

# CHAPTER 17



## Logistic Regression

### Introduction

The 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 for use when the response variable has only two possible outcomes. For example,

- *HauteLook.com* is an online destination offering limited-time flash sale events. A response variable of interest to their sales division is whether a member buys or does not buy the daily flash sale item.
- For JP Morgan Chase & Co. recruiting leadership, a response variable of interest is whether a candidate accepts or declines a job offer.

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 with the same  $p$ , this is the *binomial setting*. What is *new* in this chapter is that the data now include at least one *explanatory variable*  $x$  and the probability  $p$  of a success depends on the value of  $x$ . The explanatory variables can either be categorical or quantitative. For example, the probability a customer purchases the flash sale item could depend on the age and gender of the customer, as well as the type of clothing item on sale and the percent discount. The probability a candidate accepts a job offer from JP Morgan Chase & Co. could depend on the salary amount, the level of guaranteed bonuses, and whether or not the offer includes a non-compete clause.

Because it is now a probability that depends on explanatory variables, inference methods are needed to ensure that the probability  $0 \leq p \leq 1$ . Logistic regression is a statistical method for describing these kinds of relationships.<sup>1</sup>

### CHAPTER OUTLINE

- 17.1 The Logistic Regression Model**
- 17.2 Inference for Logistic Regression**
- 17.3 Multiple Logistic Regression**

**REMINDER** →  
binomial setting, p. 245

## 17.1 The Logistic Regression Model

In general, the data for simple logistic regression are  $n$  independent cases, each consisting of a value of the explanatory variable  $x$  and either a success or a failure for that trial. For example,  $x$  may be the salary amount, and “success” means that this applicant accepted the job offer. Every observation may have a different value of  $x$ .

To introduce logistic regression, however, it is convenient to start with the special case in which the explanatory variable  $x$  is also a yes-or-no variable. The data then contain a number of outcomes (success or failure) for each of the two values of  $x$ . There are also just two values of  $p$ , one for each value of  $x$ . Assuming the count of successes for each value of  $x$  has a binomial distribution, we are on familiar ground as described in Chapters 5 and 8. Here is an example.



**Clothing Color and Tipping** What are the factors that affect a customer’s tipping behavior? Studies have shown that a server’s gender and various aspects of a server’s appearance have an effect on tipping, unrelated to the quality of service. Some of these same studies have shown that the effect is different for male and female customers.

Because the color red has been shown to increase the physical attractiveness of women, a group of researchers decided to see if the color of clothing a female server wears has an effect on the tipping behavior.<sup>2</sup> Although they considered both male and female customers, we focus on the 418 male customers in the study.

The response variable is whether or not the male customer left a tip. The explanatory variable is whether the female server wore a red top or not. Let’s express this condition numerically using an indicator variable,

$$x = \begin{cases} 1 & \text{if the server wore a red top} \\ 0 & \text{if the server wore a different colored top} \end{cases}$$

The female servers in the study wore a red top for 69 of the customers and wore a different colored top for the other 349.

The probability that a randomly chosen customer will tip has two values,  $p_1$  for those whose server wore a red top and  $p_0$  for those whose server wore a different colored top. The number of customers who tipped a server wearing red top has the binomial distribution  $B(69, p_1)$ . The number of customers who tipped a server wearing a different color top has the  $B(349, p_0)$  distribution.

### Binomial distributions and odds

We begin with a review of some ideas associated with binomial distributions.

#### EXAMPLE 17.1 Proportion of Tippers



**CASE 17.1** In Chapter 8, we used sample proportions to estimate population proportions. For this study, 40 of the 69 male customers tipped a server who was wearing red and 130 of the 349 customers tipped a server who was wearing a different color. Our estimates of the two population proportions are

$$\text{red: } \hat{p}_1 = \frac{40}{69} = 0.5797$$

and

$$\text{not red: } \hat{p}_0 = \frac{130}{349} = 0.3725$$

That is, we estimate that 58.0% of the male customers will tip if the server wears red, and 37.3% of the male customers will tip if the server wears a different color.

 **REMINDER**  
odds, p. 582

Logistic regression works with **odds** rather than proportions. The odds are the ratio of the proportions for the two possible outcomes. If  $p$  is the probability of a success, then  $1 - p$  is the probability of a failure, and

$$\text{odds} = \frac{p}{1 - p} = \frac{\text{probability of success}}{\text{probability of failure}}$$

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

### EXAMPLE 17.2 Odds of Tipping

**CASE 17.1** The proportion of tippers among male customers who have a server wearing red is  $\hat{p}_1 = 0.5797$ , so the proportion of male customers who are not tippers when their server wears red is

$$1 - \hat{p}_1 = 1 - 0.5797 = 0.4203$$

The estimated odds of a male customer tipping when the server wears red are, therefore,

$$\begin{aligned}\text{odds} &= \frac{\hat{p}_1}{1 - \hat{p}_1} \\ &= \frac{0.5797}{1 - 0.5797} = 1.3793\end{aligned}$$

For the case when the server does not wear red, the odds are

$$\begin{aligned}\text{odds} &= \frac{\hat{p}_0}{1 - \hat{p}_0} \\ &= \frac{0.3725}{1 - 0.3725} = 0.5936\end{aligned}$$

When people speak about odds, they often round to integers or fractions. Because 1.3793 is approximately 7/5, we could say that the odds that a male customer tips when the server wears red are 7 to 5. In a similar way, we could describe the odds that a male customer does *not* tip when the server wears red as 5 to 7.

#### APPLY YOUR KNOWLEDGE

**17.1 Energy drink commercials.** A study was designed to compare Red Bull 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.

**17.2 Use of audio/visual sharing through social media.** In Case 8.3 (page 438), we studied data on large and small food and beverage companies and the use of audio/visual sharing through social media. Here are the data:

Observed numbers of companies			
Size	Use A/V sharing		Total
	Yes	No	
Small	150	28	178
Large	27	25	52
Total	177	53	230

What proportion of the small companies use audio/visual sharing? What proportion of the large companies use audio/visual sharing? Convert each of these proportions to odds.

### Model for logistic regression

In Chapter 8, we learned how to compare the proportions of two groups (such as large and small companies) using  $z$  tests and confidence intervals. Simple logistic regression is another way to make this comparison, but it extends to more general settings with a success-or-failure response variable.

In simple linear regression we modeled the mean  $\mu$  of the response variable  $y$  as a linear function of the explanatory variable:  $\mu = \beta_0 + \beta_1 x$ . When  $y$  is just 1 or 0 (success or failure), the mean is the probability  $p$  of a success. Simple 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 model removes this difficulty by working with the natural logarithm of the odds. We use the term **log odds** or **logit** for this transformation of  $p$ . We model the log odds as a linear function of the explanatory variable:

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

As  $p$  moves from 0 to 1, the log odds move through all negative and positive numerical values. Here is a summary of the logistic regression model.

