



# Logistic Regression

# 14

CHAPTER

## 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. In this chapter we describe similar methods that are used when the response variable is a categorical variable with two possible values, such as a student applicant receives or does not receive financial aid, a patient lives or dies during emergency surgery, or your cell phone coverage is acceptable or not.

In general, we call the two outcomes of the response variable “success” and “failure” and represent them by 1 (for a success) and 0 (for a failure). The mean is then the proportion of 1s,  $p = P(\text{success})$ . If our data are  $n$  independent observations, we have the *binomial setting*. What is *new* in this chapter is that the data now include at least one *explanatory variable*  $x$  and the probability  $p$  depends on the value of  $x$ . For example, suppose that we are studying whether a student applicant receives ( $y = 1$ ) or is denied ( $y = 0$ ) financial aid. Here,  $p$  is the probability that an applicant receives aid, and possible explanatory variables include (a) the financial support of the parents, (b) the income and savings of the applicant, and (c) whether the applicant has received financial aid before. Just as in multiple linear regression, the explanatory variables can be either categorical or quantitative. Logistic regression is a statistical method for describing these kinds of relationships.<sup>1</sup>

- 14.1 The Logistic Regression Model
- 14.2 Inference for Logistic Regression

← **LOOK BACK**  
binomial setting,  
p. 322

## 14.1 The Logistic Regression Model

When you complete this section, you will be able to

- Find the odds from a single probability.
- Describe the statistical model for logistic regression with a single explanatory variable.
- Find the odds ratio for comparing two proportions.

### 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

**14.1 A break from Facebook.** Example 8.1 (page 488) describes a Pew Internet survey of 1006 adults living in the United States. The 525 people who reported that they were Facebook users were asked, “Have you ever voluntarily taken a break from Facebook for a period of several weeks or more?” A total of 320 responded, “Yes, I have done this.”

In the notation of Chapter 5,  $p$  is the proportion of U.S. adult Facebook users who took a break from Facebook. The number of adults who took a break in an SRS of size  $n$  has the binomial distribution with parameters  $n$  and  $p$ . The sample size of Facebook users is  $n = 525$  and the number who took a break is the count  $X = 320$ . The sample proportion is

$$\hat{p} = \frac{320}{525} = 0.6095$$

← **LOOK BACK**  
odds, p. 632

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

**14.2 Odds of taking a break.** For the Facebook data, the proportion of adults who took a break in the sample of Facebook users is  $\hat{p} = 0.6095$ , so the proportion of adults who did not take a break is

$$1 - \hat{p} = 1 - 0.6095 = 0.3905$$

Therefore, the odds of taking a break are

$$\begin{aligned}\text{odds} &= \frac{\hat{p}}{1 - \hat{p}} \\ &= \frac{0.6095}{0.3905} \\ &= 1.561\end{aligned}$$

When people speak about odds, they often round to integers or fractions. If we round 1.561 to  $1.5 = 3/2$ , we would say that the odds are approximately 3 to 2 that a Facebook user took a break. In a similar way, we could describe the odds that a Facebook user did *not* take a break as 2 to 3.

### USE YOUR KNOWLEDGE

- 14.1 Odds of drawing an ace.** If you deal one card from a standard deck, the probability that the card is an ace is  $4/52 = 1/13$ . Find the odds of drawing an ace.
- 14.2 Given the odds, find the probability.** If you know the odds, you can find the probability by solving the odds equation for the probability. So,  $\hat{p} = \text{odds}/(\text{odds} + 1)$ . If the odds of an outcome are 4 (or 4 to 1), what is the probability of the outcome?

### Odds for two groups

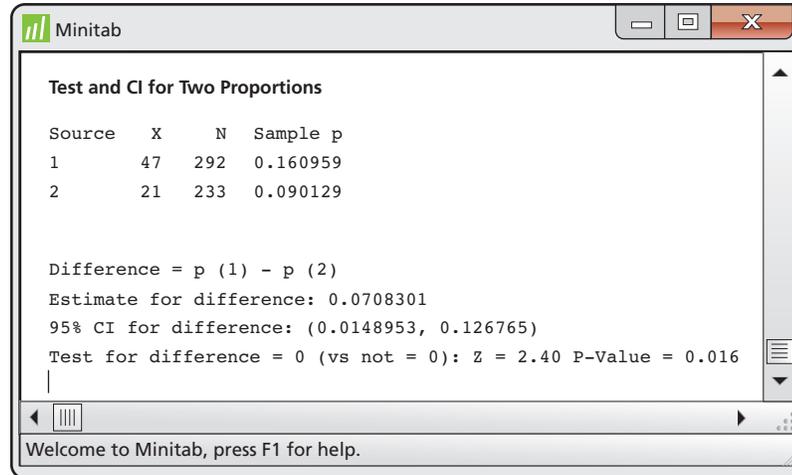
The Facebook users sample of 525 adults contained 292 women and 233 men, with 47 women and 21 men who responded that they increased their use of Facebook over the past year. Using the methods of Chapter 8, we can compare the proportions of male and female Facebook users who increased their use using a confidence interval (page 490) or significance test (page 495).

### EXAMPLE



**14.3 Comparing the proportions of male and female Facebook users who increased their use.** Figure 14.1 contains output for this comparison. The sample proportion for women is 0.160959 (16%), and the sample proportion for men is 0.090129 (9%). The difference is 0.07083, and the 95% confidence interval is (0.015, 0.127). We can summarize this result by saying, “In this sample of Facebook users, the percent who increased their use is 7% higher among women than among men. This difference is statistically significant ( $z = 2.40$ ,  $P = 0.016$ ).”

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



**FIGURE 14.1** Minitab output for the comparison of two proportions (female versus male Facebook users who took a break), for Example 14.3.

**indicator variable** this is with an **indicator variable**. For our problem we will use an indicator of whether or not the adult is a woman:

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

The response variable is the proportion of Facebook users who took a break. For use in a logistic regression, we perform two transformations on this variable. First, we convert to odds. For women,

$$\begin{aligned} \text{odds} &= \frac{\hat{p}}{1 - \hat{p}} \\ &= \frac{0.160959}{1 - 0.160959} \\ &= 0.19184 \end{aligned}$$

Similarly, for men we have

$$\begin{aligned} \text{odds} &= \frac{\hat{p}}{1 - \hat{p}} \\ &= \frac{0.090129}{1 - 0.090129} \\ &= 0.099057 \end{aligned}$$

#### USE YOUR KNOWLEDGE

**14.3 Energy drink commercials.** A study was designed to compare two energy drink commercials. Each participant was shown the commercials, A and B, in random order and asked to select the better one. There were 140 women and 130 men who participated in the study. Commercial A was selected by 65 women and by 67 men. Find the odds of selecting Commercial A for the men. Do the same for the women.

**14.4 Find the odds.** Refer to the previous exercise. Find the odds of selecting Commercial B for the men. Do the same for the women.

## Model for logistic regression

In simple linear regression we modeled the mean  $\mu_y$  of the response variable  $y$  as a linear function of the explanatory variable:  $\mu_y = \beta_0 + \beta_1 x$ . When  $y$  is just 1 or 0 (success or failure), the mean is the probability  $p$  of a success. Logistic regression models the mean  $p$  in terms of an explanatory variable  $x$ . We might try to relate  $p$  and  $x$  as in simple linear regression:  $p = \beta_0 + \beta_1 x$ . Unfortunately, this is not a good model. Whenever  $\beta_1 \neq 0$ , extreme values of  $x$  will give values of  $\beta_0 + \beta_1 x$  that fall outside the range of possible values of  $p$ ,  $0 \leq p \leq 1$ .

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** or **logit** for this transformation. As we did with linear regression, we use  $y$  for the response variable. So for women,

$$y = \log(\text{odds}) = \log(0.19184) = -1.6511$$

and for men,

$$y = \log(\text{odds}) = \log(0.099057) = -2.3121$$

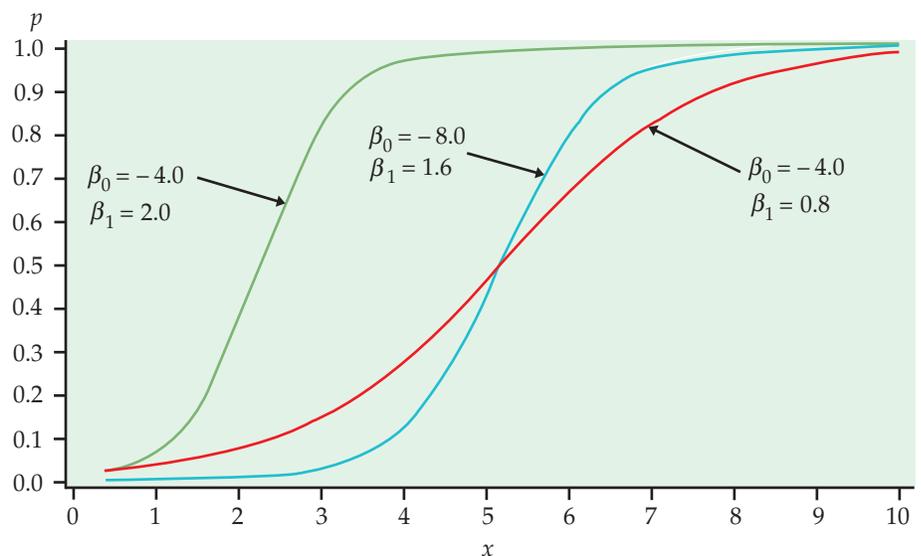
In these expressions for the log odds we use  $y$  as the observed value of the response variable, the log odds of having increased Facebook use. We are now ready to build the logistic regression model.

