is an extension of the survival analysis method presented in Chapter 11, and it also uses the linear model approach.

14.1 Simple Logistic Regression

Joseph Berkson did much to advance the use of logistics in the 1940s and 1950s (Berkson 1944; 1951). However, it was D. R. Cox (1969) who popularized the logit transformation for modeling binary data. Since the 1980s, logistic regression has become one of the more widely used analysis techniques in public health and the biomedical sciences because it allows for an examination of the relation between disease status (presence or absence) and a set of possible risk factors for the disease based on data from cross-sectional, case-control, or cohort studies.

Let's consider a simple example to introduce the topic because it allows us to show the logistic regression in terms of statistics that we already know.

Example 14.1

Suppose that we wish to determine whether or not there is a relationship between a male's pulmonary function test (PFT) results and air pollution level at his residence — lead in the air serving as a proxy for overall air pollution. The data for this situation are shown in Table 14.1 (Forthofer and Lehnen 1981).

Table 14.1 Pulmonary function test results by ambient air pollution.

	Pollution (Lead) Level	
Pulmonary Function Test Results	**Low**	**High**
Normal	368	82
Abnormal	19	10
Total	387	92
Proportions Normal	0.9509	0.8913
Odds (normal)	19.367	8.200
Logits (normal)	2.964	2.104

We cannot use ordinary linear regression for this situation because the dependent variable — PFT results categorized as normal or nonnormal — has only two levels, and, hence, the assumption of a continuous and normally distributed dependent variable does not hold. We can use categorical data analysis, since the independent variable, lead level categorized as low or high, is discrete. More generally, if there were several independent variables, some of which were continuous, then the categorical data approach would no longer be appropriate.

One categorical data approach is to compare the odds of having a normal PFT between those exposed to low and those exposed to high levels of air pollution — that is, to calculate the odds ratio and then test the hypothesis that the odds ratio is equal to one. In the following, we shall consider the relation between logistic regression and the odds ratio.

In logistic regression the underlying model is that the natural logarithm, written as ln, of the odds of a normal (or nonnormal) PFT is a linear function of a constant and the effect of lead pollution. The logarithm of the odds is also referred to as the *log odds* or *logit*. In this example, a larger logit value indicates a more favorable outcome because it indicates a greater proportion of males having a normal PFT. Hence those with low exposures to lead (logit = 2.964) have a more favorable outcome than those with higher exposure to lead (logit = 2.104) for this sample.

This model is

$$\ln\left(\frac{\pi_{i1}}{\pi_{i2}}\right) = constant + i^{th}\ lead\ \ pollution\ effect$$

where π_{i1} is the probability of a normal PFT and π_{i2} is the probability of a nonnormal PFT for the ith lead level. The ratio of π_{i1} to π_{i2} is the odds of a normal PFT for the ith level of lead.

Substituting symbols for all the terms in the preceding equation yields

$$\ln\left(\frac{\pi_{i1}}{\pi_{i2}}\right) = \mu + \alpha_i \tag{14.1}$$

where μ represents the constant and α_i is the effect of the ith level of lead. This model has the same structure that we used in the ANOVA where we are measuring the effect of the levels of a variable from a reference level. For the lead variable, we consider the high level of pollution to be the reference level. This means that α_2 is

taken to be 0 and that μ is the logit for the high lead level as can be seen from the following two equations:

$$\ln\left(\frac{\pi_{11}}{\pi_{12}}\right) = \mu + \alpha_i$$

$$\ln\left(\frac{\pi_{21}}{\pi_{22}}\right) = \mu.$$

It is clear from the second of these two equations that μ is the logit of a normal PFT for those exposed to the high lead pollution level. If we subtract the second equation from the first, we see that α_1 is simply the difference of the two logits — that is,

$$\ln\left(\frac{\pi_{11}}{\pi_{12}}\right) - \ln\left(\frac{\pi_{21}}{\pi_{22}}\right) = \alpha_1.$$

Since the difference of two logarithms is the logarithm of the ratio, we have

$$\alpha_1 = \ln\left(\frac{\pi_{11} \cdot \pi_{22}}{\pi_{12} \cdot \pi_{21}}\right).$$

Thus, α_1 is the natural logarithm of the odds ratio, and this is one of the reasons that logistic regression is so useful. In this example, the estimate of α_1 is 0.860 and the estimate of μ is 2.104. If we take the exponential of the estimate of α_1, we obtain the value 2.362, the estimated odds ratio. This value is much greater than one, and it strongly supports the idea that those with the lower lead exposure have the greater proportion of a normal PFT. The estimate of the constant term is the logit for the high level of lead, and the exponential of the estimate of μ is 8.2, the odds of a normal PFT result for those with high lead exposures. Thus the logistic regression model leads to parameters that are readily interpretable.

14.1.1 Proportion, Odds, and Logit

Before proceeding with the extension of the logistic regression model to multiple independent variables, it is helpful to examine the relationship between probabilities (π_i), odds [$o_i = \pi_i/(1 - \pi_i)$] and logits [$\lambda_i = \ln(o_i)$] shown in Table 14.2.

Note that when the probability is 0.5, the odds equal 1 or are even. As the probabilities increase toward 1, the odds increase quite rapidly. As the probabilities decrease toward 0, the odds also approach 0. When the odds equal 1, the logit is 0. As the odds decrease below 1, the logit takes a negative value, approaching negative infinity. As the odds increase above 1, the logit takes a positive value, approaching positive infinity.

The relationship between probabilities and logits is graphically shown in Figure 14.1. The relationship is essentially linear for probabilities between 0.3 and 0.7 and nonlinear

Table 14.2 A comparison of probabilities, odds, and log odds (logits).

Probabilities (π_i)	0.01	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	0.99
Odds (o_i)	0.01	0.11	0.25	0.43	0.67	1.00	1.50	2.33	4.00	9.00	99.00
Logits (λ_i)	−4.59	−2.20	−1.39	−0.85	−0.41	0.00	0.41	0.85	1.39	2.20	4.59

Figure 14.1 Plot of probabilities versus logits.

for lower and greater probabilities. A unit change in the logit results in greater differences in probabilities at levels in the middle than at high and low levels.

Manipulating the formula for the odds allows us to express probabilities in terms of odds

$$\pi_i = \frac{o_i}{1 + o_i}. \tag{14.2}$$

Since $o_i = \exp(\lambda_i)$, we can also express probabilities in terms of logits

$$\pi_i = \frac{\exp(\lambda_i)}{1 + \exp(\lambda_i)}. \tag{14.3}$$

This expression for the probability is one that is often seen in the literature when dealing with logistic regression.

14.1.2 Estimation of Parameters

Example 14.1 showed that the estimation of parameters for the case where both the outcome variable and the exposure variable have two levels is quite simple. However, the estimation of parameters in logistic regression becomes more complex when we incorporate continuous independent variables and discrete variables with multiple levels in the model.

It turns out that the least squares estimation procedure doesn't yield the best estimates for the parameters in logistic regression. Instead of least squares, logistic regression uses the *maximum likelihood* procedure to obtain the parameter estimates. The maximum likelihood approach finds estimates of the model parameters that have the greatest likelihood of producing the observed data. The estimation procedure usually begins with the least squares estimates of coefficients and then uses an iterative algorithm to successively find new sets of coefficients that have higher likelihood of producing the observed data. Computer programs typically show the number of iterations required to find the estimated coefficients with the greatest likelihood. However, it is beyond the scope of

this book to provide the details of the estimation. For more information on logistic regression, see the excellent book by Hosmer and Lemeshow (1999a).

14.1.3 Computer Output

The following two examples are applications of logistic regression with a single independent variable. In the first example, the independent variable has only two levels, whereas in the second example, the independent variable is continuous.

Example 14.2

Table 14.3 presents a summary of the computer output for a logistic regression analysis of the data used in Example 14.1 (see **Program Note 14.1** on the website).

The estimates for the intercept and the effect of the low lead level are 2.104 and 0.860, respectively. These estimates are the same as in Example 14.1. Table 14.3 also shows the standard errors for the coefficients, test statistics, *p*-value, and confidence interval for the odds ratio. These will be explained in the next section.

Table 14.3 Estimates resulting from the fitted logit model for the PFT data in Table 14.1.

Variable	Coefficient	Standard Error	Wald Statistic	*p*-value	Odds Ratio	95% Confidence Interval
Intercept	2.104	0.335	6.28	<0.001	—	—
Lead (low)	0.860	0.409	2.10	0.036	2.362	(1.06, 5.27)

Likelihood ratio: chi-square = 4.025, df = 1, *p*-value = 0.045

In the following example, we consider the case with a continuous covariate in the model.

Example 14.3

We want to explore the relationship between diabetes (presence or absence) and body mass index (BMI) using individuals from the DIG200 dataset. In the DIG200 dataset, BMI is a continuous variable and will serve as the independent variable. The symbolic representation of this model is

$$\ln\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 x_1$$

where x_1 represents the value of the BMI. For simplicity we rounded the values of BMI to the nearest whole number. The results of fitting the logistic regression model are shown in Table 14.4. See **Program Note 14.1** on the website for fitting this model.

The fitted logit model is

$$\ln\left(\frac{\hat{\pi}}{1-\hat{\pi}}\right) = -3.034 + 0.075 x_1.$$

Table 14.4 Estimates resulting from the logistic regression analysis of diabetes on body mass index, DIG200.

Variable	Coefficient	Standard Error	Wald Statistic	*p*-value	Odds Ratio	95% Confidence Interval
Intercept	–3.034	0.893	–3.40	0.001	—	—
BMI	0.075	0.032	2.35	0.019	1.08	(1.01, 1.15)

Log-Likelihood: Intercept only –116.652
BMI term added –113.851
Likelihood ratio: chi-square = 5.602, df = 1, *p*-value = 0.018

This estimated equation means that for a 1 kg/m^2 increase in BMI, the log odds of having diabetes increases by 0.075 units. However, a 5 kg/m^2 increase in BMI may be more meaningful than a change of 1 kg/m^2. A 5 kg/m^2 increase in BMI increases the log odds by 0.375 (= 5 ∗ 0.075) units. The estimated change in the odds is easily calculated by exp(0.375) = 1.45. This value means that the estimated odds of diabetes increases by 45 percent for every 5 kg/m^2 increase in BMI.

14.1.4 Statistical Inference

In this section we are interested in examining if a significant relationship exists between the dependent variable and independent variable(s) contained in the logistic model. The two tests commonly used in the tests of hypotheses in logistic regression are the *Wald test* and the *likelihood ratio test* (LRT). We are interested in testing the null hypothesis that the coefficient of the independent variable is equal to zero versus the alternative hypothesis that the coefficient is nonzero — that is,

$$H_0\text{: } \beta_1 = 0 \text{ versus } Ha\text{: } \beta_1 \neq 0.$$

We begin with the Wald test.

