CHAPTER 16

Logistic Regression

Introduction

The simple and multiple linear regression methods we studied in Chapters 10 and 11 are used to model the relationship between a quantitative response variable and one or more explanatory variables. A key assumption for these models is that the deviations from the model fit are normally distributed. In this chapter we describe similar methods that are used when the response variable has only two possible values.

Our response variable has only two values: success or failure, live or die, acceptable or not. If we let the two values be 1 and 0, the mean is the proportion of ones, $p = P(\text{success})$. With n independent observations, we have the *binomial setting* (page 335). What is *new* here is that we have data on an *explanatory variable* x. We study how p depends on x. For example, suppose we are studying whether a patient lives, ($y = 1$) or dies ($y = 0$) after being admitted to a hospital. Here, p is the probability that a patient lives and possible explanatory variables include (a) whether the patient is in good condition or in poor condition, (b) the type of medical problem that the patient has, and (c) the age of the patient. Note that the explanatory variables can be either categorical or quantitative. Logistic regression[1] is a statistical method for describing these kinds of relationships.

16.1 The Logistic Regression Model

Binomial distributions and odds

In Chapter 5 we studied binomial distributions and in Chapter 8 we learned how to do statistical inference for the proportion p of successes in the binomial setting. We start with a brief review of some of these ideas that we will need in this chapter.

EXAMPLE 16.1 Example 8.1 (page 537) describes a survey of 17,096 students in U.S. four-year colleges. The researchers were interested in estimating the proportion of students who are frequent binge drinkers. A student who reports drinking five or more drinks in a row three or more times in the past two weeks is called a frequent binge drinker. In the notation of Chapter 5, p is the proportion of frequent binge drinkers in the entire population of college students in U.S. four-year colleges. The number of frequent binge drinkers in an SRS of size n has the binomial distribution with parameters n and p. The sample size is $n = 17{,}096$ and the number of frequent binge drinkers in the sample is 3314. The sample proportion is

$$\hat{p} = \frac{3314}{17{,}096} = 0.1938$$

odds Logistic regressions work with **odds** rather than proportions. The odds are simply the ratio of the proportions for the two possible outcomes. If $\hat{p}$ is

the proportion for one outcome, then $1 - \hat{p}$ is the proportion for the second outcome:

$$\text{ODDS} = \frac{\hat{p}}{1 - \hat{p}}$$

A similar formula for the population odds is obtained by substituting p for $\hat{p}$ in this expression.

EXAMPLE 16.2 For the binge-drinking data the proportion of frequent binge drinkers in the sample is $\hat{p} = 0.1938$, so the proportion of students who are not frequent binge drinkers is

$$1 - \hat{p} = 1 - 0.1938 = 0.8062$$

Therefore, the odds of a student being a frequent binge drinker are

$$\begin{aligned}\text{ODDS} &= \frac{\hat{p}}{1 - \hat{p}} \\ &= \frac{0.1938}{0.8062} \\ &= 0.24\end{aligned}$$

When people speak about odds, they often round to integers or fractions. Since 0.24 is approximately 1/4, we could say that the odds that a college student is a frequent binge drinker are 1 to 4. In a similar way, we could describe the odds that a college student is *not* a frequent binge drinker as 4 to 1.

In Example 8.9 (page 557) we compared the proportions of frequent binge drinkers among men and women college students using a confidence interval. There we found that the proportion for men was 0.227 (22.7%) and that the proportion for women was 0.170 (17.0%). The difference is 0.057, and the 95% confidence interval is (0.045, 0.069). We can summarize this result by saying, "The proportion of frequent binge drinkers is 5.7% higher among men than among women."

Another way to analyze these data is to use logistic regression. The explanatory variable is gender, a categorical variable. To use this in a regression (logistic or otherwise), we need to use a numeric code. The usual way to do this is with an **indicator variable.** For our problem we will use an indicator of whether or not the student is a man:

indicator variable

$$x = \begin{cases} 1 & \text{if the student is a man} \\ 0 & \text{if the student is a woman} \end{cases}$$

The response variable is the proportion of frequent binge drinkers. For use in a logistic regression, we perform two transformations on this variable. First, we convert to odds. For men,

$$\begin{aligned}\text{ODDS} &= \frac{\hat{p}}{1 - \hat{p}} \\ &= \frac{0.227}{1 - 0.227} \\ &= 0.294\end{aligned}$$

Similarly, for women we have

$$\begin{aligned}\text{ODDS} &= \frac{\hat{p}}{1-\hat{p}} \\ &= \frac{0.170}{1-0.170} \\ &= 0.205\end{aligned}$$

Model for logistic regression

In simple linear regression we modeled the mean μ of the response variable y as a linear function of the explanatory variable: $\mu = \beta_0 + \beta_1 x$. With logistic regression we are interested in modeling the mean of the response variable p in terms of an explanatory variable x. We could try to relate p and x through the equation $p = \beta_0 + \beta_1 x$. Unfortunately, this is not a good model. As long as $\beta_1 \neq 0$, extreme values of x will give values of $\beta_0 + \beta_1 x$ that are inconsistent with the fact that $0 \leq p \leq 1$.

log odds

The logistic regression solution to this difficulty is to transform the odds ($p/(1-p)$) using the natural logarithm. We use the term **log odds** for this transformation. We model the log odds as a linear function of the explanatory variable:

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x$$

Figure 16.1 graphs the relationship between p and x for some different values of β_0 and β_1. For logistic regression we use *natural* logarithms. There are tables of natural logarithms, and many calculators have a built-in function for this transformation. As we did with linear regression, we use y for the

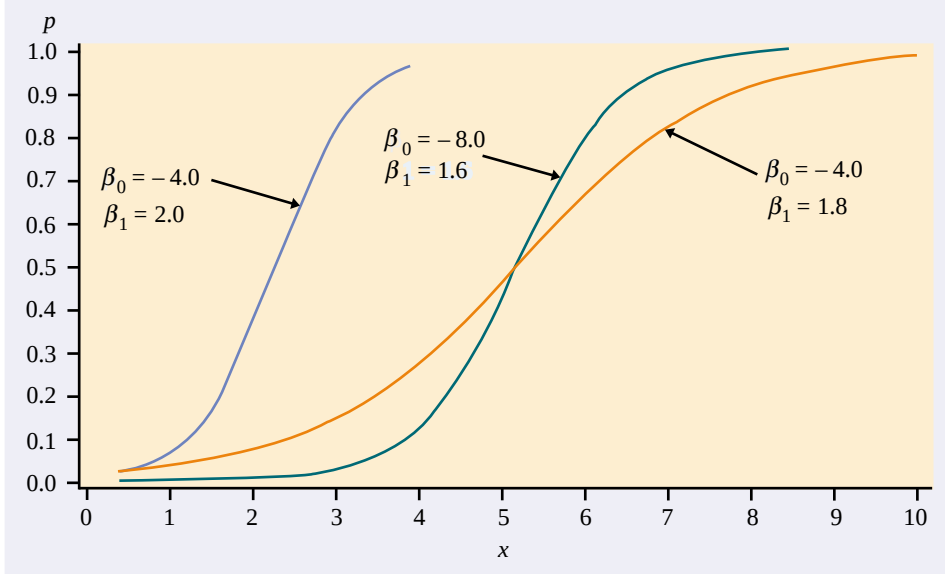

FIGURE 16.1 Plot of p versus x for selected values of β_0 and β_1.

response variable. So for men,

$$y = \log(\text{ODDS}) = \log(0.294) = -1.23$$

and for women,

$$y = \log(\text{ODDS}) = \log(0.205) = -1.59$$

In these expressions we use y as the observed value of the response variable, the log odds of being a frequent binge drinker. We are now ready to build the logistic regression model.