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 14.2 graphs the relationship between  $p$  and  $x$  for some different values of  $\beta_0$  and  $\beta_1$ . For logistic regression we use *natural* logarithms. There are tables of natural logarithms, and many calculators have a built-in function for this transformation.

**FIGURE 14.2** Plot of  $p$  versus  $x$  for different logistic regression models.



## USE YOUR KNOWLEDGE

**14.5 Find the odds.** Refer to Exercise 14.3. Find the log odds for the men and the log odds for the women.

**14.6 Find the odds.** Refer to Exercise 14.4. Find the log odds for the men and the log odds for the women.

## 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 regression model are  $\beta_0$  and  $\beta_1$ .

## EXAMPLE

**14.4 Model for increased use of Facebook.** For our Facebook use example, there are  $n = 525$  Facebook users in the sample. The explanatory variable is gender, which we have coded using an indicator variable with values  $x = 1$  for women and  $x = 0$  for men. The response variable is also an indicator variable. Thus, the Facebook user either increased his or her use of Facebook or did not increase his or her use. Think of the process of randomly selecting a Facebook user and recording the value of  $y$ , and whether or not the Facebook user increased his or her use. The model says that the probability ( $p$ ) that this user increased his or her use can depend upon the user's gender ( $x = 1$  or  $x = 0$ ). So there are two possible values for  $p$ , say  $p_{\text{women}}$  and  $p_{\text{men}}$ .

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 women,

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

and for men,

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

Note that there is a  $\beta_1$  term in the equation for women because  $x = 1$ , but it is missing in the equation for men 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  $\beta_0$  and  $\beta_1$  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

**14.5 Log odds for increasing Facebook use.** In the Facebook example, we found the log odds for women,

$$y = \log\left(\frac{\hat{p}_{\text{women}}}{1 - \hat{p}_{\text{women}}}\right) = -1.6511$$

and for men,

$$y = \log\left(\frac{\hat{p}_{\text{men}}}{1 - \hat{p}_{\text{men}}}\right) = -2.3121$$

The logistic regression model for women is

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

and for men it is

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

To find the estimates  $b_0$  and  $b_1$ , we match the female and male 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 men:

$$b_0 = -2.1321$$

and the estimate of the slope is the difference between the log odds for the women and the log odds for the men:

$$b_1 = -1.6511 - (-2.3121) = 0.6610$$

The fitted logistic regression model is

$$\log(\text{odds}) = -1.6511 + 0.6610x$$

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

$$\frac{\text{odds}_{\text{women}}}{\text{odds}_{\text{men}}} = e^{0.6610} = 1.94$$

The transformation  $e^{0.6610}$  undoes the logarithm and transforms the logistic regression slope into an **odds ratio**, in this case the ratio of the odds that a woman increases her use of Facebook to the odds that a man increases his use of Facebook. In other words, we can multiply the odds for men by the odds ratio to obtain the odds for women:

$$\text{odds}_{\text{women}} = 1.94 \times \text{odds}_{\text{men}}$$

In this case, we would say that the odds for women are about twice the odds for men.

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 men as 1 and women as 0, the sign of the slope would be reversed, the fitted equation would be  $\log(\text{odds}) = -1.6511 - 0.6610x$ , and the odds ratio would be  $e^{-0.6610} = 0.5163$ . The odds for women are about half 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



**14.6 Is a movie going to be profitable?** The MOVIES data file (described on page 637) includes both the movie's budget and the total U.S. revenue. For this example, we will classify each movie as “profitable” ( $y = 1$ ) if U.S. revenue is larger than the budget and nonprofitable ( $y = 0$ ) otherwise. This is our response variable. The data file contains several explanatory variables, but we will focus here on the natural logarithm of the opening-weekend revenue. Figure 14.3 is a scatterplot of the data with a scatterplot smoother (page 96). The probability that a movie is profitable increases with the log opening-weekend revenue. Because the curve suggested by the smoother is reasonably close to an S-shaped curve like those in Figure 14.2, we fit the logistic regression model

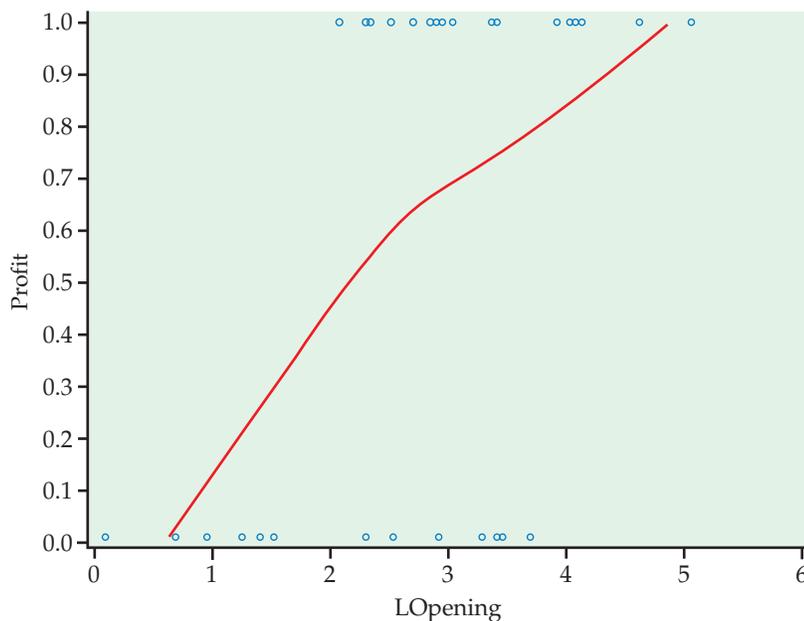
$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x$$

where  $p$  is the probability that the movie is profitable and  $x$  is the log opening-weekend revenue. The model for estimated log odds fitted by software is

$$\log(\text{odds}) = b_0 + b_1 x = -3.1658 + 1.3083x$$

The odds ratio is  $e^{b_1} = 3.700$ . This means that if log opening-weekend revenue  $x$  increases by one unit (roughly \$2.71 million), the odds that the movie will be profitable increase by 3.7 times.

**FIGURE 14.3** Scatterplot of profit (Yes = 1, No = 0) versus the log opening-weekend revenue (LOpening) with a smooth function, for Example 14.6.



**USE YOUR KNOWLEDGE**

- 14.7 Find the logistic regression equation and the odds ratio.** Refer to Exercises 14.3 and 14.5. Find the logistic regression equation and the odds ratio.
- 14.8 Find the logistic regression equation and the odds ratio.** Refer to Exercises 14.4 and 14.6. Find the logistic regression equation and the odds ratio.

## 14.2 Inference for Logistic Regression

When you complete this section, you will be able to

- For a logistic regression with a single explanatory variable, use software output to identify the estimates of the regression parameters and write the equation for the fitted model.
- For a logistic regression with a single explanatory variable, use software output to identify the 95% confidence interval for the regression slope and the significance test results for the null hypothesis that the slope is zero.
- For a logistic regression with a single explanatory variable, use software output to identify the odds ratio and the 95% confidence interval for the odds ratio. Interpret the odds ratio.
- For a logistic regression with several explanatory variables, use software output to identify the estimates of the regression parameters and write the equation for the fitted model.
- For a logistic regression with several explanatory variables, use software output to identify the significance test results for the null hypothesis that all regression slopes are zero.
- For a logistic regression with several explanatory variables, use software output to identify the 95% confidence intervals for the regression coefficients and the significance test results for the null hypothesis that each of the regression coefficients is zero.
- For a logistic regression with several explanatory variables, use software output to identify the odds ratio and the 95% confidence interval for the odds ratio for each explanatory variable. Interpret the odds ratios.