The test statistic for the Wald test is obtained by dividing the maximum likelihood estimate (MLE) of the slope parameter $\hat{\beta}_1$ by the estimate of its standard error, *se* ($\hat{\beta}_1$). Under the null hypothesis, this ratio follows a standard normal distribution.

Example 14.4

Let us reexamine the material from Example 14.2. As shown in Table 14.3, the value of $\hat{\beta}_1$ is 0.860 and *se* ($\hat{\beta}_1$) is 0.409. Therefore, the Wald test statistic is calculated as follows:

$$\frac{\hat{\beta}_1}{se\left(\hat{\beta}_1\right)} = \frac{0.860}{0.409} = 2.10.$$

If the null hypothesis is true, this statistic follows the standard normal distribution. The *p*-value for this test is 0.036 [= 2 ∗ Prob(Z > 2.10)], suggesting that β_1 is significantly different from zero at the 0.05 level. These values are shown in Table 14.3.

We can use the confidence interval for the odds ratio to determine whether or not the odds ratio equals one. If the confidence interval does not contain one, then we

conclude that the odds ratio is statistically significant. The use of the confidence interval is equivalent to testing the hypothesis that $\beta_1 = 0$. The $100 * (1 - \alpha)$ percent confidence interval for the odds ratio $[\exp(\beta_1)]$ is calculated by

$$[\exp\{\hat{\beta}_1 - z_{1-\alpha/2} \cdot se(\hat{\beta}_1)\}, \exp\{\hat{\beta}, + z_{1-\alpha/2} \cdot se(\hat{\beta}_1)\}].$$

Using the estimates in Table 14.3, the 95 percent confidence interval for the odds ratio is

$$[\exp\{0.860 - 1.96 * (0.409)\}, \exp\{0.860 + 1.96 * (0.409)\}]$$

or from 1.059 to 5.269. Since the interval does not contain one, the odds ratio is considered to be statistically significant at the 0.05 level. Note that the confidence interval for the odds ratio is not symmetric around the sample estimate. We also did not use the usual approach and base the confidence interval on the estimated odds ratio itself and its estimated standard error because the estimated odds ratio does not follow a normal distribution.

The LRT is used to test the hypothesis that an independent variable is zero. The LRT test statistic is

$$\chi^2_{LR} = -2\ln\left(\frac{\text{Likelihood of the reduced model}}{\text{Likelihood of the full model}}\right) \tag{14.4}$$

a quantity that follows the chi-square distribution under the null hypothesis. The degrees of freedom for the chi-square distribution is the difference between the number of parameters in the full model and the number of parameters in the reduced model. In the simple case of only one covariate in the model, the null hypothesis is that the covariate's coefficient is equal to zero. Although the Wald test's *p*-values are commonly reported, we recommend the use of the *p*-values from the likelihood ratio test (Hauck and Donner 1977; Jennings 1986). The following example demonstrates the use of the LRT.

Example 14.5

Let us revisit the results of the logistic model for the BMI data shown Table 14.4 in Example 14.3. We wish to determine whether or not there is a significant relationship between the independent variable BMI and the presence or absence of diabetes. We shall test the hypothesis of no relationship at the 0.05 level. In symbols, the hypothesis is

$$H_0 : \beta_1 = 0 \text{ versus } H_a : \beta_1 \neq 0.$$

We begin with the model containing only the constant term and compare it to a model containing both the constant and the BMI variable. The log of the likelihood for the constant only model is –116.652, and the log of the likelihood for the model with the BMI variable is –113.851. The test statistic is found by applying Equation (14.4) — that is,

$$X^2_{LR} = -2[\ln(\text{likelihood of reduced model}) - \ln(\text{likelihood of full model})]$$

$= -2 * [-116.652 - (-113.851)] = -2 * (-2.801)$, which is 5.602. There is one degree of freedom for this test of hypothesis because the full model contains only one covariate and the reduced model does not contain any covariates. In this case, the p-value for a chi-square value of 5.602 with one degree of freedom is 0.018. Therefore, we reject the null hypothesis and conclude that β_1 is significantly different from zero — that is, the occurrence of diabetes is related to the BMI variable at the 0.05 level.

14.2 Multiple Logistic Regression

Regression models are useful because they help us explore the relationships between a dependent or response variable and one or more independent or predictor variables of interest. In particular, logistic regression models allow medical researchers to help clinicians in the choice of an appropriate treatment strategy for individual patients.

14.2.1 Model and Assumptions

In the previous section we introduced the simple logistic regression model with only one independent variable. For multiple logistic regression with k independent variables, $x_1, x_2, \ldots, x_k$, the model, taking the form of Equation (14.3), is

$$\pi = \frac{\exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k)}{1 + \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k)}. \tag{14.5}$$

By obtaining estimates for the betas in the linear combination, $\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$, we can calculate the estimated or predicted probability of the outcome of interest.

We present two examples here. The first example includes discrete independent variables only, whereas the second example has both discrete and continuous independent variables.

Example 14.6

We reconsider Example 14.1 and now introduce a covariate. In the example, we found a significant lead effect, a finding that is somewhat surprising, since lead has not been shown to have a negative impact on the respiratory system in other studies. However, during the period 1974–1975 when this study was performed, automobile emissions were a major source of lead pollution. Thus, a possible explanation for this finding is that lead pollution is serving as a proxy for nitrogen dioxide or other pollutants that have adverse respiratory effects. Another possible explanation is that we have not controlled for possible confounding variables. Smoking status is a key variable that has been ignored in the analysis so far. Table 14.5 shows the smoking status by lead level and PFT result.

We begin by considering a model containing the main effects of lead and smoking. Because smoking status contains four levels, we must create three dummy variables in order to obtain a symbolic representation of this model. The dummy variables can be expressed as shown in Table 14.6.

Table 14.5 Pulmonary function test (PFT) results by smoking status and lead exposure.

Lead Level	Smoking Status	PFT Results			Proportions	Odds	Logits
		Normal	Abnormal	Total	Normal	(normal)	(normal)
Low	Heavy	84	3	87	0.9655	28.000	3.332
	Light	75	6	81	0.9260	12.500	2.526
	Former	49	6	55	0.8910	8.167	2.100
	Never	160	4	164	0.9756	40.000	3.689
High	Heavy	16	3	19	0.8421	5.333	1.674
	Light	21	2	23	0.9130	10.500	2.351
	Former	12	2	14	0.8571	6.000	1.792
	Never	33	3	36	0.9167	11.000	2.398

Table 14.6 Dummy variables for the smoking status variable.

	Smoking Status	D_1	D_2	D_3
Smoking status 0	Heavy	0	0	0
Smoking status 1	Light	1	0	0
Smoking status 2	Former	0	1	0
Smoking status 3	Never	0	0	1

We will use the heavy smoking status as the reference category and measure the effects of the other smoking categories from it. Thus, the dummy variable D_1 is 1 when the smoking status is light and 0 otherwise, D_2 is 1 when the smoking status is former and 0 otherwise, and D_3 1 when the smoking status is never and 0 otherwise. Statistical software packages can create these dummy variables for the user (see **Program Note 14.2** on the website for more details).

Therefore, the estimated logit can be expressed as

$$\ln\left(\frac{\hat{\pi}_i}{1-\hat{\pi}_i}\right) = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 D_{1i} + \hat{\beta}_3 D_{2i} + \hat{\beta}_4 D_{3i}.$$

The estimated values of the logit model's parameters are the following:

$$\ln\left(\frac{\hat{\pi}_i}{1-\hat{\pi}_i}\right) = 2.18 + 0.84 x_{1i} - 0.29 D_{1i} - 0.77 D_{2i} + 0.50 D_{3i}.$$

The addition of the smoking variable has not changed the parameter estimates much. The estimate of the constant was previously 2.104 (versus 2.18 now), and the previous estimate of the low lead effect was 0.860 (versus 0.84 now).

In this situation, the estimate of β_1 is the natural logarithm of the odds ratio if the high and low lead levels had contained the same distributions of the smoking status variable. Examination of Table 14.5 shows that the distributions of the smoking status variable are similar for the high and low lead levels. Hence it is not surprising that the estimates of the odds ratio for high lead levels compared to low lead levels are approximately the same for the simple model shown in Table 14.3 and the model shown in Table 14.7. Individuals in residences with low lead levels are 2.3 times more likely to have normal PFT results compared to individuals in residences with high lead levels after adjusting for smoking status.

Table 14.7 Estimates for the logit model parameters and odds ratio for the data in Table 14.5.

Variable	Coefficient	Standard Error	Wald Statistic	*p*-value	Odds Ratio	95% Confidence Interval
Intercept	2.178	0.510	4.27	<0.0001	—	—
Lead (low)	0.837	0.414	2.03	0.043	2.31	(1.03, 5.20)
Smoke 1[a]	−0.289	0.562	−0.51	0.607	0.75	(0.25, 2.25)
Smoke 2[b]	−0.767	0.567	−1.35	0.176	0.46	(0.15, 1.41)
Smoke 3[c]	0.508	0.572	0.89	0.375	1.66	(0.54, 5.10)

Likelihood ratio: chi-square = 9.914, df = 4, *p*-value = 0.042
Goodness of fit tests: Pearson chi-square = 2.276, df = 3, *p*-value = 0.517
Deviance chi-square = 2.256, df = 3, *p*-value = 0.521

[a]light smoker; [b]former smoker; [c]never smoked

There is no suggestion that the effect of any of the three levels of the smoking variables differ from the effect of the heavy smoking level. The 95 percent confidence intervals for the odds ratios of the smoking effects all contain one, and none of the Wald statistics suggest statistical significance at the 0.05 level. The likelihood ratio test statistic shown in the output is used to test the hypothesis that all four model coefficients (the lead effect and the three smoking effects) are simultaneously equal to zero. We reject this hypothesis at the 0.05 level. The lead variable still appears to be related to the PFT variable. We will explain the other two test statistics later in this chapter.

Example 14.7

Suppose that we would like to develop a logistic regression model to predict diabetes using BMI, treatment, and race using the DIG200 dataset. The literature suggests that individuals with larger values of BMI, who are on a placebo, and who are nonwhite are more likely to have diabetes. As we did in Example 14.3, we rounded the values of BMI to the nearest whole number. Table 14.8 shows information about the three predictor variables and the presence or absence of diabetes.