#### Simple Logistic Regression Model

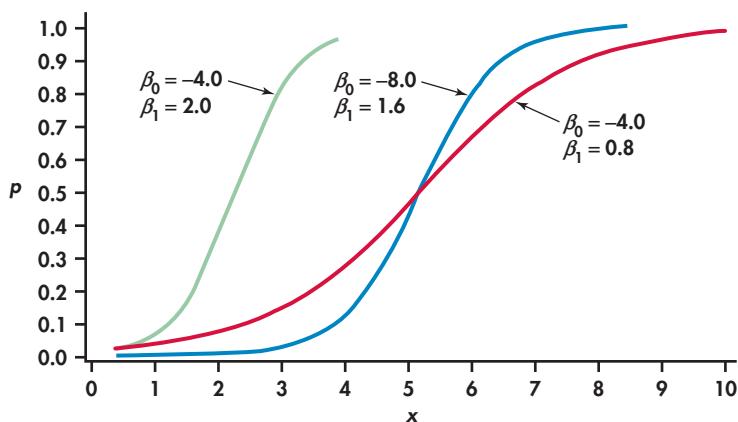
The statistical model for simple logistic regression is

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

where  $p$  is a binomial proportion and  $x$  is the explanatory variable. The parameters of the logistic model are  $\beta_0$  and  $\beta_1$ .

Figure 17.1 graphs the relationship between  $p$  and  $x$  for some different values of  $\beta_0$  and  $\beta_1$ . The logistic regression model uses *natural* logarithms. There are tables of natural logarithms and most calculators have a built-in function for the natural logarithm, often labeled “ln.”

**FIGURE 17.1** Plot of  $p$  versus  $x$  for selected values of  $\beta_0$  and  $\beta_1$ .



Returning to the tipping study, for servers wearing red we have

$$\log(\text{odds}) = \log(1.3793) = 0.3216$$

and for servers not wearing red we have

$$\log(\text{odds}) = \log(0.5936) = -0.5215$$

Verify these results with your calculator, remembering that “log” in these equations is the natural logarithm.

#### APPLY YOUR KNOWLEDGE

**17.3 Log odds choosing Commercial A.** Refer to Exercise 17.1. Find the log odds for the men and the log odds for the women choosing Commercial A.

**17.4 Log odds for use of audio/visual sharing.** Refer to Exercise 17.2. Find the log odds for the small and large companies.

### Fitting and interpreting the logistic regression model

We must now fit the logistic regression model to data. In general, the data consist of  $n$  observations on the explanatory variable  $x$ , each with a success-or-failure response. Our tipping example has an indicator (0 or 1) explanatory variable. Logistic regression with an indicator explanatory variable is a special case but is important in practice. We use this special case to understand a little more about the model.

#### EXAMPLE 17.3 Logistic Model for Tipping Behavior

**CASE 17.1** In the tipping example, there are  $n = 418$  observations. The explanatory variable is whether the server wore a red top or not, which we coded using an indicator variable with values  $x = 1$  for servers wearing red and  $x = 0$  for servers wearing a different color. There are 69 observations with  $x = 1$  and 349 observations with  $x = 0$ . The response variable is also an indicator variable:  $y = 1$  if the customer left a tip and  $y = 0$  if not. The model says that the probability  $p$  of leaving a tip depends on the color of the server’s top ( $x = 1$  or  $x = 0$ ). There are two possible values for  $p$ —say,  $p_1$  for servers wearing red and  $p_0$  for servers wearing a different color. The model says that for servers wearing red

$$\log\left(\frac{p_1}{1 - p_1}\right) = \beta_0 + \beta_1$$

and for servers wearing a different color

$$\log\left(\frac{p_0}{1 - p_0}\right) = \beta_0$$

Note that there is a  $\beta_1$  term in the equation for servers wearing red because  $x = 1$ , but it is missing in the equation for servers wearing a different color because  $x = 0$ .

In general, the calculations needed to find the 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 17.4 Parameter Estimates for Tipping Behavior

**CASE 17.1** For the tipping example, we found the log odds for servers wearing red,

$$\log\left(\frac{\hat{p}_1}{1 - \hat{p}_1}\right) = 0.3216$$

and for servers wearing a different color,

$$\log\left(\frac{\hat{p}_0}{1 - \hat{p}_0}\right) = -0.5215$$

To find estimates  $b_0$  and  $b_1$  of the model parameters  $\beta_0$  and  $\beta_1$ , we match the two model equations in Example 17.3 with the corresponding data equations. Because

$$\log\left(\frac{p_0}{1 - p_0}\right) = \beta_0 \quad \text{and} \quad \log\left(\frac{\hat{p}_0}{1 - \hat{p}_0}\right) = -0.5215$$

the estimate  $b_0$  of the intercept is simply the log(odds) for servers wearing a different color,

$$b_0 = -0.5215$$

Similarly, the estimated slope is the difference between the log(odds) for servers wearing red and the log(odds) for servers wearing a different color,

$$b_1 = 0.3216 - (-0.5215) = 0.8431$$

The fitted logistic regression model is

$$\text{log(odds)} = -0.5215 + 0.8431x$$

The slope in this logistic regression model is the difference between the log(odds) for servers wearing red and the log(odds) for servers wearing a different color. Most people are not comfortable thinking in the log(odds) scale, so interpretation of the results in terms of the regression slope is difficult.

#### EXAMPLE 17.5 Transforming Estimates to the Odds Scale

**CASE 17.1** To get to the odds scale, we take the exponential of the log(odds). Based on the parameter estimates in Example 17.4,

$$\text{odds} = e^{-0.5215 + 0.8431x} = e^{-0.5215} \times e^{0.8431x}$$

From this, the ratio of the odds for a server wearing red ( $x = 1$ ) and for a server wearing a different color ( $x = 0$ ) is

$$\frac{\text{odds}_{\text{red}}}{\text{odds}_{\text{other}}} = e^{0.8431} = 2.324$$

The transformation  $e^{0.8431}$  undoes the natural logarithm and transforms the logistic regression slope into an **odds ratio**, in this case, the comparison of odds that a male customer tips when a server is wearing red to the odds that a male customer tips when a server is wearing a different color. In other words, we can multiply the odds of tipping when a server wears a different color by the odds ratio to obtain the odds of tipping for a server wearing red:

$$\text{odds}_{\text{red}} = 2.324 \times \text{odds}_{\text{other}}$$

In this case, the odds of tipping when a server wears red are about 2.3 times the odds when a server wears a different color.

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 servers wearing a different color as 1 and servers wearing red as 0, the sign of the slope would be reversed, and the fitted equation would be  $\log(\text{odds}) = 0.3216 - 0.8431x$ , and the odds ratio would be  $e^{-0.8431} = 0.430$ . The odds of tipping for servers wearing a different color are roughly 43% of the odds for servers wearing red.