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 of the slope 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 distribution 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**  $\beta_1$  is

$$b_1 \pm z^* SE_{b_1}$$

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**  $e^{\beta_1}$  is obtained by transforming the confidence interval for the slope:

$$(e^{b_1 - z^* SE_{b_1}}, e^{b_1 + z^* SE_{b_1}})$$

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**

$$z = \frac{b_1}{SE_{b_1}}$$

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.

$$X^2 = z^2$$

← **LOOK BACK**  
chi-square statistic,  
p. 538

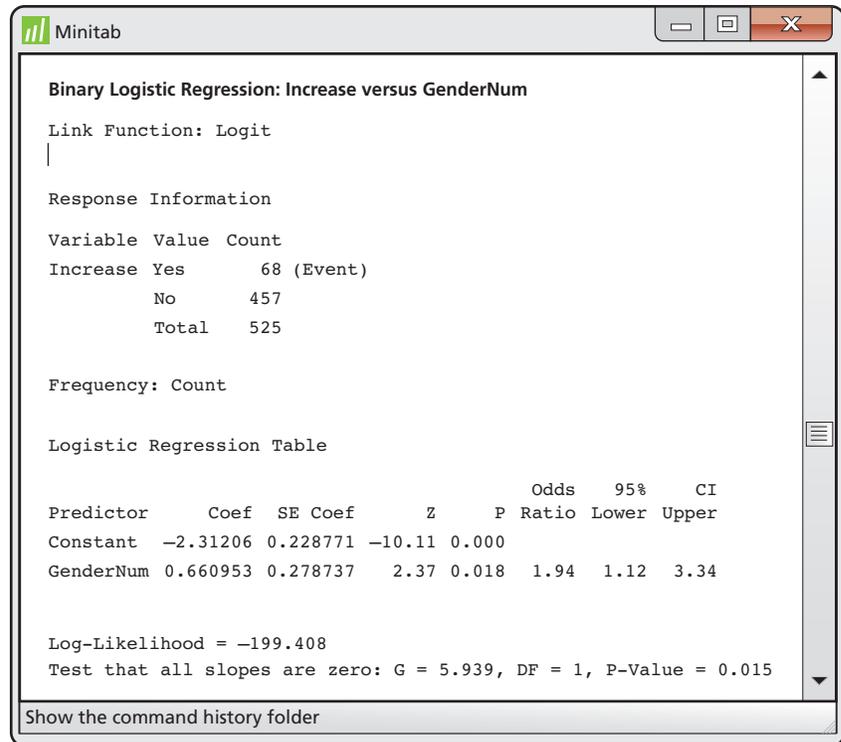
This statistic is called a chi-square statistic. When the null hypothesis is true, it has a distribution that is approximately a  $\chi^2$  distribution with 1 degree of freedom, and the  $P$ -value is calculated as  $P(\chi^2 \geq X^2)$ . Because the square of a standard Normal random variable has a  $\chi^2$  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  $\beta_1$  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.

### EXAMPLE



**14.7 Software output.** Figure 14.4 gives the output from Minitab for the Facebook increased use example described in Example 14.5. The parameter estimates are given as  $b_0 = -2.31206$  and  $b_1 = 0.660953$ . The standard errors are 0.228771 and 0.278737, respectively.



**FIGURE 14.4** Logistic regression output from Minitab for the Facebook increased use data, for Example 14.7.

The 95% confidence interval for the slope is

$$\begin{aligned} b_1 \pm z^*SE_{b_1} &= 0.660953 \pm (1.96)(0.278737) \\ &= 0.660953 \pm 0.546325 \end{aligned}$$

We are 95% confident that the slope is between 0.1146 and 1.2073.

The output also provides the odds ratio 1.94 and a 95% confidence interval, 1.12 to 3.34. For this problem we would report, “Female Facebook users are more likely to increase their use of Facebook than male Facebook users (odds ratio = 1.94, 95% CI = 1.12 to 3.34).”

### USE YOUR KNOWLEDGE



FACETIM



FACETIM

**14.9 Verify the calculation of the odds ratio.** Refer to Example 14.7. Verify that the odds ratio, 1.94, is  $e^{b_1}$ .

**14.10 Verify the calculation of the confidence interval.** Refer to Example 14.7. Verify that the 95% confidence interval for the odds ratio, 1.12 to 3.34, is

$$(e^{b_1 - z^*SE_{b_1}}, e^{b_1 + z^*SE_{b_1}})$$

where  $z^* = 1.96$ . Explain why we use this value of  $z^*$  in the calculation.

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

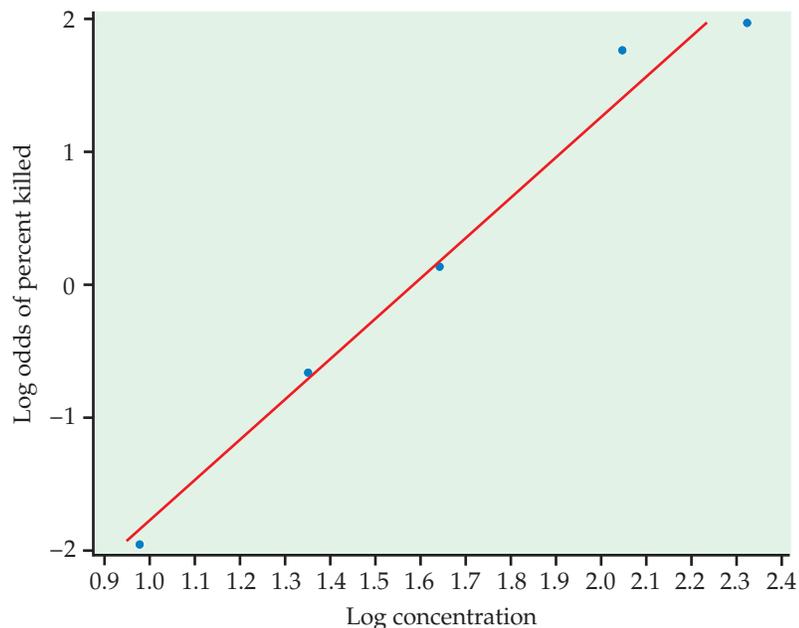


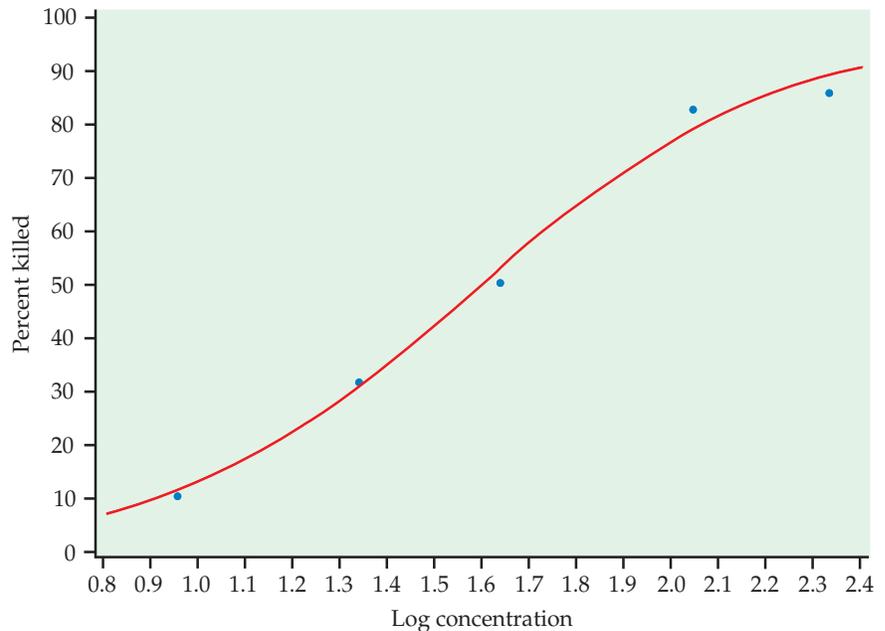
**14.8 An insecticide for aphids.** An experiment was designed to examine how well the insecticide rotenone kills an aphid, called *Macrosiphoniella sanborni*, that feeds on the chrysanthemum plant.<sup>2</sup> The explanatory variable is the concentration (in log of milligrams per liter) of the insecticide. 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

If we transform the response variable (by taking log odds) and use least squares, we get the fit illustrated in Figure 14.5. The logistic regression fit is given in Figure 14.6. It is a transformed version of Figure 14.5 with the fit calculated using the logistic model.

**FIGURE 14.5** Plot of log odds of percent killed versus log concentration for the insecticide data, for Example 14.8.





**FIGURE 14.6** Plot of the percent killed versus log concentration with the logistic fit for the insecticide data, for Example 14.8.