We will consider the placebo level of the treatment variable to be the reference level and measure the effect of the digoxin treatment from it. We will also consider

Table 14.8 Patient characteristics by diabetes status, DIG200.

Characteristics		Diabetes: Yes	Diabetes: No
Mean BMI[a] ± SD[b]		28.0 ± 5.5	26.1 ± 4.6
Range		(18 – 43)	(15 – 45)
Treatment:	Placebo	34	66
	Digoxin	20	80
Race:	White	42	131
	Nonwhite	12	15

[a]BMI—Body Mass Index rounded to the nearest whole number
[b]SD—Standard Deviation

Table 14.9 Logistic regression analysis of diabetes on BMI, treatment, and race, DIG200.

Variable	Coefficient	Standard Error	Wald Statistic	*p*-value	Odds Ratio	95% Confidence Interval
Intercept	−2.948	0.914	−3.22	0.001		
BMI (kg/m^2)	0.081	0.033	2.45	0.014	1.08	(1.02, 1.16)
Treatment (digoxin)	−0.796	0.339	−2.35	0.019	0.45	(0.23, 0.88)
Race (nonwhite)	0.904	0.440	2.05	0.040	2.47	(1.04, 5.85)

Likelihood ratio: chi-square = 15.471, df = 3, *p*-value = 0.001

Goodness of fit tests: Pearson chi-square = 44.485, df = 57, *p*-value = 0.886
Deviance chi-square = 57.816, df = 57, *p*-value = 0.445
H-L chi-square = 2.532, df = 8, *p*-value = 0.960

the white race to be the reference level and measure the nonwhite effect from it. The results of the logistic regression analysis are shown in Table 14.9 (see **Program Note 14.3** on the website).

The fitted logit model is

$$\ln\left(\frac{\hat{\pi}}{1-\hat{\pi}}\right) = -2.948 + 0.081x_1 - 0.796x_2 + 0.904x_3.$$

From Table 14.9, we see that the odds of having diabetes is higher for larger values of BMI even after adjusting for treatment and race. The estimated adjusted odds ratios are greater than one for the BMI and race variables, whereas the adjusted odds ratio is below one for the treatment variable. This indicates that patients receiving digoxin are less likely (specifically 45 percent less likely) to have diabetes compared to patients on the placebo after adjusting for BMI and race.

All three of the independent variables are statistically significant at the 0.05 level. The likelihood ratio chi-square statistic (= 15.47 with three degrees of freedom) suggests that the three coefficients associated with the independent variables are not simultaneously equal to zero at the 0.05 level.

The probability of diabetes given an individual's BMI, treatment group, and race can also be estimated based on the estimated model parameters using Equation (14.5)

$$\hat{\pi}_i = \frac{\exp(-2.948 + 0.081x_{1i} - 0.796x_{2i} + 0.904x_{3i})}{1 + exp(-2.948 + 0.081x_{1i} - 0.796x_{2i} + 0.904x_{3i})}.$$

As an example let us consider calculating the probability of having diabetes given a patient with a BMI of 24 kg/m^2, on digoxin treatment, and being of a nonwhite race. The calculation is

$$\hat{\pi} = \frac{\exp[-2.948 + 0.081(24) - 0.796(1) + 0.0904(1)]}{1 + \exp[-2.948 + 0.081(24) - 0.796(1) + 0.904(1)]} = 0.290.$$

Therefore, the odds of diabetes given a patient with a BMI of 24 kg/m^2, on digoxin treatment, and being nonwhite is [0.290/(1 − 0.290)] = 0.408. We explain the other two test statistics in the following sections.

14.2.2 Residuals

In logistic regression, we can get a feel for how well the model agrees with the data by comparing the observed and predicted logits or probabilities for all possible covariate patterns. For example, in Example 14.6 the eight possible covariate patterns are listed again in Table 14.10 along with observed and predicted logits and probabilities. The observed logits and probabilities come from Table 14.5.

Table 14.10 List of covariate patterns for PFT data in Example 14.6.

Covariate Pattern	Lead Level	Smoking Status			Logit		Probability	
					Observed	Predicted	Observed	Predicted
(i)	x_i	D_{1i}	D_{2i}	D_{3i}	l_i	$\hat{\lambda}_i$	p_i	$\hat{\pi}_i$
1	1	0	0	0	3.332	3.015	0.9655	0.9532
2	1	1	0	0	2.526	2.726	0.9260	0.9385
3	1	0	1	0	2.100	2.248	0.8910	0.9045
4	1	0	0	1	3.689	3.523	0.9756	0.9713
5	0	0	0	0	1.674	2.178	0.8421	0.8982
6	0	1	0	0	2.351	1.889	0.9130	0.8686
7	0	0	1	0	1.789	1.411	0.8571	0.8038
8	0	0	0	1	2.398	2.686	0.9167	0.9362

In multiple linear regression, the residuals provided useful information about possible problems with the model. We can also use the residuals in logistic regression to examine the fit of the logistic model. Two common forms of residuals used in logistic regression are *Pearson residuals* and *deviance residuals*. These residuals are useful for identifying outlying and influential points (Pregibon 1981). The *Pearson residual* is defined as

$$r_i = \frac{y_i - n_i\hat{\pi}_i}{\sqrt{n_i\hat{\pi}_i(1-\hat{\pi}_i)}}$$

where n_i is the number of observations with the ith covariate pattern, y_i is the number of observations with the outcome of interest among n_i observations, and $\hat{\pi}_i$ is the predicted probability of the outcome of interest for the ith covariate pattern. The form of the Pearson residual is familiar — dividing the difference in the observed and predicted cell counts by the standard error of the observed count. We did the same calculations in converting statistics to a standard normal variable. Note that we can also express the numerator of r_i as $y_i - \hat{y}_i$, where $\hat{y}_i$ is equal to $n_i\,\hat{\pi}_i$.

Some recommend a slightly different form of the Pearson residual. For example, according to Collett (2003), a better procedure is to divide the raw residual, $y_i - \hat{y}_i$, by its standard error, $se(y_i - \hat{y}_i)$. This standard error is complicated to derive, but it is used in many of the logistic regression programs. Residuals based on the $se(y_i - \hat{y}_i)$ are known as the *standardized Pearson residuals*.

The *deviance residual* is defined as

$$d_i = \text{sgn}(y_i - n_j\hat{\pi}_j)\left[2y_i\ln\left(\frac{y_i}{n_i\hat{\pi}_i}\right) + 2(n_i - y_i)\ln\left(\frac{n_i - y_i}{n_i(1-\hat{\pi}_i)}\right)\right]^{1/2}$$

where sgn is plus if the quantity in the parenthesis is positive and negative if the quantity is negative.

Table 14.11 Pearson and deviance residuals for the multiple logistic regression model from Example 14.6.

	Residual	
Covariate Pattern	**Pearson**	**Deviance**
1	0.54	0.57
2	−0.47	−0.46
3	−0.34	−0.34
4	0.33	0.34
5	−0.81	−0.75
6	0.63	0.67
7	0.50	0.52
8	−0.48	−0.46
Sum of Squares	**2.28**	**2.25**

Since there are only eight covariate patterns for the PFT data in Example 14.6, we can easily show the Pearson and deviance residuals in Table 14.11 (see **Program Note 14.2** on the website).

14.2.3 Goodness-of-Fit Statistics

We can also use the residuals in testing the goodness of fit of the model. A Pearson test statistic can be calculated by summing the squares of the residuals, that is, Σr_i^2. A similar test statistic based on the deviance residuals is then Σd_i^2. If the model fits, both of these statistics follow a chi-square distribution with degrees of freedom equal to number of covariate patterns minus the number of parameters in the model plus one.

Let's now test the goodness of fit of the model. The null and alternative hypotheses are

H_0: the model fits the data *versus* Hâ: the model does not fit the data.

Because we estimated four parameters in the model and there are eight covariate patterns, there are three degrees of freedom for the chi-square test. If we test the hypothesis that the model fits at the 0.05 level, a value of the test statistic greater than 7.81 is required to reject the null hypothesis. Since both test statistics (values of 2.28 and 2.25 for the Pearson statistic and the deviance statistic, respectively) are smaller than this critical value, we fail to reject the hypothesis that the model fits.

In logistic situations with continuous independent variables, it is likely that the number of distinct covariate patterns will be close to the number of observations. The next example considers this situation.

Example 14.8

We are going to plot the Pearson and deviance residuals by individual for the multiple logistic regression model considered for the diabetes data in Example 14.7 (see **Program Note 14.3** on the website).

Since these residuals have, in effect, been divided by their standard errors — it is hard to see that this statement applies to the deviance residuals, but it does —

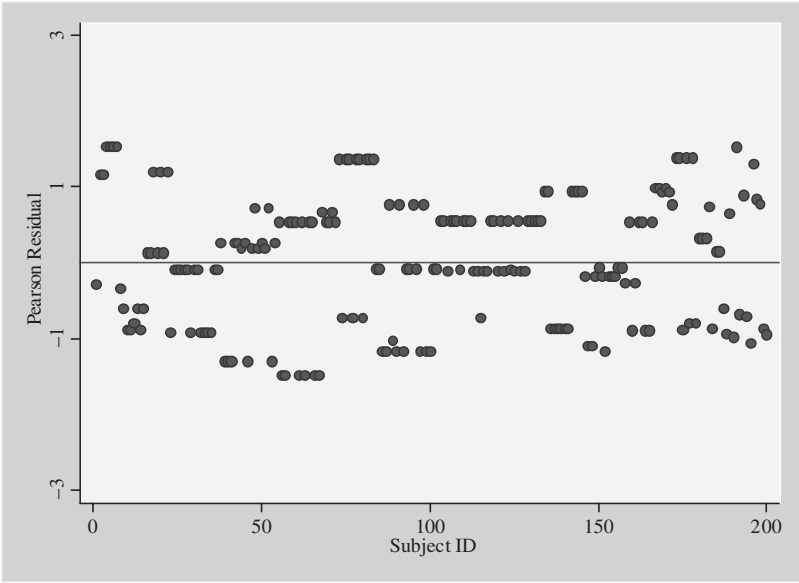

Figure 14.2 Pearson residual by subject for the data in Example 14.7.