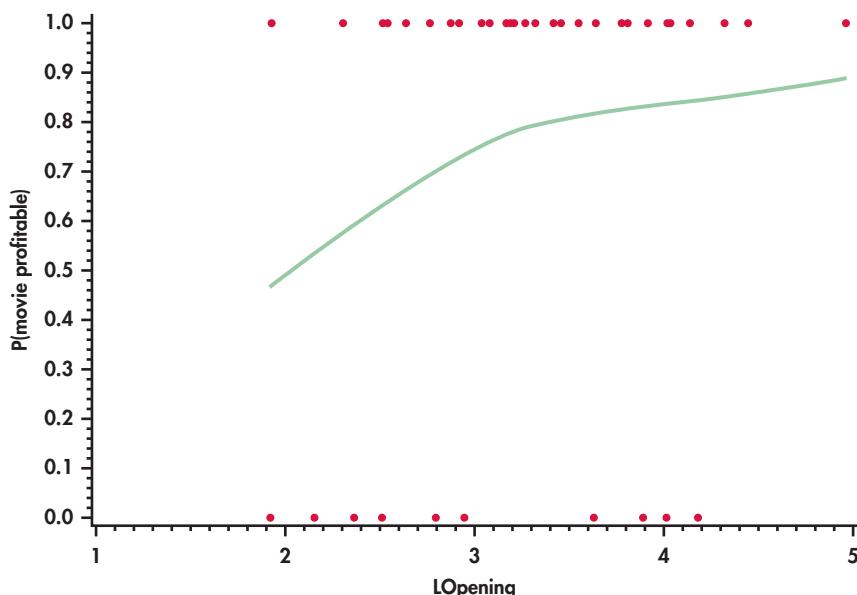
Of course, it is often the case that the explanatory variable is quantitative rather than an indicator variable. We must then use software to fit the logistic regression model. Here is an example.

### EXAMPLE 17.6 Will a Movie Be Profitable?



The MOVIES data set (described on page 550) includes both the movie's budget and the total U.S. revenue. For this example, we 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.

**FIGURE 17.2** Scatterplot of the movie profit data with a scatterplot smoother, Example 17.6. The smoother suggests the upper half of an S-shape similar to those shown in Figure 17.1.



The data set contains several explanatory variables, but we focus here on the natural logarithm of the opening-weekend revenue, LOpening. Figure 17.2 is a scatterplot of the data with a scatterplot smoother (page 68). The probability that a movie is profitable increases with the log opening weekend revenue. Because an S-shaped curve like those in Figure 17.1 is suggested by the smoother, 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

$$\text{log(odds)} = b_0 + b_1 x = -1.41 + 0.781x$$

The estimated odds ratio is  $e^{b_1} = 2.184$ . This means that if opening-weekend revenue  $x$  were roughly  $e^1 = 2.72$  times larger (for example, \$18.1 million to \$49.2 million), the odds that the movie will be profitable increase by 2.2 times.

#### APPLY YOUR KNOWLEDGE

**17.5 Fitted model for energy drink commercials.** Refer to Exercises 17.1 and 17.3. Find the estimates  $b_0$  and  $b_1$  and give the fitted logistic model. What is the estimated odds ratio for a male to choose Commercial A ( $x = 1$ ) versus a female to choose Commercial A ( $x = 0$ )?

**17.6 Fitted model for use of audio/video sharing.** Refer to Exercises 17.2 and 17.4. Find the estimates  $b_0$  and  $b_1$  and give the fitted logistic model. What is the estimated odds ratio for small ( $x = 1$ ) versus large ( $x = 0$ ) companies?

**17.7 Interpreting an odds ratio.** If we apply the exponential function to the fitted model in Example 17.6, we get

$$\text{odds} = e^{-1.41+0.781x} = e^{-1.41} \times e^{0.781x}$$

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  $e^{0.781} = 2.184$ . This justifies the interpretation given at the end of Example 17.6.

#### probit regression

The odds ratio interpretation of the estimated slope parameter is a very attractive feature of the logistic regression model. The health sciences, for example, have used this model extensively to identify risk factors for disease and illness. There are other statistical models, such as **probit regression**, that describe binary responses, but none of them has this interpretation.

#### SECTION 17.1 Summary

- **Logistic regression** explains a success-or-failure response variable in terms of at least one explanatory variable.
- If  $p$  is a proportion of successes, then the **odds** of a success are  $p/(1-p)$ , the ratio of the proportion of successes to the proportion of failures.

- The **simple logistic regression model** relates the proportion of successes in the population to one explanatory variable  $x$  through the logarithm of the odds (or **logit**) of a success:

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

That is, each value of  $x$  gives a different proportion  $p$  of successes. The **data** are  $n$  values of  $x$ , with observed success or failure for each. The model assumes that these  $n$  success-or-failure trials are independent, with probabilities of success given by the logistic regression equation. The **parameters** of the model for one explanatory variable are  $\beta_0$  and  $\beta_1$ .

- Software fits the data to the model, producing estimates  $b_0$  and  $b_1$  of the parameters  $\beta_0$  and  $\beta_1$ .
- The **odds ratio** is the ratio of the odds of a success at  $x + 1$  to the odds of a success at  $x$ . It is found as  $e^{\beta_1}$ , where  $\beta_1$  is the slope in the logistic regression equation.

## 17.2 Inference for Logistic Regression

Statistical inference for logistic regression with one explanatory variable is 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.

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

$$X^2 = z^2$$

 **REMINDER**  
chi-square statistic,  
p. 463

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 one 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 one degree of freedom, the  $z$  statistic and the chi-square statistic give the same results for statistical inference.

### Confidence Intervals and Significance Tests for Logistic Regression

An approximate **level  $C$  confidence interval for the slope  $\beta_1$**  in the logistic regression model is

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

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

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

In these expressions  $z^*$  is the standard Normal critical value with area  $C$  between  $-z^*$  and  $z^*$ .

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

$$X^2 = \left( \frac{b_1}{\text{SE}_{b_1}} \right)^2$$

In terms of a random variable  $\chi^2$  having the  $\chi^2$  distribution with one degree of freedom, the  $P$ -value for a test of  $H_0$  against  $H_a: \beta_1 \neq 0$  is approximately  $P(\chi^2 \geq X^2)$ .