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 14.6, 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

**14.9 Software output.** Figure 14.7 gives the output from Minitab, SPSS, and JMP for the 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, and 2.32. From the output in Minitab and SPSS, we see that the fitted model is

$$\log(\text{odds}) = b_0 + b_1 x = -4.89 + 3.11x$$

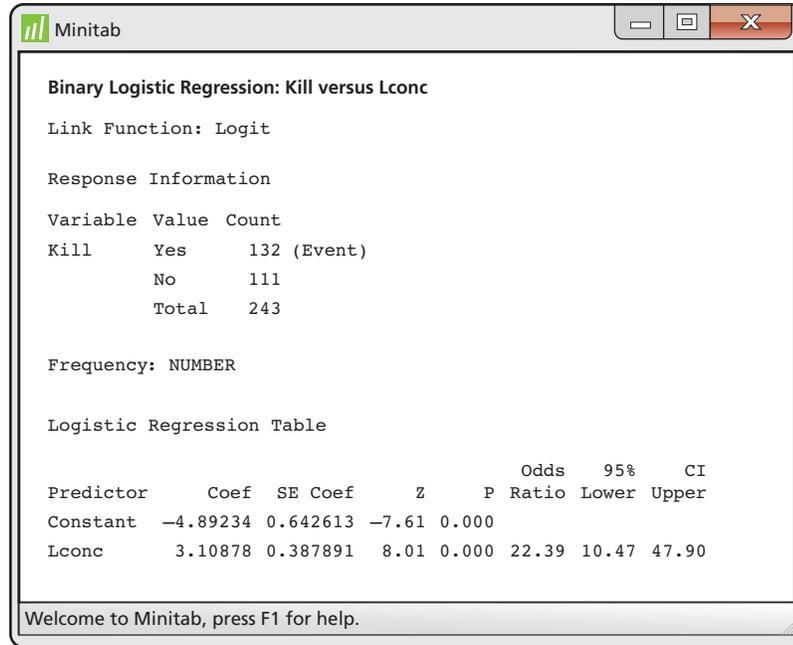
This is the fit that we plotted in Figure 14.6. The null hypothesis that  $\beta_1 = 0$  is clearly rejected ( $z = 8.01$  in Minitab, Wald  $X^2 = 64.233$  in SPSS, and  $X^2 = 64.23$  in JMP;  $P < 0.001$  for all). We calculate a 95% confidence interval for  $\beta_1$  using the estimate  $b_1 = 3.1088$  and its standard error  $SE_{b_1} = 0.3879$  given in the output:

$$\begin{aligned} b_1 \pm z^* SE_{b_1} &= 3.1088 \pm (1.96)(0.3879) \\ &= 3.1088 \pm 0.7603 \end{aligned}$$

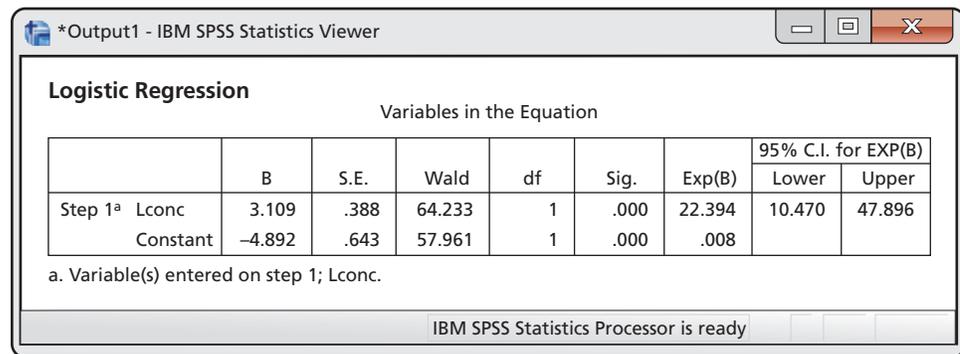
We are 95% confident that the true value of the slope is between 2.35 and 3.87.



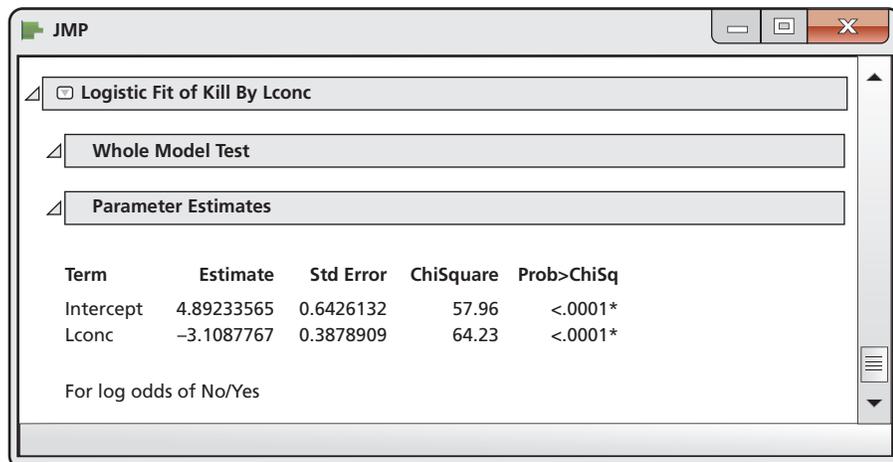
**FIGURE 14.7** Logistic regression output from (a) Minitab, (b) SPSS, and (c) JMP for the insecticide data, for Example 14.9.



(a) Minitab



(b) SPSS



(c) JMP

The odds ratio is given on the Minitab output as 22.39. 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. Minitab gives the 95% confidence interval for the odds ratio, 10.47 to 47.90. We could calculate this from the confidence interval for the slope:

$$\begin{aligned}(e^{b_1 - z^*SE_{b_1}}, e^{b_1 + z^*SE_{b_1}}) &= (e^{2.3485}, e^{3.8691}) \\ &= (10.47, 47.90)\end{aligned}$$

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 22.4 for each unit increase in the log concentration of insecticide ( $X^2 = 64.23$ ,  $P < 0.001$ ; 95% CI = 10.5 to 47.9).”

Note that JMP gives the fitted model as

$$\log(\text{odds}) = 4.89 - 3.11x$$

We see that the regression coefficients  $b_0$  and  $b_1$  are  $-1$  times the coefficients given by Minitab and SPSS. The reason for this is that JMP models the log odds that an insect is *not* killed rather than the log odds that an insect is killed, as shown in the other two outputs. *Always examine software output carefully to be sure that the results you are getting correspond exactly to the analysis that you are trying to perform.* For this analysis, we know from our graph in Figure 14.6 that the slope should be positive.



In Example 14.6 we studied the problem of predicting whether or not a movie was going to make a profit using the log opening weekend revenue as the explanatory variable. We now revisit this example and show how statistical inference is an important part of the conclusion.

### EXAMPLE



**14.10 Software output.** Figure 14.8 gives the output from Minitab for a logistic regression analysis using log opening-weekend revenue as the explanatory variable. From the Minitab output, we see that the fitted model is

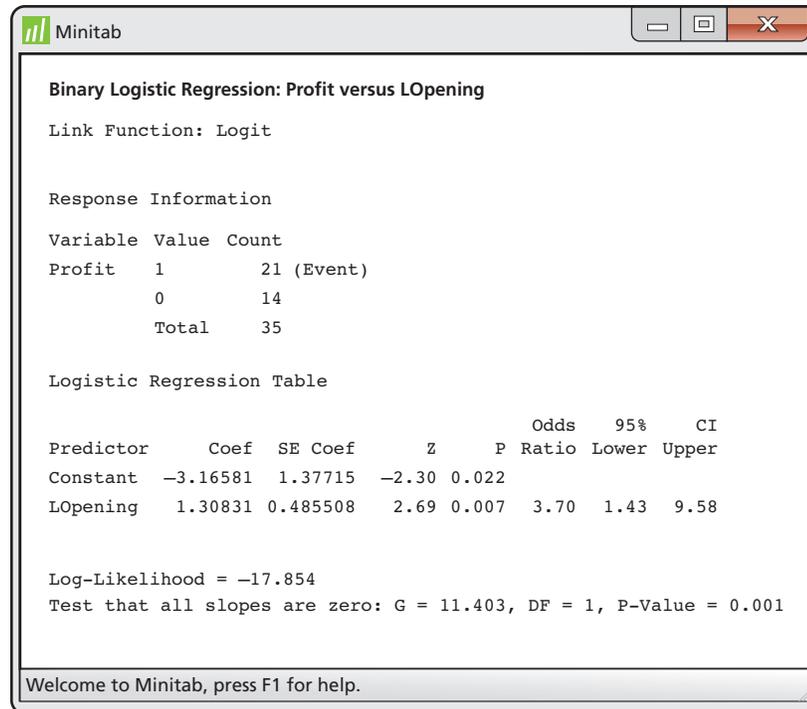
$$\log(\text{odds}) = b_0 + b_1x = -3.166 + 1.3083x$$