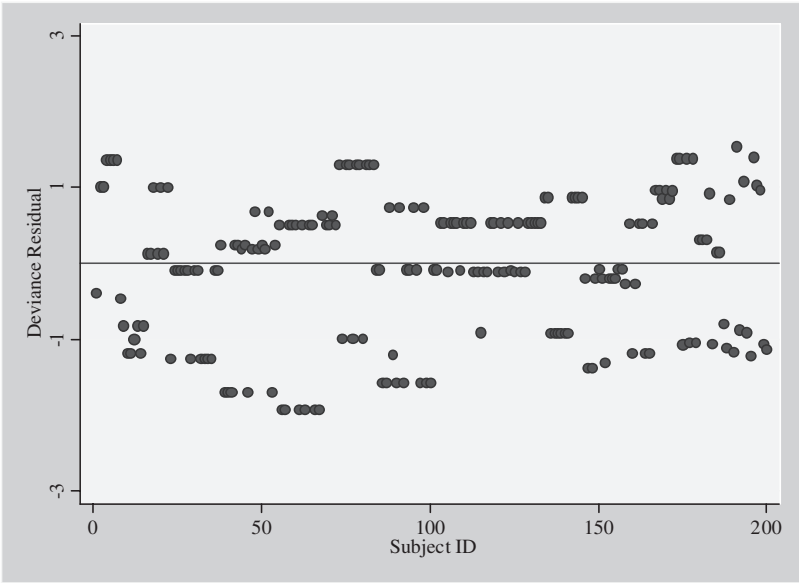

Figure 14.3 Deviance residual by subject for the data in Example 4.7.

residuals that have a value greater than two are of interest. Residuals with a value greater than two could result because of a coding error or simply represent a rare occurrence. We are looking for any patterns in the residuals, similar to the analysis of residuals in multiple linear regression. Since there don't appear to be any large residuals in Figure 14.2 or Figure 14.3, it does not appear that any of the observations require further inspection. If there were large residuals, we could try other plots such as the residuals versus the independent variables as well as doing univariate analysis on the original data looking for anomalies.

In some cases, particularly those with continuous independent variables, we prefer not to use the Pearson and deviance chi-square statistics to test the fit of the model. In these cases, we believe that other tests — for example, the *Hosmer-Lemeshow (H-L)* goodness-of-fit test — have better statistical properties (Hosmer and Lemeshow 1999a). The H-L procedure groups the data into g categories where g is usually 10. The grouping is based on the values of the predicted probabilities from the model. In one approach, the data are grouped into equal-sized ordered categories with the first category having the subjects with the smallest estimated probabilities and so forth to the last group containing the subjects with the largest estimated probabilities. In another approach suggested by Hosmer and Lemeshow, the categories are formed by specific cutpoints — for examples, 0.10, 0.20, . . . , 0.90. The first group contains all subjects with predicted probabilities less than or equal to 0.10, the second group contains all subjects with predicted probabilities greater than 0.10 and less than or equal to 0.20 and so on, to the last group that contains all the subjects with predicted probabilities greater than 0.90. The H-L test statistic is defined as

$$\hat{C} = \sum_{k=1}^{g} \frac{(o_k - n_k' \bar{\pi}_k)^2}{n_k' \bar{\pi}_k (1 - \bar{\pi}_k)}$$

where n_k' is the number of covariate patterns in the kth group, o_k is the number of subjects with the condition of interest in the n_k' covariate patterns, and $\bar{\pi}_k$ is the average predicted probability in the kth group. Based on extensive simulations, the H-L statistic follows the chi-square distribution with $g - 2$ degrees of freedom.

Let's now test the goodness-of-fit of the logistic model used in Example 14.7 at the 0.05 level. We will use the first method of grouping — that is, dividing the data into 10 equal-sized categories. As shown in Table 14.9, the H-L test statistic is 2.532 with 8 degrees of freedom. Since the H-L statistic is less than the critical value of 15.51, we fail to reject the goodness of fit of the model at the 0.05 level.

14.2.4 The ROC Curve

Another measure of how well a logistic regression model performs can be obtained by examining the area under the receiver operating characteristic (ROC) curve, originally presented in Chapter 4, for that model. Recall that the ROC curve is created by plotting 1-specificity against sensitivity at different cutoff points for determining a positive or negative test result. In the logistic model, the sensitivity and specificity can be evaluated at different levels of predicted probabilities by comparing the predicted classification with the observed classification of the dependent variable. The area under the ROC curve provides a measure of the discriminative ability of the logistic model. Hosmer and Lemeshow (1999a) suggest the following guidelines for assessing the discriminatory power of the model:

If the area under the ROC curve (AUROC) is 0.5, the model does not discriminate.
If $0.5 < \text{AUROC} < 0.7$, the model has poor to fair discrimination.
If $0.7 < \text{AUROC} < 0.8$, the model has acceptable discrimination.
If $0.8 \leq \text{AUROC} < 0.9$, the model has excellent discrimination.
If AUROC ≥ 0.9 — a very rare outcome — the model has outstanding discrimination.

Example 14.9

Let's consider the logistic regression model including lead levels and smoking status as predictors of the PFT results shown in Example 14.6 to see how we create the ROC curve. As shown in Table 14.10, there are eight predicted probabilities in this example and we can evaluate sensitivity and specificity at eight different cutoff points. At the lowest predicted probability of 0.8038 (high lead level and former smoker), the predicted PFT status is determined to be "normal" if the predicted probabilities are greater than or equal to 0.8038. The 2 by 2 table shown here can be formed from the cross-tabulation of the data in Table 14.5 by the predicted and observed PFT status. Sensitivity and specificity are calculated from the table using the procedure explained in Chapter 4:

Predicted PFT Status	Observed PFT Status: Normal	Observed PFT Status: Abnormal	
Normal	450	29	Sensitivity = 450/450 = 1.00
Abnormal	0	0	Specificity = 0/29 = 0.00
Total	450	29	

Similarly, we can evaluate sensitivity and specificity at the second lowest predicted probability of 0.8686 (high lead level and light smoker) as follows:

Predicted PFT Status	Observed PFT Status: Normal	Observed PFT Status: Abnormal	
Normal	438	27	Sensitivity = 438/450 = 0.973
Abnormal	12	2	Specificity = 2/29 = 0.069
Total	450	29	

For the rest of the cutoff points the sensitivity and specificity are

Cutoff Point	Sensitivity	Specificity
0.8982	0.927	0.138
0.9045	0.891	0.241
0.9363	0.782	0.448
0.9385	0.709	0.552
0.9532	0.542	0.759
0.9713	0.356	0.862
1.0000	0.000	1.000

From these data the ROC curve can be plotted. We can use a computer program to create the ROC curve and calculate AUROC, as shown in Figure 14.4 (see **Program Note 14.2** on the website).

AUROC can be interpreted as the likelihood that an individual who has a non-normal PFT result will have a higher predicted probability of having a nonnormal PFT than an individual who does not have a nonnormal PFT result (Pregibon 1981). The AUROC value for this example is approximately 0.68, a value suggesting poor to fair discrimination.

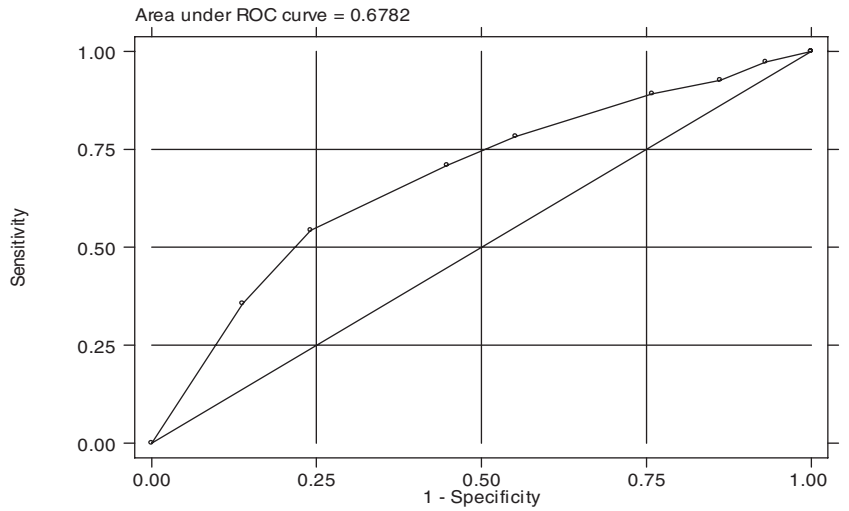

Figure 14.4 Plot of ROC curve for the logistic regression model in Example 14.6.

Many programs also report a pseudo-R^2. Statisticians tend to give less attention to this measure because it may suggest the model has poor explanatory power, whereas other measures such as the AUROC suggest good discriminatory power. The goodness-of-fit tests, the examination of residuals, and the AUROC are three tools with good acceptance by statisticians for examining multiple logistic regression models.

We have provided a few of the numerous diagnostic tools available to the researcher for examining the logistic regression model. The use of additional plots and many other statistics shown in Chapter 13 for examining the fit of the model carry over to logistic regression. To learn more about the application of these other tools, the reader is encouraged to check other sources on logistic regression (Hosmer and Lemeshow 1999; Pregibon 1981). However, we should not automatically delete those subjects identified using these diagnostic methods. Any elimination of subjects must be done very carefully and be based on substantive considerations as well as on diagnostic methods.

14.3 Ordered Logistic Regression

In previous sections, we introduced logistic regression models that have a dependent variable with a dichotomous outcome. However, more complicated forms of logistic regression are also available, and we begin this section by considering an ordinal-dependent variable with more than two levels.

Example 14.10

Let us examine the perceived health status reported in the National Health and Nutrition Examination Survey. The health status is reported as "excellent," "very good," "good," "fair," and "poor." Based on an NHANES III Phase II adult sample, 23.0 percent of U.S. adults reported that their health status is "excellent," 30.3 percent "very good," 31.3 percent "good," and 15.5 percent for the "fair or poor" categories

combined. We want to determine whether or not there is a relationship between health status and the use of vitamin or mineral supplements (1 = use, 0 = nonuse) and education reflected by the number of years of schooling. Note that if a relationship exists, it does not necessarily imply any causal relationship between the variables we are labeling as independent and the variable we are labeling as dependent. It may be that supplement use is a function of health status, or it could be that health status is a function of supplement use or there could be a mixture of relationships. We can't tell from these data the direction of the relationship even if a relationship actually exists.

Given the relatively small number of people in the fair and poor categories we have combined them into one category. Hence, we are now working with four ordered health status categories. Let's start our investigation by looking at health status and supplement use. Before examining the relationship between these two variables, we must decide how to handle this ordinal health status variable. Since there are four levels, there are really only three pieces of independent information. This means that we could create three independent functions that would contain all the information in the health status variable. One such set of functions is the following:

Pr (excellent) versus Pr (all other levels)

Pr (excellent plus very good) versus (good plus fair or poor)

Pr (excellent plus very good plus good) versus (fair or poor).