We have expressed the null hypothesis 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 17.7 Computer Output for Tipping Study



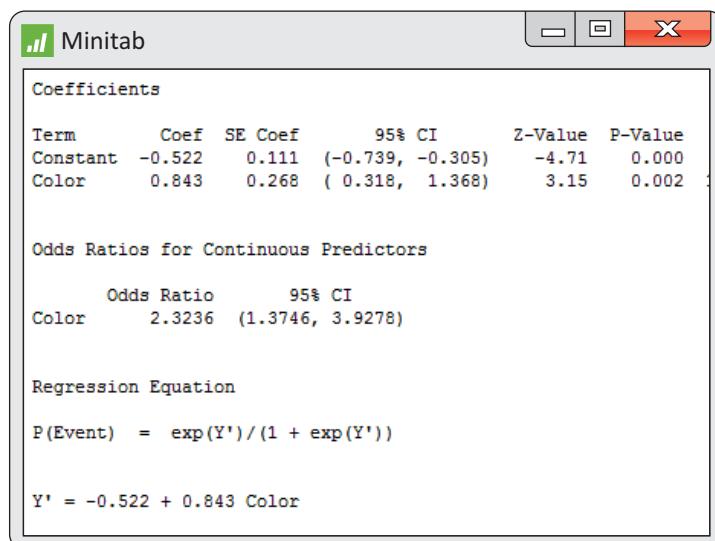
**CASE 17.1** Figure 17.3 gives the output from Minitab and SAS for the tipping study. The parameter estimates match those we calculated in Example 17.4. The standard errors are 0.1107 and 0.2678. A 95% confidence interval for the slope  $\beta_1$  is

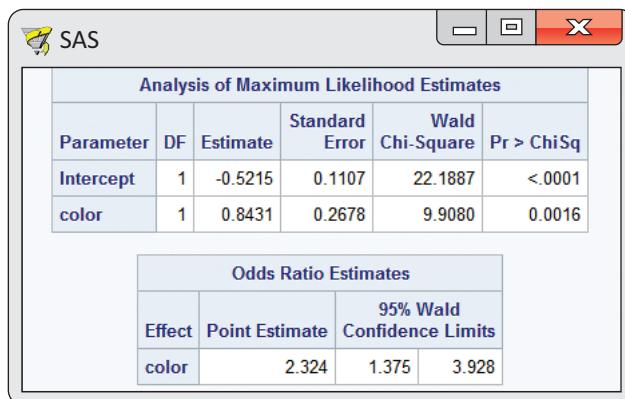
$$\begin{aligned} b_1 \pm z^* \text{SE}_{b_1} &= 0.8431 \pm (1.96)(0.2678) \\ &= 0.8431 \pm 0.5249 \end{aligned}$$

We are 95% confident that the slope is between 0.3182 and 1.368. Both Minitab and SAS output provide the odds ratio estimate and 95% confidence interval. If this interval is not provided, it is easy to compute from the interval for the slope  $\beta_1$ :

$$\begin{aligned} (e^{b_1 - z^* \text{SE}_{b_1}}, e^{b_1 + z^* \text{SE}_{b_1}}) &= (e^{0.3182}, e^{1.368}) \\ &= (1.375, 3.927) \end{aligned}$$

**FIGURE 17.3** Logistic regression output from Minitab and SAS for the tipping data, Example 17.7.



**FIGURE 17.3** (Continued)

We conclude, “Servers wearing red are more likely to be tipped than servers wearing a different color (odds ratio = 2.324, 95% CI = 1.375 to 3.928).”

It is standard to use 95% confidence intervals, and software often reports these intervals. A 95% confidence interval for the odds ratio also provides a test of the null hypothesis that the odds ratio is 1 at the 5% significance level. 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.

#### APPLY YOUR KNOWLEDGE

**CASE 17.1 17.8 Read the output.** Examine the Minitab and SAS output in Figure 17.3. Create a table that reports the estimates of  $\beta_0$  and  $\beta_1$  with the standard errors. Also report the odds ratio with its 95% confidence interval as given in this output.



**17.9 Inference for energy drink commercials.** Use software to run a logistic regression analysis for the energy drink commercial data of Exercise 17.1. Summarize the results of the inference.



**17.10 Inference for audio/visual sharing.** Use software to run the logistic regression analysis for the audio/visual sharing data of Exercise 17.2. Summarize the results of the inference.

#### Examples of logistic regression analyses

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

#### EXAMPLE 17.8 Effectiveness of an Insecticide



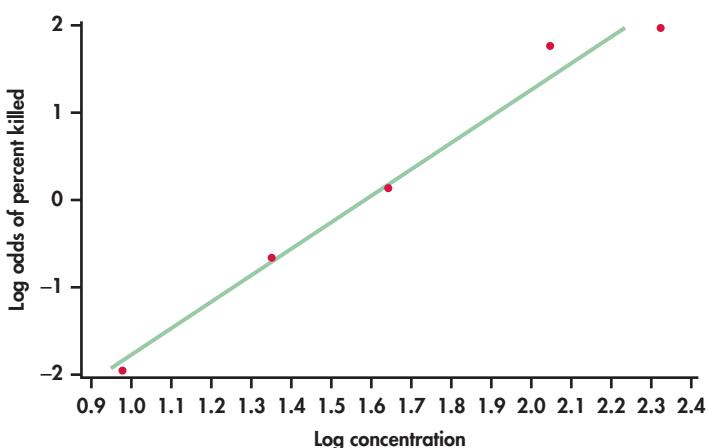
As part of a cost-effectiveness study, a wholesale florist company ran an experiment to examine how well the insecticide rotenone kills an aphid called *Macrosiphoniella sanborni* that feeds on the chrysanthemum plant.<sup>3</sup> The explanatory variable is the concentration (in log of milligrams per liter) of the insecticide. About 50 aphids

each were exposed to one of five concentrations. Each insect was either killed or not killed. Here are the data, along with the results of some calculations:

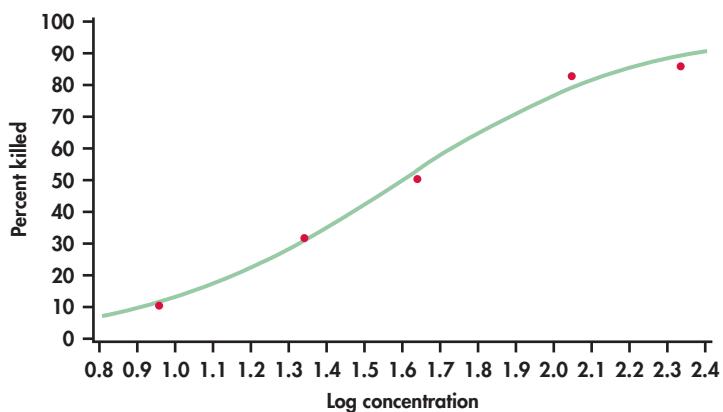
Concentration $x$ (log scale)	Number of insects	Number killed	Proportion killed $\hat{p}$	Log odds
0.96	50	6	0.1200	-1.9924
1.33	48	16	0.3333	-0.6931
1.63	46	24	0.5217	0.0870
2.04	49	42	0.8571	1.7918
2.32	50	44	0.8800	1.9924

Because there are replications at each concentration, we can calculate the proportion killed and estimate the log odds of death at each concentration. The logistic model in this case assumes that the log odds are *linearly* related to log concentration. Least-squares regression of log odds on log concentration gives the fit illustrated in Figure 17.4. There is a clear linear relationship, which justifies our use of the logistic model. The logistic regression fit for the proportion killed appears in Figure 17.5. It is a transformed version of Figure 17.4 with the fit calculated using the logistic model rather than least squares.

**FIGURE 17.4** Plot of log odds of percent killed versus log concentration for the insecticide data, Example 17.8.



**FIGURE 17.5** Plot of the percent killed versus log concentration with the logistic fit for the insecticide data, Example 17.8.