LOGISTIC REGRESSION MODEL

The **statistical model for logistic regression** is

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x$$

where p is a binomial proportion and x is the explanatory variable. The parameters of the logistic model are β_0 and β_1.

EXAMPLE 16.3 For our binge-drinking example, there are $n = 17{,}096$ students in the sample. The explanatory variable is gender, which we have coded using an indicator variable with values $x = 1$ for men and $x = 0$ for women. The response variable is also an indicator variable. Thus, the student is either a frequent binge drinker or the student is not a frequent binge drinker. Think of the process of randomly selecting a student and recording the values of x and whether or not the student is a frequent binge drinker. The model says that the probability (p) that this student is a frequent binge drinker depends upon the student's gender ($x = 1$ or $x = 0$). So there are two possible values for p, say, p_{men} and p_{women}.

Logistic regression with an indicator explanatory variable is a very special case. It is important because many multiple logistic regression analyses focus on one or more such variables as the primary explanatory variables of interest. For now, we use this special case to understand a little more about the model.

The logistic regression model specifies the relationship between p and x. Since there are only two values for x, we write both equations. For men,

$$\log\left(\frac{p_{\text{men}}}{1-p_{\text{men}}}\right) = \beta_0 + \beta_1$$

and for women,

$$\log\left(\frac{p_{\text{women}}}{1-p_{\text{women}}}\right) = \beta_0$$

Note that there is a β_1 term in the equation for men because $x = 1$, but it is missing in the equation for women because $x = 0$.

Fitting and interpreting the logistic regression model

In general, the calculations needed to find estimates b_0 and b_1 for the parameters β_0 and β_1 are complex and require the use of software. When the explanatory variable has only two possible values, however, we can easily find the estimates. This simple framework also provides a setting where we can learn what the logistic regression parameters mean.

EXAMPLE 16.4 In the binge-drinking example, we found the log odds for men,

$$y = \log\left(\frac{\hat{p}_{\text{men}}}{1-\hat{p}_{\text{men}}}\right) = -1.23$$

and for women,

$$y = \log\left(\frac{\hat{p}_{\text{women}}}{1-\hat{p}_{\text{women}}}\right) = -1.59$$

The logistic regression model for men is

$$\log\left(\frac{p_{\text{men}}}{1-p_{\text{men}}}\right) = \beta_0 + \beta_1$$

and for women it is

$$\log\left(\frac{p_{\text{women}}}{1-p_{\text{women}}}\right) = \beta_0$$

To find the estimates of b_0 and b_1, we match the male and female model equations with the corresponding data equations. Thus, we see that the estimate of the intercept b_0 is simply the log(ODDS) for the women:

$$b_0 = -1.59$$

and the slope is the difference between the log(ODDS) for the men and the log(ODDS) for the women:

$$b_1 = -1.23 - (-1.59) = 0.36$$

The fitted logistic regression model is

$$\log(\text{ODDS}) = -1.59 + 0.36x$$

The slope in this logistic regression model is the difference between the log(ODDS) for men and the log(ODDS) for women. Most people are not comfortable thinking in the log(ODDS) scale, so interpretation of the results in terms of the regression slope is difficult. Usually, we apply a transformation to help us. With a little algebra, it can be shown that

$$\frac{\text{ODDS}_{\text{men}}}{\text{ODDS}_{\text{women}}} = e^{0.36} = 1.43$$

odds ratio

The transformation $e^{0.36}$ undoes the logarithm and transforms the logistic regression slope into an **odds ratio,** in this case, the ratio of the odds that a

man is a frequent binge drinker to the odds that a woman is a frequent binge drinker. In other words, we can multiply the odds for women by the odds ratio to obtain the odds for men:

$$\text{ODDS}_{\text{men}} = 1.43 \times \text{ODDS}_{\text{women}}$$

In this case, the odds for men are 1.43 times the odds for women.

Notice that we have chosen the coding for the indicator variable so that the regression slope is positive. This will give an odds ratio that is greater than 1. Had we coded women as 1 and men as 0, the signs of the parameters would be reversed, the fitted equation would be $\log(\text{ODDS}) = 1.59 - 0.36x$, and the odds ratio would be $e^{-0.36} = 0.70$. The odds for women are 70% of the odds for men.

Logistic regression with an explanatory variable having two values is a very important special case. Here is an example where the explanatory variable is quantitative.

EXAMPLE 16.5 The CHEESE data set described in the Data Appendix includes a response variable called "Taste" that is a measure of the quality of the cheese in the opinions of several tasters. For this example, we will classify the cheese as acceptable (tasteok = 1) if Taste $\geq$ 37 and unacceptable (tasteok = 0) if Taste $<$ 37. This is our response variable. The data set contains three explanatory variables: "Acetic," "H2S," and "Lactic." Let's use Acetic as the explanatory variable. The model is

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x$$

where p is the probability that the cheese is acceptable and x is the value of Acetic. The model for estimated log odds fitted by software is

$$\log(\text{ODDS}) = b_0 + b_1 x = -13.71 + 2.25x$$

The odds ratio is $e^{b_1} = 9.48$. This means that if we increase the acetic acid content x by one unit, we increase the odds that the cheese will be acceptable by about 9.5 times.

16.2 Inference for Logistic Regression

Statistical inference for logistic regression is very similar to statistical inference for simple linear regression. We calculate estimates of the model parameters and standard errors for these estimates. Confidence intervals are formed in the usual way, but we use standard normal z^*-values rather than critical values from the t distributions. The ratio of the estimate to the standard error is the basis for hypothesis tests. Often the test statistics are given as the squares of these ratios, and in this case the P-values are obtained from the chi-square distributions with 1 degree of freedom.