Given the sample values just mentioned for the probabilities of the various health status states, we would expect the first function to be much less than one, the second function to be close to one, and the third function to be much greater than one. If we take the natural logarithm of the three functions, we would expect the first to be negative, the second close to zero, and the third to be positive.

We could then perform three separate binary logistic regressions to examine the relationships to supplement use. A logistic model that could be used to examine the relationship is

$$\ln(\textit{health-status function}_i) = \textit{constant}_i + \textit{effect of supplement use}_i.$$

However, if the effect of supplement use on health status is consistent for these three functions, we could estimate this "average" effect of supplement use by considering a single model that included the supplement use effect plus three separate constant terms. In effect, this model is

$$\ln(\textit{health-status function}_i) = \textit{constant}_i + \textit{effect of supplement use}.$$

This representation reflects the idea that the regression lines for the different outcome functions are parallel to each other but that they have different intercepts. Table 14.12 shows basic data for this analysis and for checking of the assumption of a consistent effect of supplement use — that is, the odds ratios for each of the health status functions with supplement use are similar. This assumption is called the proportional odds assumption.

Since the odds ratios of 1.19, 1.36, and 1.51 are reasonably similar, we can conclude that the proportional odds assumption seems to be acceptable. We see that

Table 14.12 Perceived health status by use and nonuse of vitamin or mineral supplements, NHANES, Phase II adult subsample (n = 988).

	Perceived Health Status				
Vitamin Use	**Excellent**	**Very Good**	**Good**	**Fair or Poor**	**Total**
User	105	139	127	53	424
Nonuser	122	160	182	100	564
Total	227	299	309	153	988[a]
	(28.0%)	(30.3%)	(31.3%)	(15.5%)	(100.1%)
Comparisons	I		II		III	
	105	319	244	180	371	53
	122	442	282	282	464	100
Odds Ratio	1.19		1.36		1.51	

[a]Excluding cases with missing values

those taking vitamin or mineral supplements are more likely to feel better about their health and vice versa. Given that the proportional odds assumption seems to hold, we can estimate the common odds ratio that summarizes the effect.

Note that it is not uncommon for an ordered logistic regression model not to satisfy the proportionality assumption (or parallel regression assumption). If this assumption is not satisfied, other alternative models should be considered, such as the multinomial logistic model (Hosmer and Lemeshow 1999a).

Table 14.13 shows the results of ordered logistic regression analysis (see **Program Note 14.4** on the website). The top panel shows the ordered logistic regression of health status on supplement use based on the reduced model that assumes the equality of the supplement coefficients for the three health-status variables. In this example, the equality of the supplement coefficients is another way of saying that the lines are parallel or that the odds are proportional for the three health-status variables.

In examining the results, we first look at the test for the goodness of fit of the model. In this case, the goodness-of-fit test examines whether or not the three coefficients for supplement use in the full model are all equal. Based on the goodness-of-fit values from the Pearson and deviance tests, we fail to reject the equality of the coefficients (or that the lines are parallel or that the odds are proportional), a result we expected, since the preceding three odds ratios were fairly similar.

The maximum likelihood estimates of coefficients include the three intercepts and the common supplement effect. The intercepts don't hold much interest for us, but their values are consistent with the expected pattern mentioned above (negative, close to zero, and positive). The estimated coefficient for vitamin use is 0.2835, and the corresponding estimated odds ratio is 1.33. This is the estimated common odds ratio for healthier status, comparing those taking supplements with those not taking supplements. The 95 percent confidence interval for the common odds ratio, the p-value for the test that the coefficient for supplement use is zero, and the g statistic (follows a chi-square distribution) all suggest that there is a significant relationship between supplement use and health status at the 0.05 level.

The preceding analysis could be done using the CMH method presented in Chapter 10. But the ordered logistic regression model allows us to include continuous

Table 14.13 Ordered logistic regression analysis of perceived health status on use of vitamin or mineral supplements and years of schooling, NHANES III, Phase II adult subsample (n = 988).

Model I (health status on vitamin use)

Predictor	Coef	SE Coef	Z	p	Odds Ratio	95% CI Lower Upper
Constant (1)	–1.3384	0.0923	–14.49	<0.001	—	—
Constant (2)	0.0063	0.0808	0.08	0.938	—	—
Constant (3)	1.5808	0.0993	15.92	<0.001	—	—
Vitamin use	0.2835	0.1160	2.44	0.015	1.33	(1.06, 1.67)

Log-likelihood = –1332.777
Test that all slopes are zero: G = 6.004, DF = 1, p-Value = 0.014

Pseudo R-Square = 0.002

Goodness-of-Fit Tests:
Pearson Chi-Square = 1.354, df = 2, p-Value = 0.508
Deviance Chi-Square = 1.357, df = 2, p-Value = 0.507

Model II (health status on vitamin use and years of schooling)

Predictor	Coef	SE Coef	Z	p	Odds Ratio	95% CI Lower Upper
Constant (1)	–4.4268	0.2768	–15.99	<0.001	—	—
Constant (2)	–2.9434	0.2586	–11.38	<0.001	—	—
Constant (3)	–1.1679	0.2452	–4.76	<0.001	—	—
Vitamin use	0.0425	0.1192	0.36	0.722	1.04	(0.83, 1.32)
Schooling	0.2476	0.0205	12.08	<0.001	1.28	(1.23, 1.33)

Score test for the proportional odds assumption:
Chi-Square = 1.594, df = 4, p-Value = 0.810

Log-likelihood = –1254.178
Test that all slopes are zero: G = 163.202, DF = 2, p-Value = <0.001

Pseudo R-Square = 0.061

Goodness-of-Fit Tests:
Pearson Chi-Square = 130.426, df = 100, p-Value = 0.022
Deviance Chi-Square = 119.519, df = 100, p-Value = 0.089

explanatory variables. The results of the logistic regression of health status on supplement use and the number of years of schooling are shown in the bottom panel of Table 14.13. First, our attention is called to goodness-of-fit statistics. Since the Pearson and deviance residual statistics are larger than the degrees of freedom, the key finding here is that this model does not provide a good fit to the data. Given that the model does not fit, there is little reason to place much emphasis on the parameter estimates. However, note that the supplement variable's effect has been greatly reduced when the years of schooling variable is considered. As just stated, it is difficult using data from a point in time to examine relationships over time. In this situation, it is even not clear what variable should be used as the response or dependent variable.

In general, if the outcome variable is ordered and has g categories, we can form $(g - 1)$ independent functions from the outcome variable. The proportional odds model assumes that the odds ratio across all $(g - 1)$ cut-points is the same. Applying the same approach as previously, the proportional odds model for the j = 1, 2, . . . , $g - 1$ functions and p explanatory variables is

$$\ln\left(\frac{\pi_{\leq j}}{\pi_{\leq j+1}}\right) = \beta_{0j} + \sum_{i=1}^{p} \beta_i x_i .$$

The functions used as the dependent variables are the logits of being in the g category or lower versus being in higher categories.

14.4 Conditional Logistic Regression

Data from matched studies can also be analyzed by a logistic regression approach. As discussed in Chapter 6, *matching* is a way of balancing certain characteristics between two groups. If matching is used in the design phase of a study, a treatment is given to one member of a matched pair and a placebo is given to the other. In case-control studies, a case with a particular outcome is matched to a control without the outcome of interest and an examination of a possible relationship to an exposure is assessed retrospectively. Matching can be one to one or one to several controls.

One way of analyzing matched studies is *conditional logistic regression*, a method illustrated in the following example.

Example 14.11

The DIG200 data set contains 27 subjects with cardiovascular disease (CVD) — cases who can be perfectly matched to 27 subjects without CVD — controls based on age, sex, and race. The matched data are shown in Table 14.14.

Table 14.14 Twenty-seven controls and matched cases of cardiovascular disease, DIG200.

Control (without CVD)						Case (with CVD)					
Set	Age	Sex	Race	SBP	MI	Set	Age	Sex	Race	SBP	MI
1	43	1	1	120	1	1	43	1	1	90	0
2	45	1	1	122	0	2	45	1	1	160	1
3	46	1	1	96	1	3	46	1	1	110	1
4	47	2	1	120	0	4	47	2	1	116	0
5	49	1	1	140	0	5	49	1	1	122	1
6	50	1	1	148	1	6	50	1	1	140	0
7	51	2	1	124	1	7	51	2	1	95	0
8	54	1	1	120	1	8	54	1	1	106	0
9	57	1	1	136	0	9	57	1	1	140	1
10	58	2	2	100	0	10	58	2	2	100	1
11	59	1	1	100	1	11	59	1	1	100	0
12	60	1	1	102	1	12	60	1	1	140	1
13	63	1	1	105	0	13	63	1	1	114	0
14	64	1	1	150	1	14	64	1	1	130	0
15	65	1	1	132	0	15	65	1	1	130	1
16	66	1	1	130	1	16	66	1	1	160	0
17	67	1	1	130	1	17	67	1	1	130	1
18	68	2	1	144	1	18	68	2	1	152	1
19	69	1	1	130	0	19	69	1	1	116	0
20	70	1	1	150	1	20	70	1	1	110	0
21	71	1	1	90	0	21	71	1	1	90	0
22	72	2	1	140	0	22	72	2	1	155	0
23	73	1	1	140	1	23	73	1	1	150	0
24	74	1	1	100	1	24	74	1	1	140	1
25	76	1	1	140	0	25	76	1	1	130	0
26	79	1	1	130	0	26	79	1	1	150	1
27	80	2	1	118	1	27	80	2	1	165	0

Let us first look at the relationship between CVD and prior MI. As we discussed in Chapter 10 (Section 10.2.5), the relationship can be summarized in the following 2 by 2 table:

	Prior MI in Controls		
Prior MI in Cases	**Yes**	**No**	**Total**
Yes	5 (c_1)	6 (d_1)	11
No	10 (d_2)	6 (c_2)	16
Total	15	12	27

Relevant information for the analysis of this table is contained in discordant cells (d_1 and d_2), and we used the McNemar chi-square test to test the hypothesis of no relationship between CVD and MI. For the preceding table, the McNemar test statistic is

$$X_M^2 = \frac{(|d_1 - d_2| - 1)^2}{d_1 + d_2} = \frac{(|6 - 10| - 1)^2}{6 + 10} = 0.56.$$

The p-value for this test statistic is 0.454. If we ignore the correction for continuity, the test statistic is 1.00 with p-value of 0.317. There is no statistically significant relationship between CVD and prior MI — that is, a previous MI is not predictive of the occurrence of CVD. The odds ratio for the preceding table is $d_1/d_2 = 6/10 = 0.6$ and the corresponding 95 percent confidence interval is (0.179, 1.82). Since the confidence interval contains the value of 1, there does not appear to be a significant relationship.