From the output, we see that because  $P = 0.007$ , we can reject the null hypothesis that  $\beta_1 = 0$ . The value of the test statistic is  $X^2 = 7.26$  with 1 degree of freedom. We use the estimate  $b_1 = 1.3083$  and its standard error  $SE_{b_1} = 0.4855$  to compute the 95% confidence interval for  $\beta_1$ :

$$\begin{aligned}b_1 \pm z^*SE_{b_1} &= 1.3083 \pm (1.96)(0.4855) \\ &= 1.3083 \pm 0.9516\end{aligned}$$

Our estimate of the slope is 1.3083, and we are 95% confident that the true value is between 0.3567 and 2.2599. For the odds ratio, the estimate on the output is 3.70. The 95% confidence interval is

$$\begin{aligned}(e^{b_1 - z^*SE_{b_1}}, e^{b_1 + z^*SE_{b_1}}) &= (e^{0.3567}, e^{2.2599}) \\ &= (1.43, 9.58)\end{aligned}$$



**FIGURE 14.8** Logistic regression output from Minitab for the movie profitability data with log opening-weekend revenue as the explanatory variable, for Example 14.10.

We estimate that an opening-weekend revenue that is one unit larger (roughly \$2.71 million) will increase the odds that a movie is profitable by about 4 times. The data, however, do not give us a very accurate estimate. The odds ratio could be as small as 1.43 or as large as 9.58 with 95% confidence. We have evidence to conclude that movies with higher opening-weekend revenues are more likely to be profitable, but establishing the relationship accurately would require more data.

Note that the SAS output (not shown), like JMP, gives the same estimates of the regression coefficients but with opposite signs. By default, this software models the odds that the movie is not profitable.

### Multiple logistic regression

The movie example that we just considered naturally leads us to the next topic. The MOVIES data file includes additional explanatory variables. Do these other explanatory variables contain additional information that will give us a better prediction of profitability? 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.

multiple logistic regression



multiple linear regression,  
p. 611

### EXAMPLE



**14.11 Software output.** As in Example 14.10, we predict the odds that a movie is profitable. The explanatory variables are log opening-weekend revenue (LOpening), number of theaters (Theaters), and the movie's IMDb



rating at the end of the first week (Opinion), which is on a 1 to 10 scale (10 being best). Figure 14.9 gives the outputs from SAS, Minitab, and SPSS. The fitted model is

$$\begin{aligned}\log(\text{odds}) &= b_0 + b_1 \text{LOpening} + b_2 \text{Theaters} + b_3 \text{Opinion} \\ &= -2.013 + 2.147 \text{LOpening} - 0.001 \text{Theaters} - 0.109 \text{Opinion}\end{aligned}$$

Note that the coefficients given by SAS have the signs reversed because SAS models the odds that the movie will not be profitable.

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

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

**FIGURE 14.9** Logistic regression output from (a) SAS, (b) Minitab, and (c) SPSS for the movie profitability data with log opening-weekend revenue, number of theaters, and the movie's IMDb rating as the explanatory variables, for Example 14.11.

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	12.7157	3	0.0053
Score	10.9325	3	0.0121
Wald	7.1248	3	0.0680

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	2.0131	3.2320	0.3880	0.5334
LOpening	1	-2.1467	0.9749	4.8488	0.0277
Theaters	1	0.00103	0.000940	1.1924	0.2748
Opinion	1	0.1095	0.4514	0.0589	0.8083

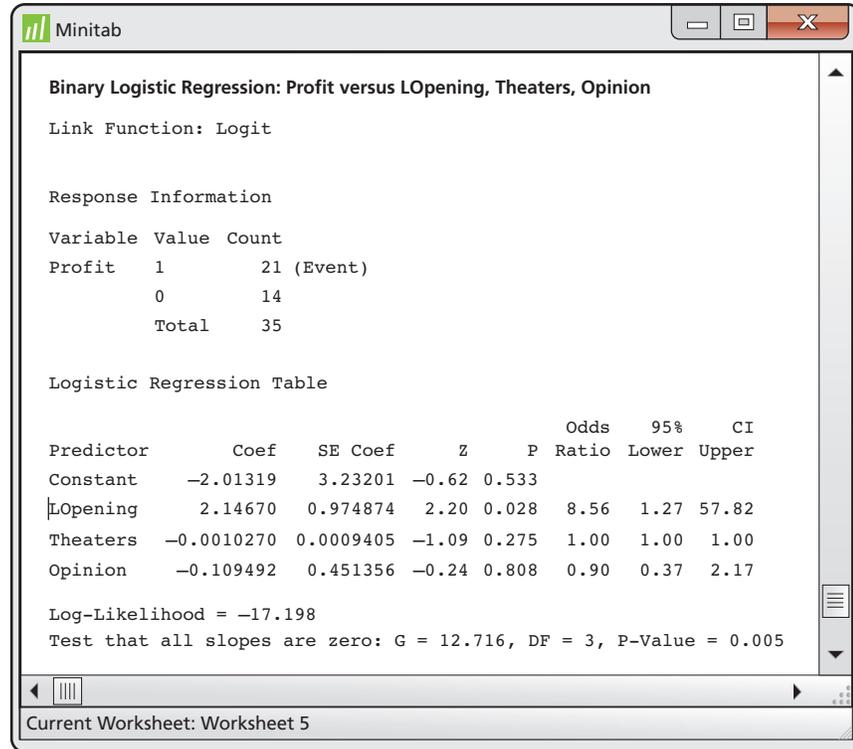
  

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
LOpening	0.117	0.117	0.790
Theaters	1.001	0.999	1.003
Opinion	1.116	0.461	2.702

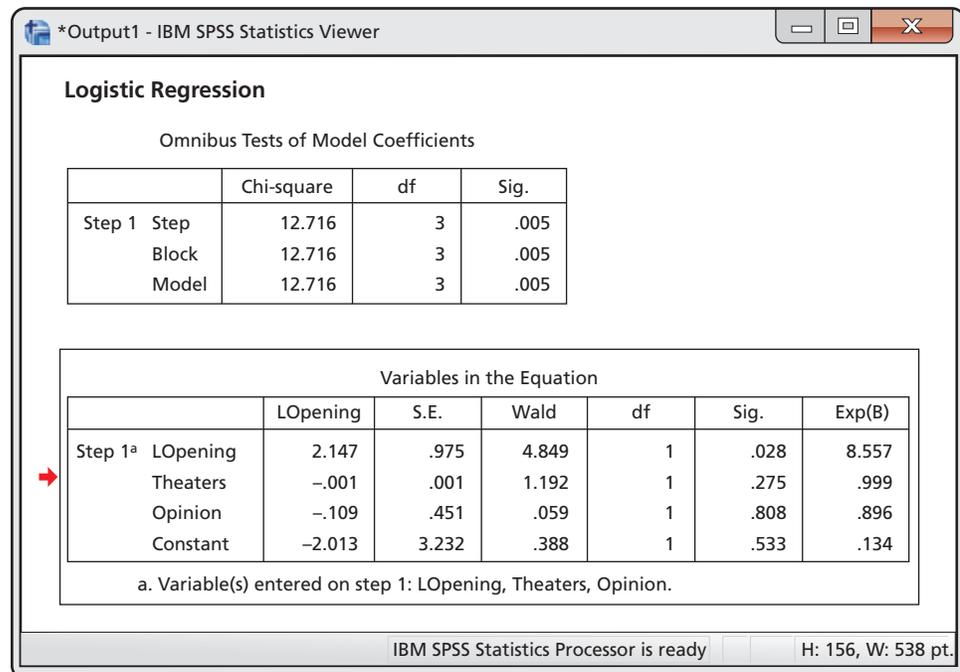
(a) SAS

(continued)

FIGURE 14.9 (continued)



(b) Minitab



(c) SPSS

is tested by a chi-square statistic with 3 degrees of freedom. (The degrees of freedom are 3 because there are 3 coefficients that are set to zero in the null hypothesis.) For Minitab, this is given in the last line of the output, and the statistic is called “G.” The value is  $G = 12.716$  and the  $P$ -value is 0.005. We reject  $H_0$  and conclude that one or more of the explanatory variables can be used to predict the odds that a movie is profitable.

We now examine the coefficients for each variable and the tests that each of these is zero *in a model that contains the other two*. The  $P$ -values are 0.028, 0.275, and 0.808. The null hypotheses  $H_0: \beta_2 = 0$  and  $H_0: \beta_3 = 0$  cannot be rejected. That is, log opening-weekend revenue is the only predictor that adds significant predictive ability once the other two are already in the model.