Confidence Intervals and Significance Tests

CONFIDENCE INTERVALS AND SIGNIFICANCE TESTS FOR LOGISTIC REGRESSION PARAMETERS

A **level C confidence interval for the slope β_1** is

$$b_1 \pm z^* \mathrm{SE}_{b_1}$$

The ratio of the odds for a value of the explanatory variable equal to $x + 1$ to the odds for a value of the explanatory variable equal to x is the **odds ratio.**

A **level C confidence interval for the odds ratio e^{β_1}** is obtained by transforming the confidence interval for the slope

$$(e^{b_1 - z^* \mathrm{SE}_{b_1}},\ e^{b_1 + z^* \mathrm{SE}_{b_1}})$$

In these expressions z^* is the value for the standard normal density curve with area C between $-z^*$ and z^*.

To test the hypothesis $H_0: \beta_1 = 0$, compute the **test statistic**

$$z = \frac{b_1}{\mathrm{SE}_{b_1}}$$

The P-value for the significance test of H_0 against $H_a: \beta_1 \neq 0$ is computed using the fact that when the null hypothesis is true, z has approximately a standard normal distribution.

Wald statistic

The statistic z is sometimes called a **Wald statistic.** Output from some statistical software reports the significance test result in terms of the square of the z statistic

$$X^2 = x^2$$

chi-square statistic

This statistic is called a **chi-square statistic.** When the null hypothesis is true, it has a distribution that is approximately a χ^2 distribution with 1 degree of freedom, and the P-value is calculated as $P(\chi^2 \geq X^2)$. Because the square of a standard normal random variable has a χ^2 distribution with 1 degree of freedom, the z statistic and the chi-square statistic give the same results for statistical inference.

We have expressed the hypothesis-testing framework in terms of the slope β_1 because this form closely resembles what we studied in simple linear regression. In many applications, however, the results are expressed in terms of the odds ratio. A slope of 0 is the same as an odds ratio of 1, so we often express the null hypothesis of interest as "the odds ratio is 1." This means that the two odds are equal and the explanatory variable is not useful for predicting the odds.

SPSS

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1	GENDERM	0.362	0.039	86.611	1	0.000	1.435
	Constant	–1.587	0.027	3520.069	1	0.000	0.205

a Variable(s) entered on step 1: GENDERM

SAS

The LOGISTIC Procedure

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	–1.5868	0.0267	3520.3120	< 0.001
genderm	1	0.3617	0.0388	86.6811	< 0.001

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
genderm	1.436	1.330	1.549

FIGURE 16.2 Logistic regression output from SPSS and SAS for the binge-drinking data, for Example 16.6.

EXAMPLE 16.6 Figure 16.2 gives the output from SPSS and SAS for the binge-drinking example. The parameter estimates are given as $b_0 = -1.5869$ and $b_1 = 0.3616$, the same as we calculated directly in Example 16.4, but with more significant digits. The standard errors are 0.0267 and 0.0388. A 95% confidence interval for the slope is

$$b_1 \pm z^* \mathrm{SE}_{b_1} = 0.3616 \pm (1.96)(0.0388)$$
$$= 0.3616 \pm 0.0760$$

We are 95% confident that the slope is between 0.2856 and 0.4376. The output provides the odds ratio 1.436 but does not give the confidence interval. This is easy to compute from the interval for the slope:

$$(e^{b_1 - z^* \mathrm{SE}_{b_1}},\ e^{b_1 + z^* \mathrm{SE}_{b_1}}) = (e^{0.2855},\ e^{0.4376})$$
$$= (1.33,\ 1.55)$$

For this problem we would report, "College men are more likely to be frequent binge drinkers than college women (odds ratio = 1.44, 95% CI = 1.33 to 1.55)."

In applications such as these, it is standard to use 95% for the confidence coefficient. With this convention, the confidence interval gives us the result of testing the null hypothesis that the odds ratio is 1 for a significance level of

0.05. If the confidence interval does not include 1, we reject H_0 and conclude that the odds for the two groups are different; if the interval does include 1, the data do not provide enough evidence to distinguish the groups in this way.

The following example is typical of many applications of logistic regression. Here there is a designed experiment with five different values for the explanatory variable.

EXAMPLE 16.7 An experiment was designed to examine how well the insecticide rotenone kills an aphid, called *Macrosiphoniella sanborni*, that feeds on the chrysanthemum plant.[2] The explanatory variable is the concentration (in log of milligrams per liter) of the insecticide. At each concentration, approximately 50 insects were exposed. Each insect was either killed or not killed. We summarize the data using the number killed. The response variable for logistic regression is the log odds of the proportion killed. Here are the data:

Concentration (log)	Number of insects	Number killed
0.96	50	6
1.33	48	16
1.63	46	24
2.04	49	42
2.32	50	44

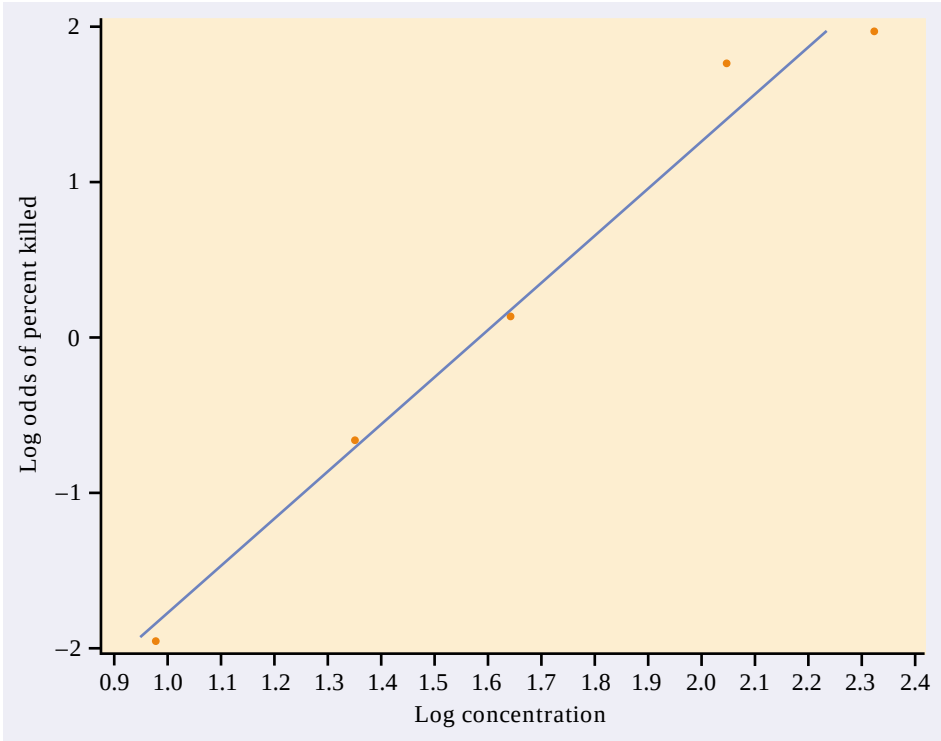

FIGURE 16.3 Plot of log odds of percent killed versus log concentration for the insecticide data, for Example 16.7.