Conditional logistic regression offers an alternative method of analysis for matched studies. For example, if we wish to examine whether or not there may be a relationship between the occurrence of CVD (1 = yes, 0 = no) and MI (1 = yes, 0 = no), we will focus on the difference of the variables within each of the 27 pairs because of the matching. The idea of focusing on the differences is similar to the use of differences in the paired t test. The CVD difference is always equal to +1 by definition. The difference in the MI variable can have the value of +1, 0, or −1 and this difference variable is now treated as a continuous variable by the computer software. We can use ordinary logistic regression using the differences as the variables. Since we are using differences, there is no need to include the constant term in the analysis.

The first panel of Table 14.15 shows the results of the logistic regression analyses of the presence and absence of cardiovascular disease on prior myocardial infarction (see **Program Note 14.5** on the website).

The estimated coefficient is −0.5108 (se = 0.5164), which gives the estimated odds ratio as exp(−0.5108) = 0.6. The 95 percent confidence interval is found from the exp(−0.5108 ± 1.96 ∗ 0.5164) or (0.22, 1.65). The odds ratio is exactly the same as found from the 2 by 2 table. The test results also turn out to be very similar to those obtained from the 2 by 2 table. The p-value for McNemar test was 0.317 compared to 0.3147 from the likelihood ratio test for the conditional logistic regression and to 0.323 based on the normal test. Note that we entered the data for 54 observations (27 pairs), but we could have entered just the 16 discordant pairs and obtained the same results, since data for concordant pairs do not contribute anything to the analysis.

Table 14.15 Conditional logistic regression analysis of matched cases of cardiovascular disease on prior myocardial infarction and systolic pressure, 57 pairs from DIG200.

Model I (CVD on prior MI)

Conditional (fixed-effects) logistic regression			Number of obs		=	54
			LR chi2 (1)		=	1.01
			Prob > chi2		=	0.3147
Log likelihood = –18.209631			Pseudo R2		=	0.0270
	Coef.	Std. Err.	z	$p > \|z\|$	Odds Ratio	[95% CI]
Prior MI	–0.5108	0.5164	–0.99	0.323	0.600	(0.218, 1.651)

Model II (CVD on prior MI and systolic blood pressure)

Conditional (fixed-effects) logistic regression			Number of obs		=	54
			LR chi2 (2)		=	2.00
			Prob > chi2		=	0.3683
Log likelihood = –17.716158			Pseudo R2		=	0.0534
	Coef.	Std. Err.	z	$p > \|z\|$	Odds Ratio	[95% CI]
Prior MI	–0.6496	0.5546	–1.17	0.242	0.522	(0.176, 1.549)
SBP	0.0187	0.0195	0.96	0.337	1.019	(0.981, 1.059)

For a simple situation like in the above 2 by 2 table, there is really no need to use the conditional logistic regression model. However, the conditional logistic model is very useful for more complicated situations where multiple predictor variables (including continuous variables) are used or for predictor variables with more than two levels. In the case of a discrete variable, such as the smoking variable in Table 14.5, we use three dummy variables like those shown in Table 14.6 to show the smoking status of a person. But now in our conditional logistic regression model, we are subtracting the smoking status of the control from that of the case. This means that we are now creating three new difference variables having either the value of +1, 0 or –1. Each of these three difference variables reflecting the smoking status would then be entered into the model and treated as if they were continuous variables.

In the model shown in the lower panel in Table 14.15, we entered two predictor variables (prior MI and systolic blood pressure). The results show that the estimated coefficient for MI changed slightly. The estimated odds ratio for prior MI adjusted for systolic blood pressure is 0.52, and its confidence interval still includes one. The normal test for prior MI has a p-value of 0.242, and the p-value for the two-degree-of-freedom test of hypothesis that both the prior MI coefficient and the SBP coefficient are simultaneously zero is 0.368. Hence we may conclude that prior MI appears to have no statistically significant effect on CVD, whether or not we adjust for SBP.

14.5 Introduction to Proportional Hazard Regression

The proportional hazards model introduced by D. R. Cox (1972) is an extension of the material in Chapter 11, and the Cox approach has become the most widely used regression model in survival analysis. In Chapter 11, we introduced the hazard function, defined as the probability of failure during an interval of time divided by the size of the

interval. Cox's regression allows the examination of the possible relationship between the hazard function and a set of independent variables. We use the following example in the introduction of the Cox model.

Example 14.12

The DIG200 data set was introduced in Chapter 3 as part of the digoxin trial. Mortality was monitored and the number of days to death or to the end of the trial for those who were still living. Mortality and the number of days of survival for 200 subjects in the DIG200 dataset are shown in Table 14.16, along with age and BMI rounded to the whole number.

Table 14.16 Survival data for 200 subjects in the Digoxin trial, DIG200.

	Placebo Group				Digoxin Group			
ID	Death	Days to Death	Age	BMI	Death	Days to Death	Age	BMI
1	0	631	70	26	1	627	45	33
2	0	1,166	74	30	0	1,501	66	29
3	1	1,025	65	26	1	431	62	27
4	0	1,508	51	30	1	149	63	23
5	0	1,727	73	28	0	1,335	72	22
6	0	1,167	52	30	1	620	31	27
7	0	1,117	62	29	0	1,157	58	23
8	0	1,544	70	23	0	1,215	55	21
9	0	1,578	52	31	1	1,216	74	26
10	0	1,192	62	22	1	165	28	29
11	1	1,075	65	28	0	880	57	28
12	0	1,052	66	28	0	1,518	63	29
13	1	338	71	33	1	586	69	27
14	0	1,131	58	27	0	1,181	60	23
15	0	1,173	50	27	0	1,136	47	31
16	0	1,432	29	41	0	1,475	79	38
17	0	1,432	68	28	1	169	73	27
18	0	970	46	22	0	1,194	58	26
20	0	1,279	71	21	0	879	71	26
21	1	940	70	19	1	562	63	30
22	0	1,328	57	24	0	1,697	61	23
23	0	1,454	51	20	0	1,591	63	28
24	1	1,516	72	27	0	1,523	58	28
25	0	1,598	84	32	1	415	50	32
26	0	1,355	57	27	0	1,542	66	33
27	0	1,013	59	18	0	1,353	61	27
28	1	901	52	24	0	1,390	77	27
29	1	50	63	20	0	1,060	71	27
30	0	1,726	50	26	0	1,748	73	27
31	0	1,188	46	26	0	1,559	57	26
32	1	825	68	25	0	1,034	68	24
33	1	33	79	30	0	1,680	51	26
34	0	1,501	88	33	1	300	65	24
35	0	1,318	54	31	1	644	56	26
36	1	538	53	34	1	132	66	27
37	1	629	79	27	0	1,528	60	27
38	1	1,359	76	31	1	951	67	26
39	1	374	78	23	0	969	49	15
40	0	887	69	36	0	958	53	26
41	1	790	63	25	0	989	66	29
42	1	966	55	27	0	1,566	66	32
43	0	1,250	60	30	0	1,157	45	24

Table 14.16 *Continued*

	Placebo Group				Digoxin Group			
ID	Death	Days to Death	Age	BMI	Death	Days to Death	Age	BMI
44	0	1,192	55	20	1	949	68	21
45	0	1,108	51	22	0	537	73	22
46	1	1,176	55	22	0	1,279	49	27
47	0	1,160	71	25	0	1,629	57	24
48	1	8	72	23	0	1,277	60	22
49	1	609	69	22	0	1,342	77	22
50	0	1,649	79	19	1	943	72	23
51	1	609	64	21	0	1,626	66	27
52	0	1,374	74	36	0	1,147	42	24
53	0	1,168	78	43	0	867	52	24
54	1	1,268	68	26	0	1,144	54	27
55	0	871	71	22	0	1,152	65	29
56	0	1,516	65	27	1	295	46	36
57	0	1,090	44	26	1	447	67	30
58	1	1,007	62	25	1	511	75	26
59	0	1,391	65	29	0	899	54	27
60	1	547	61	32	0	1,622	58	28
61	1	531	52	25	1	1,567	66	24
62	1	848	64	27	0	1,328	57	24
63	1	305	57	19	0	1,203	61	26
64	1	392	69	25	1	229	65	28
65	0	1,500	76	30	1	1,003	80	25
66	1	1,464	50	34	1	335	77	27
67	0	982	68	27	1	543	46	29
68	0	1,259	54	22	1	1,004	70	19
69	0	1,125	42	24	1	10	35	26
70	0	1,508	43	29	0	895	69	20
71	0	1,559	55	19	0	984	63	21
72	1	299	67	24	0	872	71	23
73	0	1,405	56	35	0	881	53	25
74	0	1,489	47	23	0	1,598	58	29
75	0	1,012	57	23	0	947	80	27
76	1	270	56	20	0	1,588	70	29
77	0	1,298	64	18	0	1,116	38	31
78	0	1,567	81	23	0	1,587	68	25
79	0	873	75	23	1	636	50	24
80	1	1,553	69	22	0	344	54	23
81	0	1,340	43	21	0	1,097	67	33
82	1	340	69	24	1	970	65	45
83	1	188	81	38	0	1,341	76	40
84	0	1,522	59	27	0	1,339	75	38
85	0	1,504	77	24	0	898	59	22
86	1	59	74	29	0	975	47	32
87	1	1,254	67	22	0	1,131	70	37
88	0	949	53	27	0	1,486	49	23
89	0	1,553	76	25	0	1,570	79	27
90	1	895	46	21	1	477	45	27
91	0	1,270	68	28	0	1,287	67	20
92	0	1,228	55	23	0	1,678	37	27
93	1	1,298	83	26	0	1,585	67	34
94	1	1,144	59	25	0	1,350	70	24
95	0	1,669	69	32	0	1,166	69	22
96	0	1,262	61	28	1	1,032	68	24
97	1	253	55	26	1	681	34	20
98	1	495	46	27	0	42	48	30
99	0	1,180	54	32	0	538	60	24
100	1	346	45	23	0	1,612	77	28

Death: 1 = died, 0 = survived; BMI is rounded to the whole number

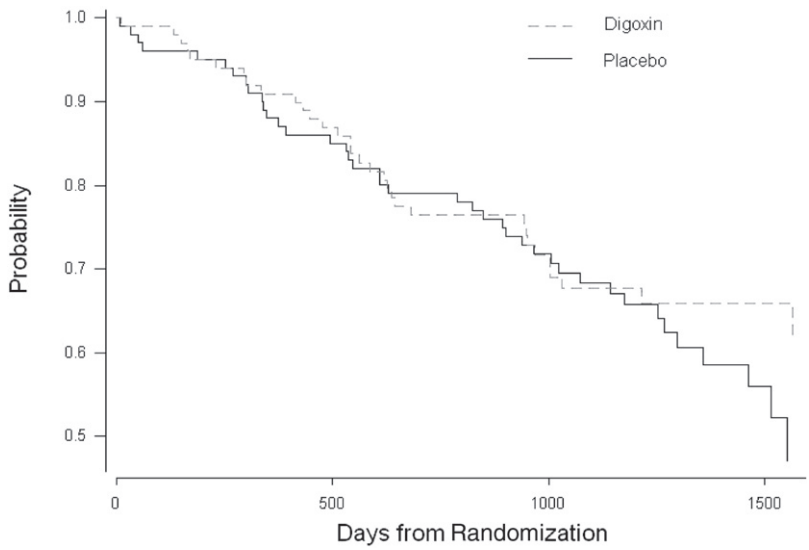

Figure 14.5 Kaplan-Meier curves for digoxon and placebo groups, DIG200.