Our initial multiple logistic regression analysis told us that the explanatory variables contain information that is useful for predicting whether or not the movie is profitable. 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 14 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  $i = 1, 2, \dots, n$  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  $\beta_0$  and  $\beta_1$ .

The **odds ratio** is  $e^{\beta_1}$ , where  $\beta_1$  is the slope in the logistic regression model.

A **level C confidence interval for the intercept**  $\beta_0$  is

$$b_0 \pm z^* SE_{b_0}$$

A **level C confidence interval for the slope**  $\beta_1$  is

$$b_1 \pm z^* SE_{b_1}$$

A **level C confidence interval for the odds ratio**  $e^{\beta_1}$  is obtained by transforming the confidence interval for the slope:

$$(e^{b_1 - z^* SE_{b_1}}, e^{b_1 + z^* SE_{b_1}})$$

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**

$$z = \frac{b_1}{SE_{b_1}}$$

and use the fact that  $z$  has a distribution that is approximately the standard Normal distribution 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  $\chi^2$  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 14 Exercises

For Exercises 14.1 and 14.2, see page 14-3; for Exercises 14.3 and 14.4, see page 14-4; for Exercises 14.5 and 14.6, see page 14-6; for Exercises 14.7 and 14.8, see page 14-9; and for Exercises 14.9 and 14.10, see page 14-11.

**14.11 How did you use your cell phone?** A Pew Internet Poll asked cell phone owners about how they used their cell phones. One question asked whether or not during the past 30 days they had used their phone while in a store to call a friend or family member for advice about a purchase they were considering. The poll surveyed 1003 adults living in the United States by telephone. Of these, 462 responded that they had used their cell phone while in a store within the last 30 days to call a friend or family member for advice about a purchase they were considering.<sup>3</sup>

- What proportion of those surveyed reported that they used their cell phone while in a store within the last 30 days to call a friend or family member for advice about a purchase they were considering?
- Find the odds for the probability that you found in (a).

**14.12 Find some odds.** For each of the following probabilities, find the odds: 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9. Make a plot of the odds versus the probabilities and describe the relationship.

**14.13 A logistic model for cell phones.** Refer to Exercise 14.11. Suppose that you want to investigate differences in cell phone use among customers of different ages. You create an indicator explanatory variable  $x$  that has the value 1 if the customer is 25 years of age or less and is 0 if the customer is over 25 years of age.

- Describe the statistical model for logistic regression in this setting.

- Explain the relationship between the regression coefficients and the odds ratios for the two groups of customers defined by  $x$ .

### 14.14 Another logistic model for cell phones and age.

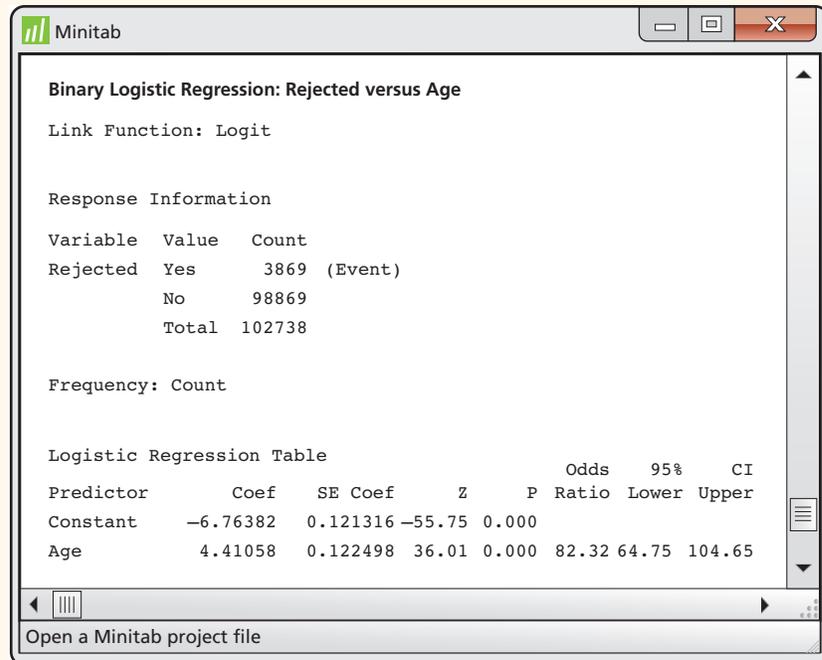
Refer to the previous exercise. Suppose that you use the actual value of age in years as the explanatory variable in a logistic regression model.

- Describe the statistical model for logistic regression in this setting.
- Interpret the regression slope in terms of an effect based on a difference in age of one year.
- This model requires an assumption that is not needed in the model that you described in the previous exercise. Explain the assumption and describe a method for examining whether or not it is a reasonable assumption to make for these data. (*Hint:* Refer to Example 14.8 and Figure 14.5, page 14-12.)

### 14.15 A logistic regression for teeth and military service.

Exercise 8.58 (page 523) describes data on the numbers of U.S. recruits who were rejected for service in a war against Spain because they did not have enough teeth. The exercise compared the rejection rate for recruits who were under the age of 20 with the rate for those who were 40 or over. To run a logistic regression for this setting we define an indicator explanatory variable  $x$  with values of 0 for age under 20 and 1 for age 40 or over. Figure 14.10 gives output from Minitab for this analysis.  **TEETH1**

- How many recruits were examined? How many were rejected and how many were not rejected?
- Write the fitted logistic regression model.



**FIGURE 14.10** Logistic regression output from Minitab for predicting recruit rejection using age in two categories, for Exercises 14.15 to 14.17.

**14.16 Inference for teeth and military service.** Refer to the previous exercise. TEETH1

- Using the information provided in the output in Figure 14.10, calculate and interpret the 95% confidence interval for the regression slope.
- Describe and interpret the results of the significance test for the regression slope. Be sure to give the null and alternative hypotheses, the test statistic, and the  $P$ -value with your conclusion.

**14.17 Odds ratio for teeth and military service.** Refer to the two previous exercises. TEETH1

- Give the odds ratio for this analysis.
- Give the 95% confidence interval for the odds ratio.
- Give a brief description of the meaning of the odds ratio in this analysis.

**14.18 Teeth and military service with six age categories.** In Exercises 14.15 to 14.17 we used logistic regression to study the relationship between being rejected for military service because a recruit did not have enough teeth and age categorized into two groups, under 20 and 40 or over. Data are available for all recruits categorized into six age groups. Let's look at a logistic regression that uses all the data to predict rejection for military service based on teeth. There

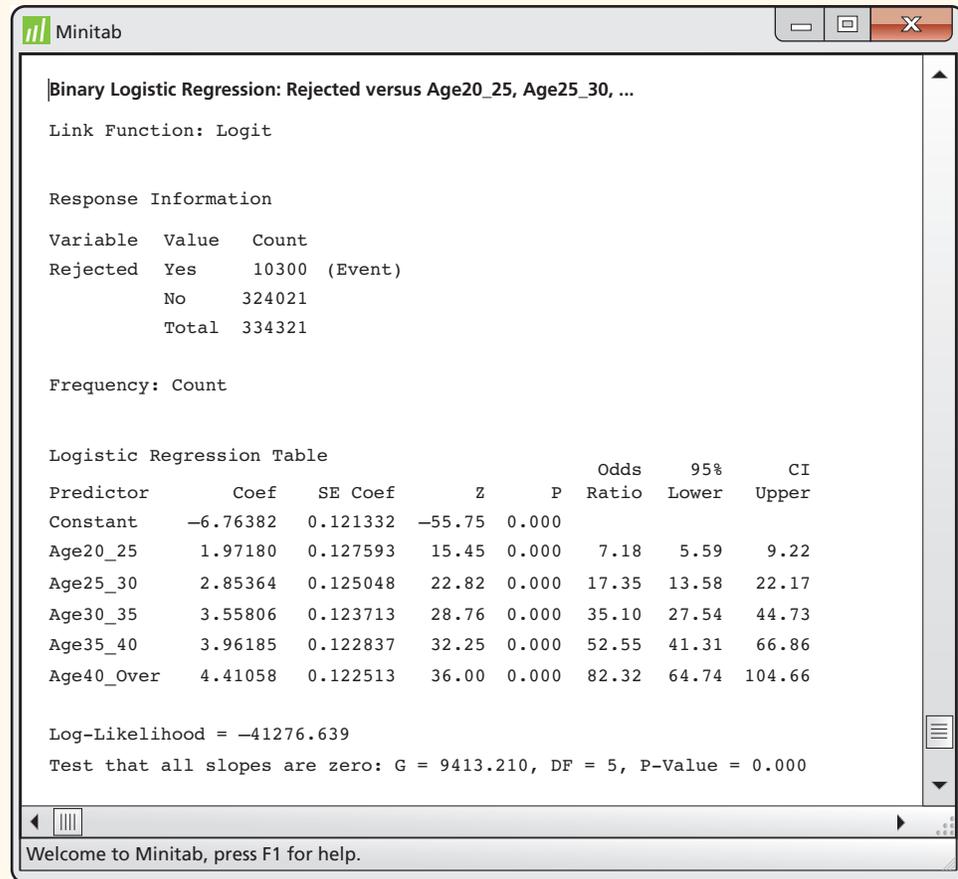
are six age groups: under 20, 20–25, 25–30, 30–35, 35–40, and 40 or over. We define indicator explanatory variables for the last five groups. This is similar to defining a single indicator explanatory variable for an analysis of two groups. TEETH2

Figure 14.11 gives the Minitab output for the logistic regression to predict rejection using the five age indicator explanatory variables.