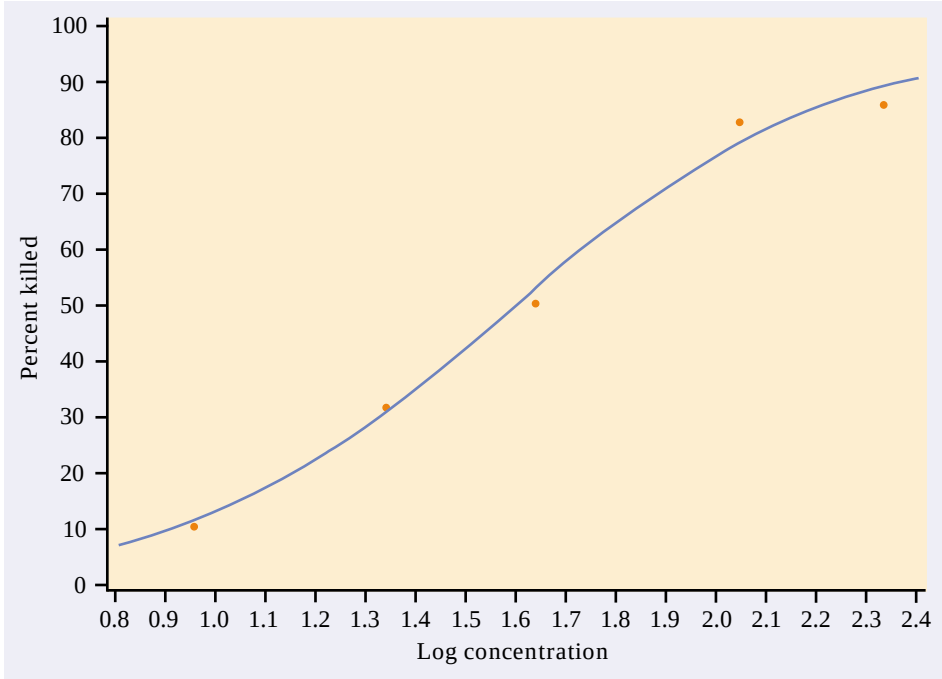

FIGURE 16.4 Plot of the percent killed versus log concentration with the logistic fit for the insecticide data, for Example 16.7.

If we transform the response variable (by taking log odds) and use least squares, we get the fit illustrated in Figure 16.3. The logistic regression fit is given in Figure 16.4. It is a transformed version of Figure 16.3 with the fit calculated using the logistic model.

One of the major themes of this text is that we should present the results of a statistical analysis with a graph. For the insecticide example we have done this with Figure 16.4 and the results appear to be convincing. But suppose that rotenone has no ability to kill *Macrosiphoniella sanborni*. What is the chance that we would observe experimental results at least as convincing as what we observed if this supposition were true? The answer is the *P*-value for the test of the null hypothesis that the logistic regression slope is zero. If this *P*-value is not small, our graph may be misleading. Statistical inference provides what we need.

EXAMPLE 16.8 Figure 16.5 gives the output from SPSS, SAS, and Minitab logistic regression analysis of the insecticide data. The model is

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x$$

where the values of the explanatory variable x are 0.96, 1.33, 1.63, 2.04, 2.32. From the output we see that the fitted model is

$$\log(\text{ODDS}) = b_0 + b_1 x = -4.89 + 3.10x$$

SPSS

Variables in the Equation

	B	S.E.	Wald	df	Sig.	Exp(B)	95.0% C.I. for EXP(B) Lower	Upper
LCONC	3.109	0.388	64.233	1	0.000	22.394	10.470	47.896
Constant	–4.892	0.643	57.961	1	0.000	0.008		

SAS

The LOGISTIC Procedure

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	–4.8923	0.6426	57.9606	< 0.001
lconc	1	3.1088	0.3879	64.2332	< 0.001

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
lconc	22.394	10.470	47.896

Minitab

```
Logistic Regression Table

                                                    Odds         95% CI
Predictor     Coef    StDev        Z        P     Ratio      Lower  Upper
Constant  -4.8923   0.6426    -7.61    0.000
1conc      3.1088   0.3879     8.01    0.000     22.39      10.47  47.90
```

FIGURE 16.5 Logistic regression output from SPSS, SAS, and Minitab for the insecticide data, for Example 16.8.

This is the fit that we plotted in Figure 16.4. The null hypothesis that $\beta_1 = 0$ is clearly rejected ($X^2 = 64.07$, $P < 0.001$). We calculate a 95% confidence interval for β_1 using the estimate $b_1 = 3.1035$ and its standard error $\text{SE}_{b_1} = 0.3877$ given in the output:

$$b_1 \pm z^* \text{SE}_{b_1} = 3.1035 \pm (1.96)(0.3877)$$
$$= 3.1035 \pm 0.7599$$

We are 95% confident that the true value of the slope is between 2.34 and 3.86.

The odds ratio is given on the output as 22.277. An increase of one unit in the log concentration of insecticide (x) is associated with a 22-fold increase in the odds that an insect will be killed. The confidence interval for the odds is obtained from the interval for the slope:

$$(e^{b_1 + z^* \text{SE}_{b_1}},\ e^{b_1 - z^* \text{SE}_{b_1}}) = (e^{2.34361},\ e^{3.86339})$$
$$= (10.42,\ 47.63)$$

Note again that the test of the null hypothesis that the slope is 0 is the same as the test of the null hypothesis that the odds are 1. If we were reporting the results in terms of the odds, we could say, "The odds of killing an insect increase by a factor of 22.3 for each unit increase in the log concentration of insecticide ($X^2 = 64.07$, $P < 0.001$; 95% CI = 10.4 to 47.6)."

In Example 16.5 we studied the problem of predicting whether or not the taste of cheese was acceptable using Acetic as the explanatory variable. We now revisit this example and show how statistical inference is an important part of the conclusion.