In the DIG200 dataset, there are 72 deaths: 40 deaths in the placebo group and 32 deaths in the treatment group. To compare survival experience of the two groups, we can use the methods discussed in Chapter 11. As explained in Chapter 11, we need to treat those subjects who were still living at the end of the follow-up period as censored observations. Figure 14.5 shows the Kaplan-Meier survival curves by treatment group.

The Kaplan-Meier curves do not show a noticeable difference in the survival experience between the placebo and treatment group, although survival appears to favor the treatment group slightly after 1200 days. In addition, the hazard plots shown in Figure 14.6 do not show an appreciable difference between the two groups except for later time periods.

Descriptive statistics in Table 14.17 show slightly better survival probabilities for the treatment group. However, the *p*-value of the log-rank test for comparing the two survival distributions is 0.398, indicating that there is no statistically significant benefit to being treated with digoxin.

The Cox proportional hazards regression model offers an alternative method to compare the survival experience of the two groups. The model focuses on the hazards in the two groups. Let $h_0(t)$ be the hazard at time t for the placebo group and $h_1(t)$ be the hazard at time t for the digoxin group. Then the ratio of these two hazards, the *hazard ratio*, can be modeled under the assumption that it is constant at all survival times, t. It implies that

$$\frac{h_1(t)}{h_0(t)} = \phi.$$

The hazard function in the denominator is called the *baseline hazard*. We already encountered this proportional hazards assumption in applying the CMH and log rank

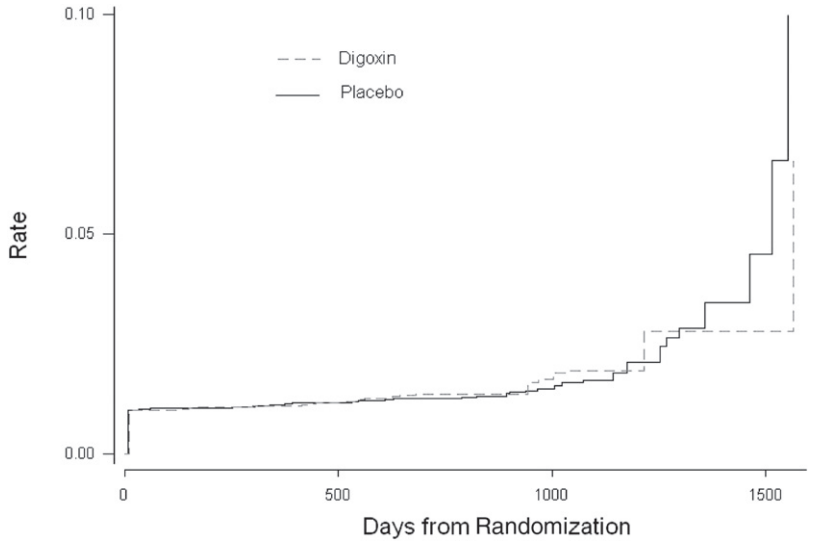

Figure 14.6 Hazard rate plot for digoxin and placebo groups, DIG200.

Table 14.17 Descriptive analysis of survival for digoxin and placebo groups, DIG200.

Descriptor	Digoxin Group (n = 100)	Placebo Group (n = 100)
Number of Deaths	32	40
Survival Probabilities at		
360 days	0.909	0.880
900 days	0.763	0.749
1440 days	0.660	0.583
Survival Percentiles		
25th	949 days	895 days
50th	—	1553 days
Log Rank Test	Chi-square	0.715
	p-value	0.398

tests in Chapter 11. Since the Cox procedure is based on this assumption, it behooves us to examine this assumption. Based on the plot of the hazard rates in Figure 14.6, it appears as if the ratio of the rates is a constant at least as far out as 1300 days. After that, the ratio changes slightly from around one to less than one. We can separate our investigation into two parts, before and after 1300 days, if we want to be safe. If we limit the analysis to the first 1300 days, the log-rank test chi-square value is 0.0062 with a p-value of 0.938. We conclude that there is no difference in survival between the placebo and digoxin groups. There is one death in the digoxin group and four deaths in the placebo group after 1300 days. For purposes of demonstration, we will simply consider the entire follow-up period in our analysis.

Since hazards are always positive, we can substitute e^{β} where β is a parameter with no restrictions (can be positive, zero, or negative) for the quantity ϕ. Using this notation, we can express Cox's regression model as

$$\ln\left(\frac{h_1(t)}{h_0(t)}\right) = \beta x$$

where x is an indicator variable (0 if an individual received a placebo or 1 if an individual received the digoxin treatment). Note that this linear model has no intercept term unlike the general regression model. No intercept is necessary here because we are only concerned with estimating the hazard ratio.

Just as when using the Kaplan-Meier procedure, in the Cox regression model we also must specify the censored observations — that is, those who were still living at the end of follow-up period, when entering the data for analysis. The results of fitting this model to the survival data for the two groups are shown in Table 14.18 (Model I). (see **Program Note 14.6** on the website.) The estimated coefficient is −0.2007 with standard error of 0.2377. The estimated hazard ratio is exp(−0.2007) = 0.82, suggesting that the hazard is 18 percent lower for the treatment group. This is consistent with the slightly favorable survival probabilities for the treatment group shown previously. The 95 percent confidence interval for the estimated hazard ratio is exp(0.2007 ± 1.96 ∗ 0.2377), or (0.51, 1.30). Since the confidence interval contains the value of one, there is not sufficient evidence to conclude that the use of digoxin lowers the risk of dying. Finally, notice that the p-values for the Wald test statistic (0.399) and likelihood ratio test statistic (0.397) are very close to the p-value for the log rank test (0.398), and they all cause us to fail to reject the null hypothesis of no treatment effect.

The Cox model allows us to incorporate additional predictor variables besides the treatment variable. To demonstrate the inclusion of additional variables, we next carry out Cox's regression analyses of survival experience in DIG200 considering treatment status and two continuous variables, age and body mass index. Model II in Table 14.18 considers treatment group status and age as predictor variables. Model III includes treatment group status, age, and body mass index in the model. Since the digoxin trial randomly allocated patients into the two groups, we do not expect that incorporation of age and BMI would change the difference in survival between

Table 14.18 The fit of the proportional hazards regression of survival on digoxin treatment, age, and body mass index: DIG200.

Model I (survival on digoxin treatment)

					LR chi^2 (1)	= 0.72
Log likelihood = −353.81122					Prob > chi^2	= 0.3972
	Coef.	Std. Err.	z	$p > \|z\|$	Odds Ratio	[95% CI]
Treatment	−0.2007	0.2377	−0.84	0.399	0.8182	(0.5134, 1.3037)

Model II (survival on digoxin treatment and age)

					LR chi^2 (2)	= 0.82
Log likelihood = −353.76215					Prob > chi^2	= 0.6653
	Coef.	Std. Err.	z	$p > \|z\|$	Odds Ratio	[95% CI]
Treatment	−0.2015	0.2378	−0.85	0.397	0.8175	(0.5130, 1.3029)
Age	−0.0033	0.0105	−0.31	0.754	0.9967	(0.9765, 1.0174)

Model III (survival on digoxin treatment, age, and body mass index)

					LR chi^2 (3)	= 0.93
Log likelihood = −353.70398					Prob > chi^2	= 0.8178
	Coef.	Std. Err.	z	$p > \|z\|$	Odds Ratio	[95% CI]
Treatment	−0.1980	0.2380	−0.83	0.405	0.8204	(0.5146, 1.3080)
Age	−0.0031	0.0106	−0.29	0.771	0.9969	(0.9765, 1.0178)
BMI	−0.0085	0.0249	−0.34	0.735	0.9916	(0.9443, 1.0413)

treatment and control groups. We are considering these two additional models to illustrate the usefulness of the Cox approach.

In Model II, the estimated hazard ratio for digoxin treatment, for a fixed age, is 0.82, which is the same as the unadjusted hazard ratio in Model I. The estimated hazard ratio for age, in the same group, is 1.00, suggesting that age variable does not make any difference at all. The comparison in the two likelihood values suggests there is no significant age effect. Similarly, in Model III, addition of body mass index to the model does not make a difference. The estimated hazard ratio for digoxin treatment, holding age and BMI constant, is still 0.82.

In general, the proportional hazards model considering k independent variables is expressed in terms of the hazard function

$$h(t) = h_0(t) \cdot \exp(\beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k)$$

where $h_0(t)$ is referred to as the baseline hazard and is multiplied by the exponential of the k independent variables. This model can also be expressed as

$$\ln\left(\frac{h(t)}{h_0(t)}\right) = \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k .$$

The natural log of the hazard ratio is linearly related to the sum of the k independent variables. This equation is similar to the formula for the logit model we used in logistic regression. Independent variables may be discrete or continuous. Discrete independent variables with more than two levels require dummy coding as in the general regression model. For more detailed treatment of proportional hazards regression, we refer to more advanced books (Collett 1994; Cox and Oakes 1984; Hosmer and Lemeshow 1999a).

Conclusion

In this chapter, we showed that logistic regression is a part of the larger general linear model approach for analyzing data. Logistic regression is an important method, particularly in epidemiology, as it allows the investigator to examine the relation between a binary dependent variable and a set of continuous and discrete independent variables. The interpretation of the model parameters in terms of the odds and odds ratios is a key attraction of the logistic regression procedure. Many of the diagnostic procedures used to examine the appropriateness and fit of the multiple linear regression model have also been adapted to logistic regression, making it an even more attractive method. We also briefly introduced the Cox's proportional hazards model as a method that goes beyond the survival analysis methods in Chapter 11. This model allows us to examine multiple factors to determine whether or not there appears to be an association with the length of survival.