When the explanatory variable has several values, we can often use graphs like those in Figures 17.4 and 17.5 to visually assess whether the logistic regression model seems appropriate. Just as a scatterplot of  $y$  versus  $x$  in simple linear

regression should show a linear pattern, a plot of log odds versus  $x$  in logistic regression should be close to linear. Just as in simple linear regression, outliers in the  $x$  direction should be avoided because they may overly influence the fitted model.

The graphs strongly suggest that insecticide concentration affects the kill rate in a way that fits the logistic regression model. Is the effect statistically significant? 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. As usual, we must add inference to our data analysis.

### EXAMPLE 17.9 Does Concentration Affect the Kill Rate?



**FIGURE 17.6** Logistic regression output from JMP and Minitab for the insecticide data, Example 17.9.

JMP

Logistic Fit of RESPONSE By LCONC

Freq: COUNT

Iterations

Whole Model Test

Parameter Estimates

Term	Estimate	Std Error	ChiSquare	Prob>ChiSq	Odds Ratio
Intercept	-4.8923357	0.6426132	57.96	<.0001*	.
LCONC	3.10877674	0.3878909	64.23	<.0001*	22.3936343

For log odds of KILL/NOKILL

Minitab

Coefficients

Term	Coef	SE Coef	95% CI	Z-Value	P-Value
Constant	-4.892	0.643	(-6.152, -3.633)	-7.61	0.000
LCONC	3.109	0.388	( 2.349, 3.869)	8.01	0.000

Odds Ratios for Continuous Predictors

	Odds Ratio	95% CI
LCONC	22.3936	(10.4701, 47.8957)

Regression Equation

$$P(\text{KILL}) = \exp(Y') / (1 + \exp(Y'))$$

$$Y' = -4.892 + 3.109 \text{ LCONC}$$

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

$$\log(\text{odds}) = b_0 + b_1x = -4.8923 + 3.1088x$$

or

$$\frac{\hat{p}}{1 - \hat{p}} = e^{-4.8923 + 3.1088x}$$

Figure 17.5 is a graph of the fitted  $\hat{p}$  given by this equation against  $x$ , along with the data used to fit the model. JMP gives the statistic  $X^2$  under the heading “ChiSquare.” The null hypothesis that  $\beta_1 = 0$  is clearly rejected ( $X^2 = 64.23$ ,  $P < 0.0001$ ).

The estimated odds ratio is 22.394. 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 given in the Minitab output: (10.470, 47.896).

Remember 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 ratio is 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 ( $X^2 = 64.23$ ,  $P < 0.0001$ ; 95% CI = 10.5 to 47.9).”

#### APPLY YOUR KNOWLEDGE

**17.11 Find the 95% confidence interval for the slope.** Using the information in the output of Figure 17.6, find a 95% confidence interval for  $\beta_1$ .

**17.12 Find the 95% confidence interval for the odds ratio.** Using the estimate  $b_1$  and its standard error in the output of Figure 17.6, find the 95% confidence interval for the odds ratio and verify that this agrees with the interval given by Minitab.

**17.13  $X^2$  or  $z$ .** The Minitab output in Figure 17.6 does not give the value of  $X^2$ . The column labeled “Z-Value” provides similar information.

(a) Find the value under the heading “Z-Value” for the predictor LCONC. Verify that this value 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  $X^2$ .

In Example 17.6, we studied the problem of predicting whether a movie will be profitable using the log opening-weekend revenue as the explanatory variable. We now revisit this example to include the results of inference.

#### EXAMPLE 17.10 Predicting a Movie’s Profitability



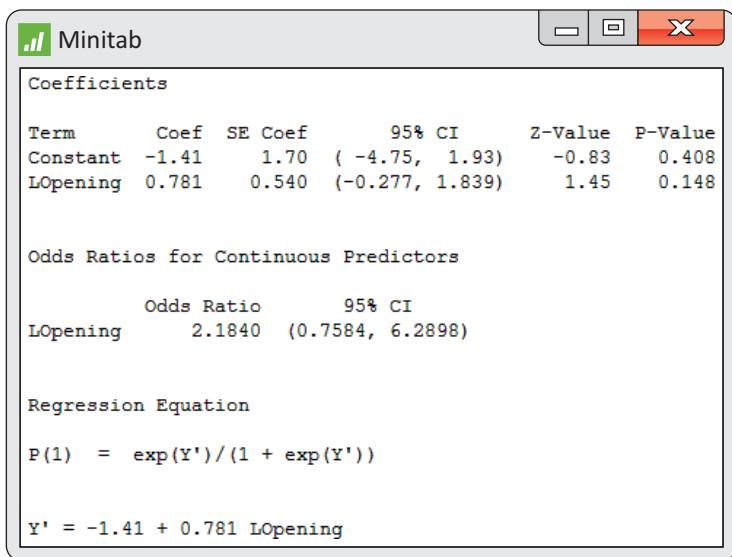
Figure 17.7 gives the output from Minitab for a logistic regression analysis using log opening-weekend revenue as the explanatory variable. The fitted model is

$$\log(\text{odds}) = b_0 + b_1x = -1.41 + 0.781x$$

This agrees up to rounding with the result reported in Example 17.6.

From the output, we see that because  $P = 0.148$ , we cannot reject the null hypothesis that the slope  $\beta_1 = 0$ . The value of the test statistic is  $z = 1.45$ , calculated from the estimate  $b_1 = 0.781$  and its standard error  $SE_{b_1} = 0.540$ . Minitab

**FIGURE 17.7** Logistic regression output from Minitab for the movie profit data with log opening-weekend revenue as the explanatory variable, Example 17.10.



reports the odds ratio as 2.184, with a 95% confidence interval of (0.7584, 6.2898). Notice that this confidence interval contains the value 1, which is another way to assess  $H_0: \beta_1 = 0$ . In this case, we don't have enough evidence to conclude that this explanatory variable, by itself, is helpful in predicting the probability that a movie will be profitable.

We estimate that a one-unit increase in the log opening-weekend revenue will increase the odds that the movie is profitable about 2.2 times. The data, however, do not give us a very accurate estimate. We do not have strong enough evidence to conclude that movies with higher opening-weekend revenues are more likely to be profitable. Establishing the true relationship accurately would require more data.

## SECTION 17.2 Summary

- Software fits the data to the model, producing estimates  $b_0$  and  $b_1$  of the parameters  $\beta_0$  and  $\beta_1$ . Software also produces standard errors for these estimates.
- A level  $C$  confidence interval for the slope  $\beta_1$  is

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

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

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

In these expressions,  $z^*$  is the standard Normal critical value with area  $C$  between  $-z^*$  and  $z^*$ .

- The null hypothesis that  $x$  does not help predict  $p$  in the logistic regression model is  $H_0: \beta_1 = 0$  or  $H_0: e^{\beta_1} = 1$  in terms of the odds ratio. To test this hypothesis, compute the **test statistic**

$$X^2 = \left( \frac{b_1}{\text{SE}_{b_1}} \right)^2$$

In terms of a random variable  $\chi^2$  having a  $\chi^2$  distribution with 1 degree of freedom, the  $P$ -value for a test of  $H_0$  against  $H_a: \beta_1 \neq 0$  is approximately  $P(\chi^2 \geq X^2)$ .