EXAMPLE 16.9 Figure 16.6 gives the output from Minitab for a logistic regression analysis using Acetic as the explanatory variable. The fitted model is

$$\log(\text{ODDS}) = b_0 + b_1 x = -13.71 + 2.25x$$

This agrees up to rounding with the result reported in Example 16.5. From the output we see that because $P = 0.0285$, we can reject the null hypothesis that $\beta_1 = 0$. The value of the test statistic is $X^2 = 4.79$ with 1 degree of freedom. We use the estimate $b_1 = 2.2490$ and its standard error $\text{SE}_{b_1} = 1.0271$ to compute the 95% confidence interval for β_1:

$$\begin{aligned} b_1 \pm z^* \text{SE}_{b_1} &= 2.2490 \pm (1.96)(1.0271) \\ &= 2.2490 \pm 2.0131 \end{aligned}$$

Our estimate of the slope is 2.25 and we are 95% confident that the true value is between 0.24 and 4.26. For the odds ratio, the estimate on the output is 9.48. The 95% confidence interval is

$$\begin{aligned} (e^{b_1 + z^* \text{SE}_{b_1}},\ e^{b_1 - z^* \text{SE}_{b_1}}) &= (e^{0.23588},\ e^{4.26212}) \\ &= (1.27,\ 70.96) \end{aligned}$$

We estimate that increasing the acetic acid content of the cheese by one unit will increase the odds that the cheese will be acceptable by about 9 times. The data, however, do not give us a very accurate estimate. The odds ratio could be as small as a little more than 1 or as large as 71 with 95% confidence. We have evidence to conclude that cheeses with higher concentrations of acetic acid are more likely to be acceptable, but establishing the true relationship accurately would require more data.

```
Logistic Regression Table

                                                    Odds        95% CI
Predictor     Coef    StDev        Z        P     Ratio     Lower  Upper
Constant  -13.705    5.932   -2.31     0.21
acetic      2.249    1.027    2.19    0.029      9.48      1.27  70.96
```

FIGURE 16.6 Logistic regression output from Minitab for the cheese data with Acetic as the explanatory variable, for Example 16.9.

SPSS

Omnibus Tests of Model Coefficients

	Chi-square	df	Sig.
Model	16.334	3	0.001

Variables in the Equation

	B	S.E.	Wald	df	Sig.	Exp(B)	95.0% C.I. for EXP(B)	
							Lower	Upper
ACETIC	0.584	1.544	0.143	1	0.705	1.794	0.087	37.001
H2S	0.685	0.404	2.873	1	0.090	1.983	0.898	4.379
LACTIC	3.468	2.650	1.713	1	0.191	32.084	0.178	5776.637
Constant	–14.260	8.287	2.961	1	0.085	0.000		

SAS

Testing Global Null Hypothesis: BETA = 0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	16.3344	3	0.0010

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	–14.2604	8.2869	2.9613	0.0853
acetic	1	0.5845	1.5442	0.1433	0.7051
h2s	1	0.6848	0.4040	2.8730	0.0901
lactic	1	3.4684	2.6497	1.7135	0.1905

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
acetic	1.794	10.087	37.004
h2s	1.983	0.898	4.379
lactic	32.086	0.178	>999.999

Minitab

```
Logistic Regression Table

                                                    Odds        95%   CI
Predictor      Coef    StDev       Z         P    Ratio    Lower     Upper
Constant  -14.260    8.287   -1.72    0.085
acetic      0.584    1.544    0.38    0.705     1.79     0.09     37.01
h2s        0.6849   0.4040    1.69    0.909     1.98     0.90      4.38
lactic      3.468    2.650    1.31    0.191    32.09     0.18   5777.85

Log-Likelihood = -9.230
Test that all slopes are zero: G = 16.334, DF = 3, P-Value = 0.001
```

FIGURE 16.7 Logistic regression output from SPSS, SAS, and Minitab for the cheese data with Acetic, H2S, and Lactic as the explanatory variables, for Example 16.10.

Multiple logistic regression

multiple logistic regression

The cheese example that we just considered naturally leads us to the next topic. The data set includes three variables: Acetic, H2S, and Lactic. We examined the model where Acetic was used to predict the odds that the cheese was acceptable. Do the other explanatory variables contain additional information that will give us a better prediction? We use **multiple logistic regression** to answer this question. Generating the computer output is easy, just as it was when we generalized simple linear regression with one explanatory variable to multiple linear regression with more than one explanatory variable in Chapter 11. The statistical concepts are similar, although the computations are more complex. Here is the example.

EXAMPLE 16.10 As in Example 16.9, we predict the odds that the cheese is acceptable. The explanatory variables are Acetic, H2S, and Lactic. Figure 16.7 gives the outputs from SPSS, SAS, and Minitab for this analysis. The fitted model is

$$\begin{aligned}\log(\text{ODDS}) &= b_0 + b_1\ \text{Acetic} + b_2\ \text{H2S} + b_3\ \text{Lactic}\\ &= -14.26 + 0.58\ \text{Acetic} + 0.68\ \text{H2S} + 3.47\ \text{Lactic}\end{aligned}$$

When analyzing data using multiple regression, we first examine the hypothesis that all of the regression coefficients for the explanatory variables are zero. We do the same for logistic regression. The hypothesis

$$H_0: \beta_1 = \beta_2 = \beta_3 = 0$$

is tested by a chi-square statistic with 3 degrees of freedom. For Minitab, this is given in the last line of the output and the statistic is called "G." The value is $G = 16.33$ and the P-value is 0.001. We reject H_0 and conclude that one or more of the explanatory variables can be used to predict the odds that the cheese is acceptable. We now examine the coefficients for each variable and the tests that each of these is 0. The P-values are 0.71, 0.09, and 0.19. None of the null hypotheses, $H_0: \beta_1 = 0$, $H_0: \beta_2 = 0$, and $H_0: \beta_3 = 0$, can be rejected.

Our initial multiple logistic regression analysis told us that the explanatory variables contain information that is useful for predicting whether or not the cheese is acceptable. Because the explanatory variables are correlated, however, we cannot clearly distinguish which variables or combinations of variables are important. Further analysis of these data using subsets of the three explanatory variables is needed to clarify the situation. We leave this work for the exercises.

CHAPTER 16 | Summary

If $\hat{p}$ is the sample proportion, then the **odds** are $\hat{p}/(1-\hat{p})$, the ratio of the proportion of times the event happens to the proportion of times the event does not happen.

The **logistic regression model** relates the log of the odds to the explanatory variable:

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_i$$

where the response variables for $i = 1, 2, \ldots, n$ are independent binomial random variables with parameters 1 and p_i; that is, they are independent with distributions $B(1,\ p_i)$. The explanatory variable is x.

The **parameters** of the logistic model are β_0 and β_1.

The **odds ratio** is e^{β_1}, where β_1 is the slope in the logistic regression model.

A **level C confidence interval for the intercept** β_0 is

$$b_0 \pm z^* \mathrm{SE}_{b_0}$$

A **level C confidence interval for the slope** β_1 is

$$b_1 \pm z^* \mathrm{SE}_{b_1}$$

A **level C confidence interval for the odds ratio** e^{β_1} is obtained by transforming the confidence interval for the slope

$$(e^{b_1 - z^* \mathrm{SE}_{b_1}},\ e^{b_1 + z^* \mathrm{SE}_{b_1}})$$

In these expressions z^* is the value for the standard normal density curve with area C between $-z^*$ and z^*.