- Use the output to find the fitted model.
- Is there a pattern in the values of the regression slopes? If yes, describe it.

**14.19 Inference for the multiple logistic regression model.** Refer to the previous exercise. TEETH2

- Describe and interpret the significance test that tests the null hypothesis that all regression coefficients are zero.
- Using the information provided in the output in Figure 14.11, calculate and interpret the 95% confidence interval for each of the regression slopes.
- Describe and interpret the results of the significance test for each regression slope. Be sure to give the null and alternative hypotheses, the test statistic, and the  $P$ -value with your conclusion.



**FIGURE 14.11** Logistic regression output from Minitab for predicting recruit rejection using age in six categories, for Exercises 14.18 to 14.21.

**14.20 Odds ratios for the multiple logistic regression model.** Refer to the two previous exercises. TEETH2

- Give the odds ratio for each explanatory variable.
- Give the 95% confidence interval for each odds ratio.
- Give a brief description of the meaning of each odds ratio in this analysis.

**14.21 Compare the multiple logistic regression analysis with the two-way table.** The data analyzed in Figure 14.11 were analyzed in Exercise 9.22 and Figure 9.7 (page 557) using a  $2 \times 6$  table of counts. Compare these two approaches to the analysis of these data. Describe some strengths and weaknesses of each approach. Which do you prefer? Give reasons for your answer. TEETH2

**14.22 Exergaming in Canada.** Exergames are active video games such as rhythmic dancing games, virtual bicycles, balance board simulators, and virtual sports simulators that require a screen and a console. A study of exergaming by students in grades 10 and 11 in Montreal, Canada, examined many factors related to participation in exergaming.<sup>4</sup> Of the 358 students who reported that they stressed about their health, 29.9% said that they

were exergamers. Of the 851 students who reported that they did not stress about their health, 20.8% said that they were exergamers. Analyze these data using logistic regression and write a summary of your analytical approach, your results, and your conclusions.

**14.23 More exergaming in Canada.** Refer to the previous exercise. Another explanatory variable reported in this study was the amount of television watched per day. Of the 54 students who reported that they watched no TV, 11.1% were exergamers; for the 776 students who watched some TV but less than two hours, 20.6% were exergamers; and for the 370 students who watched two or more hours, 31.1% were exergamers. Use logistic regression to examine the relationship between TV watching and exergaming. Write a summary of your analytical approach, your results, and your conclusions.

**14.24 What's wrong?** For each of the following, explain what is wrong and why.

- If  $b_1 = 5$  in a logistic regression analysis with one explanatory variable, we estimate that the probability of an event is multiplied by 5 when the value of the explanatory variable increases by 1 unit.

(b) The intercept  $\beta_0$  is equal to the odds of an event when  $x = 0$ .

(c) The odds of an event are 1 minus the probability of the event.

**14.25 What's wrong?** For each of the following, explain what is wrong and why.

(a) For a multiple logistic regression with 4 explanatory variables, the null hypothesis that the regression coefficients of all the explanatory variables are zero is tested with an  $F$  test.

(b) For a logistic regression we assume that the model has a Normally distributed error term.

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

(d) In multiple logistic regression we do not need to worry about correlation among explanatory variables when interpreting model coefficient estimates.

 **14.26 Interpret the fitted model.** If we apply the exponential function to the fitted model in Example 14.6 (page 14-8), we get

$$\text{odds} = e^{-11.0391 + 3.1709x} = e^{-11.0391} \times e^{3.1709x}$$

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  $e^{3.1709} = 23.83$ . This justifies the interpretation given at the end of Example 14.6.

**14.27 Will a movie be profitable?** In Example 14.6 (page 14-8), we developed a model to predict whether a movie is profitable based on log opening-weekend revenue. What are the predicted odds of a movie being profitable if the opening-weekend revenue is

(a) \$25 million dollars ( $L_{\text{Opening}} = 3.219$ )?

(b) \$45 million dollars ( $L_{\text{Opening}} = 3.807$ )?

(c) \$65 million dollars ( $L_{\text{Opening}} = 4.174$ )?

**14.28 Converting odds to probability.** Refer to the previous exercise. For each opening-weekend revenue, compute the estimated probability that the movie is profitable.

**14.29 Salt intake and cardiovascular disease.** In Example 9.14 (page 549), the relative risk of developing cardiovascular disease (CVD) for people with low- and high-salt diets was estimated. Let's reanalyze these data using the methods in this chapter. Here are the data:

Developed CVD	Salt in Diet		Total
	Low	High	
Yes	88	112	200
No	1081	1134	2215
Total	1169	1246	2415

(a) For each salt level find the probability of developing CVD.

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

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

**14.30 Salt in the diet and CVD.** Refer to the previous exercise. Use  $x = 1$  for the high-salt diet and  $x = 0$  for the low-salt diet.

(a) Find the estimates  $b_0$  and  $b_1$ .

(b) Give the fitted logistic regression model.

(c) What is the odds ratio for high-salt versus low-salt diet?

(d) When the probability of an event is very small, the odds ratio and relative risk are similar. Compare this odds ratio with the relative risk estimate in Example 9.14. Are they close? Explain your answer.

**14.31 Give a 99% confidence interval for  $\beta_1$ .** Refer to Example 14.9 (page 14-14). Suppose that you wanted to report a 99% confidence interval for  $\beta_1$ . Show how you would use the information provided in the outputs shown in Figure 14.7 to compute this interval.  INSECTS

**14.32 Give a 95% confidence interval for the odds ratio.** Refer to Example 14.9 and the outputs in Figure 14.7 (page 14-14). 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.  INSECTS

 **14.33  $z$  and the  $X^2$  statistic.** The Minitab output in Figure 14.7 (page 14-14) does not give the value of  $X^2$ . The column labeled "Z" provides similar information.  INSECTS

(a) Find the value under the heading "Z" for the predictor  $L_{\text{conc}}$ . 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 close to  $X^2$  (with no roundoff error, these two quantities will be equal). The two-sided  $P$ -value for  $z$  is the same as  $P$  for  $X^2$ .

(c) Draw sketches of the standard Normal distribution 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.

**14.34 Finding the best model?** In Example 14.11 (page 14-17), we looked at a multiple logistic regression for movie profitability based on three explanatory variables. Complete the analysis by looking at the 3 models that include two explanatory variables and the 3 models that include only one variable. Create a table that includes the parameter estimates and their  $P$ -values as well as the overall  $X^2$  statistic and degrees of freedom. Based on the results, which model do you feel is the best? Explain your answer. 

**14.35 Tipping behavior in Canada.** The Consumer Report on Eating Share Trends (CREST) contains data from all provinces of Canada detailing away-from-home food purchases by roughly 4000 households per quarter. Researchers recently restricted their attention to restaurants at which tips would normally be given.<sup>5</sup> From a total of 73,822 observations, “high” and “low” tipping variables were created based on whether the observed tip rate was above 20% or below 10%, respectively. They then used logistic regression to identify explanatory variables associated with either “high” or “low” tips. The following table summarizes what they termed the stereotype-related variables for the low-tip analysis.

Explanatory variable	Odds ratio
Senior adult	1.099
Sunday	1.098
English as second language	1.142
French-speaking Canadian	1.163
Alcoholic drinks	0.713
Lone male	0.858

All coefficients were significant at the 0.01 level. Write a short summary explaining these results in terms of the odds of leaving a low tip.

**14.36 What purchases will be made?** A poll of 1000 adults aged 18 or older asked about purchases they intended to make for the upcoming holiday season.<sup>6</sup> A total of 463 adults listed gift card as a planned purchase.