## 17.3 Multiple Logistic Regression

### *multiple logistic regression*

The MOVIES data set includes several explanatory variables. Example 17.10 examines the model where log opening-weekend revenue alone is used to predict the odds that the movie will have a total U.S. box-office revenue greater than the movie budget. Perhaps combining log opening-weekend revenue with other explanatory variables will give us a helpful prediction. We use **multiple logistic regression** to investigate this. 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 analysis.

### EXAMPLE 17.11 Multiple Logistic Regression



As in Example 17.10, we predict the odds that a movie will be profitable. The explanatory variables are log opening-weekend revenue (LOpening), the length of the movie (Minutes), and the movie rating (Rating1). For the movie rating, we use an indicator variable

$$\text{Rating1} = \begin{cases} 1 & \text{if the rating is PG-13 or R} \\ 0 & \text{if the rating is G or PG} \end{cases}$$

Figure 17.8 gives the SAS output. From the output, we see that the fitted model is

$$\begin{aligned} \log(\text{odds}) &= b_0 + b_1 \text{LOpening} + b_2 \text{Minutes} + b_3 \text{Rating1} \\ &= 1.8532 + 1.6019 \text{LOpening} - 0.0607 \text{Minutes} + 1.2225 \text{Rating1} \end{aligned}$$

**FIGURE 17.8** Multiple logistic regression output from SAS for the movie profit data with log opening-weekend revenue, number of theaters, and movie rating as the explanatory variables, Example 17.11.

 A screenshot of a SAS software window titled "SAS". The window displays three tables of statistical results.
 

Testing Global Null Hypothesis: BETA=0				
Test	Chi-Square	DF	Pr > ChiSq	
Likelihood Ratio	10.9757	3	0.0119	
Score	10.1684	3	0.0172	
Wald	7.8555	3	0.0491	

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	1.8532	2.2531	0.6765	0.4108
LOpening	1	1.6019	0.7022	5.2043	0.0225
Minutes	1	-0.0607	0.0232	6.8436	0.0089
Rating1	1	1.2225	0.9833	1.5457	0.2138

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
LOpening	4.962	1.253	19.650
Minutes	0.941	0.899	0.985
Rating1	3.396	0.494	23.328

When analyzing data using multiple regression, we first examine the hypothesis that *all* 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 three degrees of freedom. SAS provides results for three different calculations of this statistic. In all three approaches, the *P*-value is  $< 0.05$ . We reject  $H_0$  and conclude that one or more of the explanatory variables can be used to predict the odds that the movie is profitable.

Next, examine the coefficients for each variable and the tests that each of these is 0 *in a model that contains the other two*. The *P*-values are 0.0225, 0.0089, and 0.2138. The null hypothesis  $H_0: \beta_3 = 0$  cannot be rejected. That is, log opening-weekend revenue and the movie's length add significant predictive ability once the other two explanatory variables are already in the model.

---

Because the explanatory variables are correlated, however, we cannot conclude that log opening-weekend revenue and the movie's length make up the best predictive model. 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.

### SECTION 17.3 Summary

- In **multiple logistic regression** the response variable has two possible values, as in logistic regression, but there can be several explanatory variables.
- As in multiple regression, there is an **overall test** for all the explanatory variables. The null hypothesis that the coefficients for all the explanatory variables are zero is tested by a statistic that has a distribution that is approximately  $\chi^2$  with degrees of freedom equal to the number of explanatory variables. The *P*-value is approximately  $P(\chi^2 \geq X^2)$ .
- Hypotheses about **individual coefficients**,  $H_0: \beta_j = 0$  or  $H_0: e^{\beta_j} = 1$  in terms of the odds ratio, are tested by a statistic that is approximately  $\chi^2$  with 1 degree of freedom. The *P*-value is approximately  $P(\chi^2 \geq X^2)$ . As in multiple regression, these tests assess the contribution of each explanatory variable given the other explanatory variables are already in the model.

## CHAPTER 17 Review Exercises

For Exercises 17.1 and 17.2, see pages 17-3 to 17-4; for 17.3 and 17.4, see page 17-5; for 17.5 to 17.7, see page 17-8; for 17.8 to 17.10, see page 17-11; and for 17.11 to 17.13, see page 17-14.

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

(a) For a multiple logistic regression with four 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 error term in our model has a Normal distribution.  
 (c) In logistic regression with two explanatory variables we use a chi-square statistic to test the null hypothesis  $H_0: b_1 = 0$  versus a two-sided alternative.

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

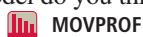
(a) If  $b_1 = 2$  in a logistic regression analysis, we estimate that the probability of an event is multiplied by 2 when the value of the explanatory variable changes by 1.

- (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.

**17.16 Is a movie profitable?** In Example 17.6 (pages 17-7 to 17-8), we developed a model to predict whether a movie will be profitable based on log opening-weekend revenue. What are the predicted odds of a movie being profitable if the opening-weekend revenue is  
 (a) \$20 million dollars?  
 (b) \$35 million dollars?  
 (c) \$50 million dollars?

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

**17.18 Finding the best model?** In Example 17.11 (pages 17-16 to 17-17), we looked at a multiple logistic regression for movie profitability based on three explanatory variables. Complete the analysis by looking at the three models with two explanatory variable models and the three models with single variables. Create a table that includes the parameter estimates and their  $P$ -values as well as the  $X^2$  statistic and degrees of freedom. Based on the results, which model do you think is the best? Explain your answer.



MOVPROF

**17.19 Tipping behavior in Canada.** The Consumer Report on Eating Share Trends (CREST) contains data that cover all provinces of Canada and that describe 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>4</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. Here is a table summarizing what they termed the stereotype-related variables for the high-tip analysis:

Explanatory variable	Odds ratio
Senior adult	0.7420*
Sunday	0.9970
English as second language	0.7360*
French-speaking Canadian	0.7840*
Alcoholic drinks	1.1250*
Lone male	1.0220

The starred odds ratios were significant at the 0.01 level. Write a short summary explaining these results in terms of the odds of leaving a high tip.

**17.20 Sexual imagery in magazine ads.** In what ways do advertisers in magazines use sexual imagery to appeal to youth? One study classified each of 1509 full-page or larger ads as “not sexual” or “sexual,” according to the amount and style of the dress of the male or female model in the ad. The ads were also classified according to the target readership of the magazine.<sup>5</sup> 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	z
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 (general interest). The variable 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  $z$  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.

**17.21 Interpret the results.** Refer to the previous exercise. The researchers also reported odds ratios with

95% confidence intervals for this logistic regression model. Here is a summary:

Explanatory variable	Odds ratio	95% confidence limits	
		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  $z$  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.