To test the hypothesis $H_0\colon \beta_1 = 0$, compute the **test statistic**

$$z = \frac{b_1}{\mathrm{SE}_{b_1}}$$

and use the fact that z has a distribution that is approximately the standard normal when the null hypothesis is true. This statistic is sometimes called the **Wald statistic.** An alternative equivalent procedure is to report the square of z,

$$X^2 = z^2$$

This statistic has a distribution that is approximately a χ^2 distribution with 1 degree of freedom, and the P-value is calculated as $P(\chi^2 \geq X^2)$. This is the same as testing the null hypothesis that the odds ratio is 1.

In **multiple logistic regression** the response variable has two possible values, as in logistic regression, but there can be several explanatory variables.

CHAPTER 16 | Exercises

16.1 For each of the following, explain what is wrong and why.

(a) In logistic regression with one explanatory variable we can use a chi-square statistic to test the null hypothesis $H_0\colon b_1 = 0$ versus a two-sided alternative.

(b) For a logistic regression we assume that the error term in our model has a normal distribution.

(c) For a multiple logistic regression with 5 explanatory variables, the null hypothesis that the regression coefficients of all of the explanatory variables are zero is tested with an F test.

16.2 In Example 9.12 (page 591) we studied data on the success of 170 franchise firms and whether or not the owner of a franchise had an exclusive territory. Here are the data:

Observed numbers of firms

	Exclusive territory		
Success	**Yes**	**No**	**Total**
Yes	108	15	123
No	34	13	47
Total	142	28	170

(a) What proportion of the exclusive-territory firms are successful?

(b) Find the proportion for the firms that do not have exclusive territories.

(c) Convert the proportion you found in part (a) to odds. Do the same for the proportion you found in part (b).

(d) Find the log of each of the odds that you found in part (c).

16.3 Following complaints about the working conditions in some apparel factories both in the United States and abroad, a joint government and industry commission recommended in 1998 that companies that monitor and enforce proper standards be allowed to display a "No Sweat" label on their products. Does the presence of these labels influence consumer behavior?

A survey of U.S. residents aged 18 or older asked a series of questions about how likely they would be to purchase a garment under various conditions. For some conditions, it was stated that the garment had a "No Sweat" label; for others, there was no mention of such a label. On the basis of the responses, each person was classified as a "label user" or a "label nonuser."[3] Suppose we want to examine the data for a possible gender effect. Here are the data for comparing women and men:

Gender	*n*	Number of label users
Women	296	63
Men	251	27

(a) For each gender find the proportion of label users.

(b) Convert each of the proportions that you found in part (a) to odds.

(c) Find the log of each of the odds that you found in part (b).

16.4 Refer to Exercise 16.2. Use $x = 1$ for the exclusive territories and $x = 0$ for the other territories.

(a) Find the estimates b_0 and b_1.

(b) Give the fitted logistic regression model.

(c) What is the odds ratio for exclusive territory versus no exclusive territory?

16.5 Refer to Exercise 16.3. Use $x = 1$ for women and $x = 0$ for men.

(a) Find the estimates b_0 and b_1.

(b) Give the fitted logistic regression model.

(c) What is the odds ratio for women versus men?

16.6 If we apply the exponential function to the fitted model in Example 16.9, we get

$$\text{ODDS} = e^{-13.71+2.25x} = e^{-13.71} \times e^{2.25x}$$

Show that for any value of the quantitative explanatory variable x, the odds ratio for increasing x by 1,

$$\frac{\text{ODDS}_{x+1}}{\text{ODDS}_x}$$

is $e^{2.25} = 9.49$. This justifies the interpretation given after Example 16.9.

16.7 Refer to Example 16.8. Suppose that you wanted to report a 99% confidence interval for β_1. Show how you would use the information provided in the outputs shown in Figure 16.5 to compute this interval.

16.8 Refer to Example 16.8 and the outputs in Figure 16.5. Using the estimate b_1 and its standard error, find the 95% confidence interval for the odds ratio and verify that this agrees with the interval given by the software.

16.9 The Minitab output in Figure 16.5 does not give the value of X^2. The column labeled "Z" provides similar information.

(a) Find the value under the heading "Z" for the predictor lconc. Verify that Z is simply the estimated coefficient divided by its standard error. This is a z statistic that has approximately the standard normal distribution if the null hypothesis (slope 0) is true.

(b) Show that the square of z is X^2. The two-sided P-value for z is the same as P for X^2.

(c) Draw sketches of the standard normal and the chi-square distribution with 1 degree of freedom. (*Hint:* You can use the information in Table F to sketch the chi-square distribution.) Indicate the value of the z and the X^2 statistics on these sketches and use shading to illustrate the P-value.

16.10 Exercise 9.20 (page 617) presents some results of a study about how advertisers use sexual imagery to appeal to young people. The clothing worn by the model in each of 1509 ads was classified as "not sexual" or "sexual" based on

a standardized criterion. A logistic regression was used to describe the probability that the clothing in the ad was "not sexual" as a function of several explanatory variables. Here are some of the reported results:

Explanatory variable	*b*	Wald (*z*) test
Reader age	0.50	13.64
Model sex	1.31	72.15
Men's magazines	−0.05	0.06
Women's magazines	0.45	6.44
Constant	−2.32	135.92

Reader age is coded as 0 for young adult and 1 for mature adult. Therefore, the coefficient of 0.50 for this explanatory variable suggests that the probability that the model clothing is *not* sexual is higher when the target reader age is mature adult. In other words, the model clothing is more likely to be sexual when the target reader age is young adult. Model sex is coded as 0 for female and 1 for male. The explanatory variable men's magazines is 1 if the intended readership is men and 0 for women's magazines and magazines intended for both men and women. Women's magazines is coded similarly.

(a) State the null and alternative hypotheses for each of the explanatory variables.

(b) Perform the significance tests associated with the Wald statistics.

(c) Interpret the sign of each of the statistically significant coefficients in terms of the probability that the model clothing is sexual.

(d) Write an equation for the fitted logistic regression model.

16.11 Refer to the previous exercise. The researchers also reported odds ratios with 95% confidence intervals for this logistic regression model. Here is a summary:

		95% confidence limits	
Explanatory variable	Odds ratio	Lower	Upper
Reader age	1.65	1.27	2.16
Model sex	3.70	2.74	5.01
Men's magazines	0.96	0.67	1.37
Women's magazines	1.57	1.11	2.23

(a) Explain the relationship between the confidence intervals reported here and the results of the Wald *z* significance tests that you found in the previous exercise.

(b) Interpret the results in terms of the odds ratios.

(c) Write a short summary explaining the results. Include comments regarding the usefulness of the fitted coefficients versus the odds ratios in making a summary.

16.12 A poll of 811 adults aged 18 or older asked about purchases that they intended to make for the upcoming holiday season.[4] One of the questions asked what kind of gift they intended to buy for the person on whom they intended to spend the most. Clothing was the first choice of 487 people.