- What proportion of adults plan to purchase a gift card as a present?
- What are the odds that an adult will purchase a gift card as a present?
- What proportion of adults do not plan to purchase a gift card as a present?
- What are the odds that an adult will not buy a gift card as a present?
- How are your answers to parts (b) and (d) related?

**14.37 High blood pressure and cardiovascular disease.** 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.

- Find the proportion of men who died from cardiovascular disease in the high-blood-pressure group. Then calculate the odds.
- Do the same for the low-blood-pressure group.
- Now calculate the odds ratio with the odds for the high-blood-pressure group in the numerator. Describe the result in words.

**14.38 High blood pressure and cardiovascular disease.** Refer to the study of cardiovascular disease and blood pressure in Exercise 14.37. Computer output for a logistic regression analysis of these data gives the estimated slope  $b_1 = 0.7505$  with standard error  $SE_{b_1} = 0.2578$ .

- Give a 95% confidence interval for the slope.
- 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.
- Write a short summary of your results and conclusions.

**14.39 High blood pressure and cardiovascular disease.** The results describing the relationship between blood pressure and cardiovascular disease are given in terms of the change in log odds in Exercise 14.38.

- Transform the slope to the odds ratio and the 95% confidence interval for the slope to a 95% confidence interval for the odds ratio.
- Write a conclusion using the odds to describe the results.

 **14.40 An example of Simpson’s paradox.** 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		Poor condition		
	Hospital A	Hospital B	Hospital A	Hospital B	
Died	6	8	Died	57	8
Survived	594	592	Survived	1443	192
Total	600	600	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 part (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).

**14.41 Reducing the number of workers.** To be competitive in global markets, many 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:

Terminated	Over 40	
	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?

**14.42 Internet use in Canada.** A recent study used data from the Canadian Internet Use Survey (CIUS) to explore the relationship between certain demographic variables and Internet use by individuals in Canada.<sup>7</sup> The response variable refers to the use of the Internet from any location within the last 12 months. Explanatory variables included age (years), income (thousands of dollars), location (1 = urban, 0 = other), sex (1 = male, 0 = female), education (1 = at least some postsecondary education, 0 = other), language (1 = English, 0 = French), and children (1 = at least one child in household, 0 = no children). The following table summarizes the results.

Explanatory variable	$b$
Age	-0.063
Income	0.013
Location	0.367
Sex	-0.222
Education	1.080
Language	0.285
Children	0.049
Intercept	2.010

All but Children were significant at the 0.05 level.

(a) Interpret the sign of each of the coefficients (except the intercept) in terms of the probability that the individual uses the Internet.

(b) Compute the odds ratio for each of the variables in the table.

(c) What are the odds that a French-speaking, 23-year-old male, living alone in Montreal, and making \$50,000 a year his second year after college is using the Internet?

(d) Convert the odds in part (c) to a probability.

**14.43 Predicting physical activity.** Participation in physical activities typically declines between high school and young adulthood. This suggests that postsecondary institutions may be an ideal setting to address physical activity. A study looked at the association between physical activity and several behavioral and perceptual characteristics among midwestern college students.<sup>8</sup> Of 663 students who met the vigorous activity guidelines for the previous week, 169 reported eating fruit two or more times per day. Of the 471 that did not meet the vigorous activity guidelines in the previous week, 68 reported eating fruit two or more times per day. Model the log odds of vigorous activity using an indicator variable for eating fruit two or more times per day as the explanatory variable. Summarize your findings.

**14.44 Online consumer spending.** The Consumer Behavior Report is designed to provide insight into online shopping trends.<sup>9</sup> A recent report asked the question "In the past three months, how has the current state of the

economy impacted your money spending on online purchasing?” Here are the results from 3156 online consumers:

Gender	Reduced Spending	
	No	Yes
Female	586	708
Male	1074	788

- (a) What proportion of individuals reduced their spending in each gender?
- (b) What is the odds ratio for comparing female individuals to male individuals?
- (c) Write the logistic regression model for this problem using the log odds of reducing spending as the response variable and an indicator of gender as the explanatory variable.
- (d) Software gives the estimated slope  $b_1 = 0.4988$  and its standard error  $SE_{b_1} = 0.0729$ . Transform this result to the odds scale and compare it with your answer in part (b).
- (e) Construct a 95% confidence interval for the odds ratio and write a short conclusion.

**14.45 Proximity of fast-food restaurants to schools and adolescent overweight.** A California study looked at the relationship between fast-food restaurants near schools (within a 0.5-mile radius) and overweight among middle and high school students.<sup>10</sup> Overweight was determined based on each student’s responses to the California Healthy Kids Survey (CHKS). A database of latitude-longitude coordinates for schools and restaurants was used to determine proximity. Here are the data:

Fast-food nearby	$n$	$X$ (overweight)
No	238,215	65,080
Yes	291,152	83,143

Use logistic regression to study the question of whether or not overweight is related to the proximity of fast-food restaurants to schools. Write a short paragraph summarizing your conclusions.

**14.46 Overweight and fast-food restaurants, continued.** Refer to the previous exercise. In the article, the researchers state (1) “CIs were adjusted for clustering at the school level,” and (2) “All models also included controls for the following student characteristics: a female indicator, grade indicator, age indicator, race/ethnicity indicators, and physical exercise indicators. All models also included indicator variables for school location type, including large urban, midsize urban, small urban, large suburban, midsize suburban, small suburban, town, and rural.”

- (a) What violation of the distribution of the response variable is Statement 1 addressing? Explain your answer.

- (b) Explain why the researchers controlled for the variables described in Statement 2 when looking at the relationship between overweight and proximity.

The following four exercises use the GPAHI data file. 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,  $HIGPA$ , to be 1 if the GPA is 3.0 or better and 0 otherwise.  GPAHI

 **14.47 Use high school grades to predict high grade point averages.** Use a logistic regression to predict  $HIGPA$  using the three high school grade summaries as explanatory variables.  GPAHI

- (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).

 **14.48 Use SAT scores to predict high grade point averages.** Use a logistic regression to predict  $HIGPA$  using the SATM and SATCR scores as explanatory variables.  GPAHI

- (a) Summarize the results of the hypothesis test that the coefficients for both explanatory variables are zero.
- (b) Give the coefficient for the SATM score with a 95% confidence interval. Do the same for the SATCR score.
- (c) Summarize your conclusions based on parts (a) and (b).

**14.49 Use high school grades and SAT scores to predict high grade point averages.** 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.  GPAHI

- (a) Test the null hypothesis that the coefficients of the three high school grade summaries are zero; that is, test  $H_0: \beta_{HSM} = \beta_{HSS} = \beta_{HSE} = 0$ .
- (b) Test the null hypothesis that the coefficients of the two SAT scores are zero; that is, test  $H_0: \beta_{SATM} = \beta_{SATCR} = 0$ .
- (c) What do you conclude from the tests in (a) and (b)?

 **14.50 Is there an effect of gender?** In this exercise we investigate the effect of gender on the odds of getting a high GPA.  GPAHI

- (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.

## CHAPTER 14 Notes and Data Sources

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*, 2nd ed., Wiley, 2007; and D. W. Hosmer and S. Lemeshow, *Applied Logistic Regression*, 2nd ed., Wiley, 2000.
2. This example is taken from a classic 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 Finney’s book 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. See [pewinternet.org/Reports/2013/in-store-mobile-commerce.aspx](http://pewinternet.org/Reports/2013/in-store-mobile-commerce.aspx).
4. Erin K. O’Loughlin et al., “Prevalence and correlates of exergaming in youth,” *Pediatrics*, 130 (2012), pp. 806–814.
5. Based on Leigh J. Maynard and Malvern Mupandawana, “Tipping behavior in Canadian restaurants,” *International Journal of Hospitality Management*, 28 (2009), pp. 597–603.
6. These results are from the Consumer Reports National Research Center, which conducted a telephone survey of a nationally representative probability sample of households with telephones. One thousand interviews were completed among adults aged at least 18 years. Interviewing took place on October 15–18, 2009.
7. Anthony A. Noce and Larry McKeown, “A new benchmark for Internet use: A logistic modeling of factors influencing Internet use in Canada, 2005,” *Government Information Quarterly*, 25 (2008), pp. 462–476.
8. Dong-Chul Seo et al., “Relations between physical activity and behavioral and perceptual correlates among midwestern college students,” *Journal of American College Health*, 56 (2007), pp. 187–197.
9. These economic trend reports can be found at [mr.pricegrabber.com](http://mr.pricegrabber.com). These results are based on the June 2009 report.
10. Brennan Davis and Christopher Carpenter, “Proximity of fast-food restaurants to schools and adolescent obesity,” *American Journal of Public Health*, 99 (2009), pp. 505–510.