**17.22 CEO overconfidence/dominance and corporate acquisitions.** The acquisition literature suggests that takeovers occur either due to conflicts between managers and shareholders or to create a new entity that exceeds the sum of its previously separate components. Other research has offered managerial hubris as a third option, but it has not been studied empirically. Recently, some researchers revisited acquisitions over a 10-year period in the Australian financial system.<sup>6</sup> A measure of CEO overconfidence was based on the CEO's level of media exposure, and a measure of dominance was based on the CEO's remuneration relative to the firm's total assets. They then used logistic regression to see whether CEO overconfidence and dominance were positively related to the probability of at least one acquisition in a year. To help isolate the effects of CEO hubris, the model included explanatory variables of firm characteristics and other potentially important factors in the decision to acquire. The following table summarizes the results for the two key explanatory variables:

Explanatory variable	$b$	SE( $b$ )
Overconfidence	0.0878	0.0402
Dominance	1.5067	0.0057

- (a) State the null and alternative hypotheses for each of the explanatory variables.
- (b) Perform the significance tests and determine whether the variables are significant at the 0.05 level.
- (c) Estimate the odds ratio for each variable and construct a 95% confidence interval.
- (d) Write a short summary explaining the results.

**17.23 E-government use in Canada.** Electronic government (e-government) provides digital means, such as an email address or a website, for citizens to contact public officials. The vision behind e-government is to create a more citizen-focused government. One study used survey data to determine what factors are related to a citizen using an e-government website rather than visiting or calling a government office.<sup>7</sup> The dependent variable refers to whether the citizen used the website or not. Explanatory variables include sex (1 = female, 0 = male), daily Internet use (1 = yes, 0 = no), age (six ordered categories numbered 1 through 6), household income (seven ordered categories numbered 1 through 7), size of the community (six ordered categories numbered 1 through 6), and education (1 = at least some postsecondary education, 0 = other). The following table summarizes the results.

Explanatory variable	Odds ratio
Sex	0.87
Daily Internet use	4.16
Age	0.81
Income	1.01
Size	0.85
Education	0.97
Intercept	0.66

All but "Education" and "Income" were significant at the 0.05 level.

- (a) Interpret each of the odds ratios in terms of the probability that the individual uses the website.
- (b) Compute the regression coefficients for each of the variables in the table.
- (c) What are the odds that a male college graduate, who uses the Internet daily, is age category 3, household Income level 4, and community size 5 is using the Internet?

**17.24 Business Travel.** The Best Western Small Business Travel survey reported that 355 of 400 U.S. small business owners plan as many business trips this fall as last year.<sup>8</sup>

- (a) What proportion of U.S. small business owners plan as many trips as last year?
- (b) What are the odds that an owner will say that his or her company plans as many business trips as last year?
- (c) What proportion of owners said that they do not plan as many trips this year?
- (d) What are the odds that an owner will say that they are cutting back on business trips this year?
- (e) How are your answers to parts (b) and (d) related?

**17.25 Stock options.** 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.<sup>9</sup>

- (a) What proportion of the high-tech companies offer stock options to their key employees? What are the odds?
- (b) What proportion of the non-high-tech companies offer 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. Interpret the result in a few sentences.

### 17.26 Log odds for high-tech and non-high-tech firms.

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  $\beta_0$  and  $\beta_1$ .
- (c) Show that the odds ratio is equal to  $e^{b_1}$ .

### 17.27 Do the inference.

Refer to the previous exercise. Software gives 0.3347 for the standard error of  $b_1$ .

- (a) Find the 95% confidence interval for  $\beta_1$ .
- (b) Transform your interval in (a) to a 95% confidence interval for the odds ratio.
- (c) What do you conclude?

### 17.28 Suppose you had twice as many data.

Refer to Exercises 17.25 through 17.27. 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 17.25 through 17.27.

### 17.29 Poor service.

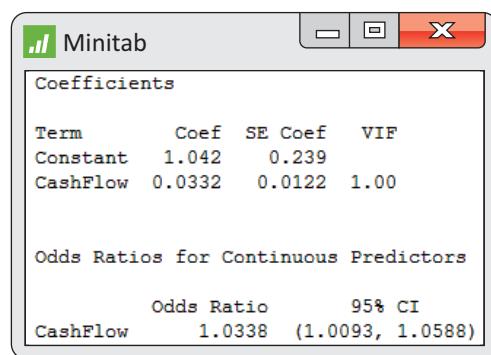
In the food service industry, some argue tipping encourages servers to provide

discriminate service. If the server expects a good tip, he or she may provide better service. In one survey, 193 servers were surveyed and asked if they ever provided poor service because they did not expect a good tip. Ninety-six replied yes.<sup>10</sup>

- (a) What proportion of the servers have provided poor service because of an expected bad tip?
- (b) What are the odds that a server will have provided bad service given an expected bad tip?
- (c) What proportion of the servers did not provide bad service?
- (d) What are the odds that a server will not have provided bad service?
- (e) How are your answers to parts (b) and (d) related?

### 17.30 Active retail companies versus failed companies.

Case 7.2 (page 389) compared the cash flow of 74 active retail firms with the cash flow for 27 firms that failed. Here we analyze the same data with a logistic regression. The outcome is whether or not the firm is active, and the explanatory variable is the cash flow. Here is the output from Minitab:



- (a) Give the fitted equation for the log odds that a firm will be active.
- (b) Describe the results of the significance test for the coefficient of cash flow.
- (c) The odds ratio is the estimated amount that the odds of being active would increase when the cash flow is increased by one unit. Report this odds ratio with the 95% confidence interval.
- (d) Write a short summary of this analysis and compare it with the analysis of these data that we performed in Chapter 7. Which approach do you prefer?

### 17.31 Analysis of a reduction in force.

To meet competition or cope with economic slowdowns, corporations sometimes undertake a “reduction in force” (RIF), in which 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:

		<b>Over 40</b>	
<b>Terminated</b>	No	Yes	
Yes	17	71	
No	564	835	

- (a) Write the logistic regression model for this problem using the log odds of a termination 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  $b_1 = 1.0371$  and its standard error  $SE_{b_1} = 0.2755$ . 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?

### 17.32 Following brands through social media.

PricewaterhouseCoopers (PwC) surveyed 1000 online shoppers in the United States and China.<sup>11</sup> One question asked if the online shopper followed brands they purchased through social media. Here are the results:

		<b>Social media</b>	
<b>Country</b>	No	Yes	
United States	487	513	
China	72	928	

- (a) What are the proportions of online shoppers who follow brands through social media in each country?
- (b) What is the odds ratio for comparing U.S. online shoppers with Chinese online shoppers?
- (c) Write the logistic regression model for this problem using the log odds of following brands through social media as the response variable and country as an indicator explanatory variable (U.S. = 1).
- (d) Software gives the estimated slope  $b_1 = -2.5043$  and its standard error  $SE_{b_1} = 0.1377$ . 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.

**17.33 Know your customers.** 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.<sup>12</sup> 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.

### 17.34 Does income relate to use of the Internet?

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?