This chapter provides an introduction to both of these topics. It is not meant to be exhaustive, particularly regarding the presentation of the proportional hazards model. Interested readers are encouraged to avail themselves of several books that focus on these topics.

EXERCISES

14.1 Data from an article by Madsen (1976) are used here to examine the relation between survival status — less than 10 years or greater than or equal to 10 years — and the type of operation — extensive (total removal of the ovaries and the uterus) and not extensive — for 299 patients with cancer of the ovary. Other factors could be included — for example, stage of the tumor, whether or not radiation was used, and whether or not the tumor had spread — in a logistic regression analysis. However, we begin our consideration with only the one independent variable. The data are

	Survival Status	
Type of Operation	**<10 years**	**≥10 years**
Extensive	29	122
Not Extensive	20	28

In a logistic regression analysis — using the logit for >10 years of survival and the not extensive type of operation as the base level — the estimates of the constant term and the regression coefficient for the type of operation (extensive) are 0.3365 and 0.3920, respectively. Provide an interpretation for these estimates. Demonstrate that your interpretations are correct by relating these estimates to the preceding table.

14.2 Based on DIG200, investigate how previous myocardial infarction (MI) is related to age, race, sex, and BMI. Summarize the computer output in a table and interpret the results. Explain the odd ratios for each independent variable. What is the predicted proportion of having had an MI for a nonwhite female 60 years of age with a BMI of 30?

14.3 The story of the Donner party, stranded in the Sierra Nevada in the winter of 1846–1847, illustrates the hardship of the pioneers' journey to California. Of the 83 members of the Donner party, only 45 survived to reach California. The following data represent sex, age, and survival status of adult members (15 years of age and older).

Person	Sex	Age	Status	Person	Sex	Age	Status
1	M	62	died	23	M	32	survived
2	F	45	died	24	F	23	survived
3	M	56	died	25	M	30	died
4	F	45	died	26	F	19	survived
5	M	20	survived	27	M	30	died
6	M	25	died	28	M	30	survived
7	M	28	died	29	F	30	survived
8	F	32	survived	30	M	57	died
9	F	25	survived	31	F	47	died
10	M	24	died	32	F	20	survived
11	M	28	died	33	M	18	survived
12	M	25	died	34	F	15	survived
13	M	51	survived	35	F	22	survived
14	F	40	survived	36	M	23	died
15	M	35	died	37	M	25	died
16	M	28	survived	38	M	23	died
17	F	25	died	39	M	18	survived
18	F	50	died	40	M	46	survived
19	M	15	died	41	M	25	survived
20	F	23	survived	42	M	60	died
21	M	28	survived	43	M	25	died
22	F	75	died				

Source: http://members.aol.com/DanMRosen/donner/survivor.htm

Run a logistic regression analysis using sex and age as predictor variables for survival and interpret the results. How does a female's odds of survival compare with a male's odds while controlling for age? How does a 45-year-old person's odds of survival compare with a 15-year-old person's odds while controlling for sex?

14.4 Woodward et al. (1995) investigated prevalence of coronary heart diseases (CHD) in men. Prevalent CHD was defined on a four-point graded scale in decreasing order of severity: myocardial infarction (MI), angina grade II, angina grade I, no CHD. One of several risk factors examined was parental history of CHD before age 60. The data are

Parental History of CHD	CHD Categories				
	MI	Angina II	Angina I	No CHD	Total
Present	104	17	45	830	996
Absent	192	30	122	3,376	3,720
Total	296	47	167	4,206	4,716

Analyze the data using an ordered logistic regression model treating the CHD categories as levels of an ordinal dependent variable and parental history of CHD as the independent variable. If the proportional odds assumption is satisfied, summarize the results and interpret the findings. What is the predicted risk of CHD for a person with no CHD?

14.5 Kellermann et al. (1993) investigated the effect of gun ownership on homicide in the home using a retrospective matched-pairs design. They compared 388 cases of homicides with control subjects matched according to neighborhood, sex, race, and age range. They presented the following information in the article:

Number of Gun Owners		
Case	Control	Odds Ratio (95% CI)
174	139	1.6 (1.2 – 2.2)

Is it possible to verify that the reported crude odds ratio is correct? If yes, verify it. If not, what information is lacking? For a multivariate analysis, the following information is shown:

	Odds Ratio (95% CI)	
Variable	Crude	Adjusted
Gun ownership	1.6 (1.2 – 2.2)	2.7 (1.6 – 4.4)
Home rented	5.9 (3.8 – 9.2)	4.4 (2.3 – 8.2)
Lived alone	3.4 (2.2 – 5.1)	3.7 (2.1 – 6.6)
Domestic violence	7.9 (5.0 – 12.7)	4.4 (2.2 – 8.8)
Any household member arrested	4.4 (3.0 – 6.0)	2.5 (1.6 – 4.1)
Any member used illicit drugs	9.0 (5.4 – 15.0)	5.7 (2.6 – 12.6)

Explain what statistical method is used to calculate the adjusted odds ratios and their confidence intervals. How would you interpret the adjusted odds ratio of 2.7 for gun ownership?

14.6 A case-control study of presenile dementia was introduced in Chapter 10 (Example 10.6). Each dementia case was individually paired with a community control of the same sex and age, and family history of dementia was ascertained in both groups, retrospectively. The following cross-tabulation of the 109 pairs by the presence or absence of family history of dementia was analyzed. Based on the McNemar chi-square test statistic, we concluded that there is evidence for an association between dementia and family history of the disease:

Family History of Dementia in Case	Family History of Dementia in Control	
	Present	Absent
Present	6	25
Absent	12	66

The following table shows the data for the 37 discordant pairs. Analyze the data using the conditional logistic regression approach and see whether the same conclusion can be drawn.

Control (without Dementia)			Case (with Dementia)		
Set	Dementia[a]	History[b]	Set	Dementia[a]	History[b]
1	0	1	1	1	0
2	0	1	2	1	0
3	0	1	3	1	0
4	0	1	4	1	0
5	0	1	5	1	0
6	0	1	6	1	0
7	0	1	7	1	0
8	0	1	8	1	0
9	0	1	9	1	0
10	0	1	10	1	0
11	0	1	11	1	0
12	0	1	12	1	0
13	0	0	13	1	1
14	0	0	14	1	1
15	0	0	15	1	1
16	0	0	16	1	1
17	0	0	17	1	1
18	0	0	18	1	1
19	0	0	19	1	1
20	0	0	20	1	1
21	0	0	21	1	1
22	0	0	22	1	1
23	0	0	23	1	1
24	0	0	24	1	1
25	0	0	25	1	1
26	0	0	26	1	1
27	0	0	27	1	1
28	0	0	28	1	1
29	0	0	29	1	1
30	0	0	30	1	1
31	0	0	31	1	1
32	0	0	32	1	1
33	0	0	33	1	1
34	0	0	34	1	1
35	0	0	35	1	1
36	0	0	36	1	1
37	0	0	37	1	1

[a]Codes: 0 = without dementia; 1 = with dementia
[b]Codes: 0 = without history; 1 = with history

14.7 The survival of 64 lymphoma patients was analyzed for two different symptom groups (A and B) in Exercise 11.4. The survival times (in months) for the two symptom groups is shown here:

A symptoms:	3.2*	4.4*	6.2	9.0	9.9	14.4	15.8	18.5	27.6*	28.5	30.1*
	31.5*	32.2*	41.0	41.8*	44.5*	47.8*	50.6*	54.3*	55.0	60.0*	60.4*
	63.6*	63.7*	63.8*	66.1*	68.0*	68.7*	68.8*	70.9*	71.5*	75.3*	75.7*
B symptoms:	2.5	4.1	4.6	6.4	6.7	7.4	7.6	7.7	7.8	8.8	13.3
	13.4	18.3	19.7	21.9	24.7	27.5	29.7	30.1*	32.9	33.5	35.4*
	37.7*	40.9*	42.6*	45.4*	48.5	48.9*	60.4*	64.4*	66.4*		

Asterisks indicate censored observations.

Analyze the data using Cox's regression method and see whether the same conclusion can be drawn as in Exercise 11.4. Do you think that the proportional hazards assumption is acceptable in your analysis?

REFERENCES

Berkson, J. "Application of the Logistic Function to Bio-Assay." *Journal of the American Statistical Association* 39:357–365, 1944.

Berkson, J. "Why I Prefer Logits to Probits." *Biometrics* 7:327–339, 1951.

Collett, D. *Modelling Survival Data in Medical Research*. London: Chapman & Hall, 1994.

Collett, D. *Modelling Binary Data*, 2nd ed. London: Chapman & Hall, 2003.

Cox, D. R. "Regression Models and Life Table (with discussion)." *Journal of Royal Statistical Society* B 74:187–220, 1972.

Cox, D. R. *Analysis of Binary Data*. London: Chapman and Hall, 1969.

Cox, D. R., and D. Oakes. *Analysis of Survival Data*. London: Chapman & Hall, 1984.

Forthofer, R. N., and R. G. Lehnen. *Public Program Analysis: A New Categorical Data Approach*. Belmont, CA: Lifetime Learning Publications, 1981, Table 7.1.

Hauck, W. W., and A. Donner. "Wald's Test as Applied to Hypotheses in Logit Analysis." *Journal of the American Statistical Association* 72:851–853, 1977.

Hosmer, D. W., and S. Lemeshow. *Applied Logistic Regression*, 2nd ed. New York: John Wiley & Sons, 1999a.

Hosmer, D. W., and S. Lemeshow. *Applied Survival Analysis: Regression Modeling of Time to Event*. New York: John Wiley & Sons, 1999b.

Jennings, D. E. "Judging Inference Adequacy in Logistic Regression." *Journal of the American Statistical Association* 81:471–476, 1986.

Kellermann, A. L., F. P. Rivara, N. B. Rushforth, J. G. Banton, D. T. Reay, J. T. Francisco, A. B. Locci, J. Prodzinski, B. B. Hackman, and G. Somes. "Gun Ownership as Risk Factor for Homicide in the Home." *New England Journal of Medicine* 329(15):1084–1091, October 7, 1993.

Madsen, M. "Statistical Analysis of Multiple Contingency Tables: Two Examples." *Scandanavian Journal of Statistics* 3:97–106, 1976.

Pregibon, D. "Logistic Regression Diagnostics." *The Annals of Statistics* 9:705–724, 1981.

Woodward, M., K. Laurent, and H. Tunstall-Pedoe. "An Analysis of Risk Factors for Prevalent Coronary Heart Disease using Proportional Odds Model." *The Statistician* 44:69–80, 1995.