(a) What proportion of adults said that clothing was their first choice?

(b) What are the odds that an adult will say that clothing is his or her first choice?

(c) What proportion of adults said that something other than clothing was their first choice?

(d) What are the odds that an adult will say that something other than clothing is his or her first choice?

(e) How are your answers to parts (a) and (d) related?

16.13 Different kinds of companies compensate their key employees in different ways. Established companies may pay higher salaries, while new companies may offer stock options that will be valuable if the company succeeds. Do high-tech companies tend to offer stock options more often than other companies? One study looked at a random sample of 200 companies. Of these, 91 were listed in the *Directory of Public High Technology Corporations*, and 109 were not listed. Treat these two groups as SRSs of high-tech and non-high-tech companies. Seventy-three of the high-tech companies and 75 of the non-high-tech companies offered incentive stock options to key employees.[5]

(a) What proportion of the high-tech companies offers stock options to their key employees? What are the odds?

(b) What proportion of the non-high-tech companies offers stock options to their key employees? What are the odds?

(c) Find the odds ratio using the odds for the high-tech companies in the numerator. Describe the result in a few sentences.

16.14 Refer to the previous exercise.

(a) Find the log odds for the high-tech firms. Do the same for the non-high-tech firms.

(b) Define an explanatory variable x to have the value 1 for high-tech firms and 0 for non-high-tech firms. For the logistic model, we set the log odds equal to $\beta_0 + \beta_1 x$. Find the estimates b_0 and b_1 for the parameters β_0 and β_1.

(c) Show that the odds ratio is equal to e^{b_1}.

16.15 Refer to Exercises 16.13 and 16.14. Software gives 0.3347 for the standard error of b_1.

(a) Find the 95% confidence interval for β_1.

(b) Transform your interval in (a) to a 95% confidence interval for the odds ratio.

(c) What do you conclude?

16.16 Refer to Exercises 16.13 to 16.15. Repeat the calculations assuming that you have twice as many observations with the same proportions. In other words, assume that there are 182 high-tech firms and 218 non-high-tech firms. The numbers of firms offering stock options are 146 for the high-tech group and 150 for the non-high-tech group. The standard error of b_1 for this scenario is 0.2366. Summarize your results, paying particular attention to what remains the same and what is different from what you found in Exercises 16.13 to 16.15.

16.17 There is much evidence that high blood pressure is associated with increased risk of death from cardiovascular disease. A major study of this association examined 3338 men with high blood pressure and 2676 men with low blood pressure. During the period of the study, 21 men in the low-blood-pressure group and 55 in the high-blood-pressure group died from cardiovascular disease.

(a) Find the proportion of men who died from cardiovascular disease in the high-blood-pressure group. Then calculate the odds.

(b) Do the same for the low-blood-pressure group.

(c) Now calculate the odds ratio with the odds for the high-blood-pressure group in the numerator. Describe the result in words.

16.18 To what extent do syntax textbooks, which analyze the structure of sentences, illustrate gender bias? A study of this question sampled sentences from 10 texts.[6] One part of the study examined the use of the words "girl," "boy," "man," and "woman." We will call the first two words juvenile and the last two adult. Here are data from one of the texts:

Gender	n	X(juvenile)
Female	60	48
Male	132	52

(a) Find the proportion of the female references that are juvenile. Then transform this proportion to odds.

(b) Do the same for the male references.

(c) What is the odds ratio for comparing the female references to the male references? (Put the female odds in the numerator.)

16.19 Refer to the study of cardiovascular disease and blood pressure in Exercise 16.17. Computer output for a logistic regression analysis of these data gives the estimated slope $b_1 = 0.7505$ with standard error $\text{SE}_{b_1} = 0.2578$.

(a) Give a 95% confidence interval for the slope.

(b) Calculate the X^2 statistic for testing the null hypothesis that the slope is zero and use Table F to find an approximate P-value.

(c) Write a short summary of the results and conclusions.

16.20 The data from the study of gender bias in syntax textbooks given in Exercise 16.18 are analyzed using logistic regression. The estimated slope is $b_1 = 1.8171$ and its standard error is $SE_{b_1} = 0.3686$.

(a) Give a 95% confidence interval for the slope.

(b) Calculate the X^2 statistic for testing the null hypothesis that the slope is zero and use Table F to find an approximate *P*-value.

(c) Write a short summary of the results and conclusions.

16.21 The results describing the relationship between blood pressure and cardiovascular disease are given in terms of the change in log odds in Exercise 16.19.

(a) Transform the slope to the odds and the 95% confidence interval for the slope to a 95% confidence interval for the odds.

(b) Write a conclusion using the odds to describe the results.

16.22 The gender bias in syntax textbooks is described in the log odds scale in Exercise 16.20.

(a) Transform the slope to the odds and the 95% confidence interval for the slope to a 95% confidence interval for the odds.

(b) Write a conclusion using the odds to describe the results.

16.23 To be competitive in global markets, many U.S. corporations are undertaking major reorganizations. Often these involve "downsizing" or a "reduction in force" (RIF), where substantial numbers of employees are terminated. Federal and various state laws require that employees be treated equally regardless of their age. In particular, employees over the age of 40 years are in a "protected" class, and many allegations of discrimination focus on comparing employees over 40 with their younger coworkers. Here are the data for a recent RIF:

	Over 40	
Terminated	**No**	**Yes**
Yes	7	41
No	504	765

(a) Write the logistic regression model for this problem using the log odds of a RIF as the response variable and an indicator for over and under 40 years of age as the explanatory variable.

(b) Explain the assumption concerning binomial distributions in terms of the variables in this exercise. To what extent do you think that these assumptions are reasonable?

(c) Software gives the estimated slope $b_1 = 1.3504$ and its standard error $SE_{b_1} = 0.4130$. Transform the results to the odds scale. Summarize the results and write a short conclusion.

(d) If additional explanatory variables were available, for example, a performance evaluation, how would you use this information to study the RIF?

16.24 The Ping Company makes custom-built golf clubs and competes in the $4 billion golf equipment industry. To improve its business processes, Ping decided to seek ISO 9001 certification.[7] As part of this process, a study of the time it took to repair golf clubs sent to the company by mail determined that 16% of orders were sent back to the customers in 5 days or less. Ping examined the processing of repair orders and made changes. Following the changes, 90% of orders were completed within 5 days. Assume that each of the estimated percents is based on a random sample of 200 orders. Use logistic regression to examine how the odds that an order will be filled in 5 days or less has improved. Write a short report summarizing your results.

16.25 To devise effective marketing strategies it is helpful to know the characteristics of your customers. A study compared demographic characteristics of people who use the Internet for travel arrangements and of people who do not.[8] Of 1132 Internet users, 643 had completed college. Among the 852 nonusers, 349 had completed college. Model the log odds of using the Internet to make travel arrangements with an indicator variable for having completed college as the explanatory variable. Summarize your findings.

16.26 The study mentioned in the previous exercise also asked about income. Among Internet users, 493 reported income of less than $50,000 and 378 reported income of $50,000 or more. (Not everyone answered the income question.) The corresponding numbers for nonusers were 477 and 200. Repeat the analysis using an indicator variable for income of $50,000 or more as the explanatory variable. What do you conclude?

16.27 A study of alcohol use and deaths due to bicycle accidents collected data on a large number of fatal accidents.[9] For each of these, the individual who died was classified according to whether or not there was a positive test for alcohol and by gender. Here are the data:

Gender	*n*	*X*(tested positive)
Female	191	27
Male	1520	515

Use logistic regression to study the question of whether or not gender is related to alcohol use in people who are fatally injured in bicycle accidents.

16.28 In Examples 16.5 and 16.9, we analyzed data from the CHEESE data set described in the Data Appendix. In those examples, we used Acetic as the explanatory variable. Run the same analysis using H2S as the explanatory variable.

16.29 Refer to the previous exercise. Run the same analysis using Lactic as the explanatory variable.

16.30 For the cheese data analyzed in Examples 16.9, 16.10, and the two exercises above, there are three explanatory variables. There are three different logistic regressions that include two explanatory variables. Run these. Summarize the

results of these analyses, the ones using each explanatory variable alone, and the one using all three explanatory variables together. What do you conclude?

The following four exercises use the CSDATA data set described in the Data Appendix. We examine models for relating success as measured by the GPA to several explanatory variables. In Chapter 11 we used multiple regression methods for our analysis. Here, we define an indicator variable, say HIGPA, to be 1 if the GPA is 3.0 or better and 0 otherwise.

16.31 Use a logistic regression to predict HIGPA using the three high school grade summaries as explanatory variables.

(a) Summarize the results of the hypothesis test that the coefficients for all three explanatory variables are zero.

(b) Give the coefficient for high school math grades with a 95% confidence interval. Do the same for the two other predictors in this model.

(c) Summarize your conclusions based on parts (a) and (b).

16.32 Use a logistic regression to predict HIGPA using the two SAT scores as explanatory variables.

(a) Summarize the results of the hypothesis test that the coefficients for both explanatory variables are zero.

(b) Give the coefficient for the SAT math score with a 95% confidence interval. Do the same for the SAT verbal score.

(c) Summarize your conclusions based on parts (a) and (b).

16.33 Run a logistic regression to predict HIGPA using the three high school grade summaries and the two SAT scores as explanatory variables. We want to produce an analysis that is similar to that done for the case study in Chapter 11.

(a) Test the null hypothesis that the coefficients of the three high school grade summaries are zero; that is, test H_0: $\beta_{\text{HSM}} = \beta_{\text{HSS}} = \beta_{\text{HSE}} = 0$.

(b) Test the null hypothesis that the coefficients of the two SAT scores are zero; that is, test H_0: $\beta_{\text{SATM}} = \beta_{\text{SATV}} = 0$.

(c) What do you conclude from the tests in (a) and (b)?

16.34 In this exercise we investigate the effect of gender on the odds of getting a high GPA.

(a) Use gender to predict HIGPA using a logistic regression. Summarize the results.

(b) Perform a logistic regression using gender and the two SAT scores to predict HIGPA. Summarize the results.

(c) Compare the results of parts (a) and (b) with respect to how gender relates to HIGPA. Summarize your conclusions.

16.35 Here is an example of Simpson's paradox, *the reversal of the direction of a comparison or an association when data from several groups are combined to form a*

single group. The data concern two hospitals, A and B, and whether or not patients undergoing surgery died or survived. Here are the data for all patients:

	Hospital A	Hospital B
Died	63	16
Survived	2037	784
Total	2100	800

And here are the more detailed data where the patients are categorized as being in good condition or poor condition:

Good condition		
	Hospital A	Hospital B
Died	6	8
Survived	594	592
Total	600	600

Poor condition		
	Hospital A	Hospital B
Died	57	8
Survived	1443	192
Total	1500	200

(a) Use a logistic regression to model the odds of death with hospital as the explanatory variable. Summarize the results of your analysis and give a 95% confidence interval for the odds ratio of Hospital A relative to Hospital B.

(b) Rerun your analysis in (a) using hospital and the condition of the patient as explanatory variables. Summarize the results of your analysis and give a 95% confidence interval for the odds ratio of Hospital A relative to Hospital B.

(c) Explain Simpson's paradox in terms of your results in parts (a) and (b).

CHAPTER 16 | Notes

1. Logistic regression models for the general case where there are more than two possible values for the response variable have been developed. These are considerably more complicated and are beyond the scope of our present study. For more information on logistic regression, see A. Agresti, *An Introduction to Categorical Data Analysis*, Wiley, 1996; and D. W. Hosmer and S. Lemeshow, *Applied Logistic Regression*, Wiley, 1989.

2. This example is taken from a classical text written by a contemporary of R. A. Fisher, the person who developed many of the fundamental ideas of statistical inference that we use today. The reference is D. J. Finney, *Probit Analysis*, Cambridge University Press, 1947. Although not included in the analysis, it is important to note that the experiment included a control group that received no insecticide. No aphids died in this group. We have chosen to call the response "dead." In the text the category is described as "apparently dead, moribund, or so badly affected as to be unable to walk more than a few steps." This is an early example of the need to make careful judgments when

defining variables to be used in a statistical analysis. An insect that is "unable to walk more than a few steps" is unlikely to eat very much of a chrysanthemum plant!

3. Marsha A. Dickson, "Utility of no sweat labels for apparel customers: profiling label users and predicting their purchases," *Journal of Consumer Affairs*, 35 (2001), pp. 96–119.

4. The poll is part of the American Express Retail Index Project and is reported in *Stores*, December 2000, pp. 38–40.

5. Based on Greg Clinch, "Employee compensation and firms' research and development activity," *Journal of Accounting Research*, 29 (1991), pp. 59–78.

6. Monica Macaulay and Colleen Brice, "Don't touch my projectile: gender bias and stereotyping in syntactic examples," *Language*, 73, no. 4 (1997), pp. 798–825.

7. Based on Robert T. Driescher, "A quality swing with Ping," *Quality Progress*, August 2001, pp. 37–41.

8. Karin Weber and Weley S. Roehl, "Profiling people searching for and purchasing travel products on the World Wide Web," *Journal of Travel Research*, 37 (1999), pp. 291–298.

9. Guohua Li and Susan P. Baker, "Alcohol in fatally injured bicyclists," *Accident Analysis and Prevention*, 26 (1994), pp. 543–548.