*For the following five exercises, you will need to construct indicator variables to use categorical variables as explanatory variables in logistic regression. Be sure to review the material in Chapter 11 on models with categorical explanatory variables (pages 571–575) before attempting these exercises.*

### 17.35 Reduction in force using logistic regression.

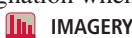
In Exercise 17.31, hypothetical data are given for a reduction in force (RIF). If there is a statistically significant difference in the RIF proportions based on age group, the employer needs to justify the difference based on other (nondiscriminatory) variables.  RIF

- (a) Run the logistic analysis to predict the odds of being riffed using age group (over 40 years of age or not) as the explanatory variable. Summarize your results.
- (b) What other variables would you add to the model in an attempt to explain the results that you described in part (a)? If these other variables can be shown to be characteristics that relate to job performance, and the age effect is no longer significant in a model that includes these variables, then the analysis provides statistical evidence that can be used to refute a claim of discrimination.

**17.36 Sexual imagery in ads.** Refer to Exercise 17.20 (page 17-18) concerning the use of sexual imagery in magazine ads. Here is the two-way table of counts for the 1509 ads.

Model dress	Magazine readership			Total
	Women	Men	General interest	
Not sexual	351	514	248	1113
Sexual	225	105	66	396
Total	576	619	314	1509

Use the model dress, expressed as the odds that the dress is sexual, as the response variable and the magazine readership as the explanatory variable. Because there are three magazine readership categories, you will need two indicator variables for this multiple logistic regression analysis. Use the last category, general interest, for the “other” designation when creating these indicator variables.



- (a) A friend has suggested that the three magazine categories be coded as 1, 2, 3 and that this single variable be used as the explanatory variable in the logistic regression. Explain why this analytical strategy is wrong.
- (b) Summarize the results of the significance testing. Do the data support the idea that the sexual content expressed in the model dress varies by the magazine readership?
- (c) Use the estimates for your model and the coding that you used for the explanatory variables to give the estimated log odds for each type of magazine readership.

### 17.37 Rerun the analysis with a different coding.

In the previous exercise, you used the last category, general interest, for the “other” designation when you constructed the indicator variables. Now use the women’s magazine readership as the “other” category and reanalyze the data. Verify that the significance testing results for the effect of the two explanatory variables is the same as in the previous exercise.



**17.38 Student athletes and gambling.** A survey of student athletes that asked questions about gambling behavior classified students according to the National

Collegiate Athletic Association (NCAA) division.<sup>13</sup> For male student athletes, the percent who reported wagering on collegiate sports are given here along with the numbers of respondents in each division:

	Division		
	I	II	III
Percent	17.2%	21.0%	24.4%
Number	5619	2957	4089

- (a) Using the numbers and percents given, calculate the numbers of students who gamble and those who do not for each NCAA division.
- (b) Use two indicator variables to code the explanatory variable, NCAA division. Let the first one be 1 for Division II and 0 otherwise; let the second be 1 for Division III and 0 otherwise. With this coding, the logistic regression model will use the intercept for Division I, the intercept plus the coefficient of the first indicator variable for Division II, and the intercept plus the coefficient of the second indicator variable for Division III.
- (c) Run the multiple logistic regression and summarize the results.

**17.39 Is there a trend?** Refer to the previous exercise. The coding of the indicator variables suggests a way to code models when you expect a pattern in the response that is based on some kind of ordering of the explanatory variable. In some settings this is called detecting a dose response.

- (a) Use the model to give the estimated log odds for each NCAA division.
- (b) Plot these estimates versus division and summarize the results. Does there appear to be a pattern in the results?
- (c) How would you model the pattern that you described in part (b)?

## 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*, 3rd ed., Wiley, 2013.
2. Nicolas Guéguen and Céline Jacob, “Clothing color and tipping: Gentlemen patrons give more tips to waitresses with red clothes,” *Journal of Hospitality & Tourism Research* 38, no. 2 (2014), pp. 275–280.
3. This example is taken from a classical text written by a contemporary of R. A. Fisher. (Fisher 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. Also, although we have chosen to call the response “killed,” 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. Nevertheless, an insect that is “unable to walk more than a few steps” is unlikely to eat very much of a chrysanthemum plant!

4. Based on Leigh J. Maynard and Malvern Mupandawana, “Tipping behavior in Canadian restaurants,” *International Journal of Hospitality Management* 28 (2009), pp. 597–603.
5. Tom Reichert, “The prevalence of sexual imagery in ads targeted to young adults,” *Journal of Consumer Affairs* 37 (2003), pp. 403–412.
6. Results from Rayna Brown and Neal Sarma, “CEO overconfidence, CEO dominance and corporate acquisitions,” *Journal of Economics and Business* 59 (2007), pp. 358–379.
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. The press release for this survey can be found at the Best Western website, [www.bestwestern.com/about-us/press-media/press-release-details.asp?NewsID=910](http://www.bestwestern.com/about-us/press-media/press-release-details.asp?NewsID=910).
9. Based on Greg Clinch, “Employee compensation and firms’ research and development activity,” *Journal of Accounting Research* 29 (1991), pp. 59–78.
10. Michael Lynn and Shou Wang, “The indirect effects of tipping policies on patronage intentions through perceived expensiveness, fairness, and quality,” *Journal of Economic Psychology* 39 (2013), pp. 62–71.
11. This result can be found at [www.pwc.com/gx/en/retail-consumer/retail-consumer-publications/global-multi-channel-consumer-survey/explore-the-data.jhtml](http://www.pwc.com/gx/en/retail-consumer/retail-consumer-publications/global-multi-channel-consumer-survey/explore-the-data.jhtml).
12. From 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.
13. Based on information in “NCAA 2003 national study of collegiate sports wagering and associated health risks,” which can be found at the NCAA website, [www.ncaa.org](http://www.ncaa.org).

## ANSWERS TO ODD-NUMBERED EXERCISES

**17.1** For men: the percent who chose Commercial A is 0.5154; the percent who chose Commercial B is 0.4846. The odds are 1.0636. For women: the percent who chose Commercial A is 0.4643; the percent who chose Commercial B is 0.5357. The odds are 0.8667.

**17.3** For men: 0.06166. For women: -0.14306.

**17.5**  $b_0 = -0.14306$ ,  $b_1 = 0.20472$ .  $\log(\text{odds}) = -0.14306 + 0.20472x$ . The odds ratio is 1.227.

$$\begin{aligned} \text{17.7} \quad \frac{\text{odds}_{x+1}}{\text{odds}_x} &= \frac{e^{-1.41+0.781(x+1)}}{e^{-1.41+0.781x}} \\ &= \frac{e^{0.781x}e^{0.781}}{e^{0.781x}} = e^{0.781} = 2.184. \end{aligned}$$

**17.9**  $\log(\text{odds}) = -0.14306 + 0.20472x$ . The odds ratio estimate is 1.227; the 95% confidence interval is (0.761, 1.979).

**17.11** (2.349, 3.869).

**17.13** (a)  $Z = 8.01$ . (b) 64.16, which agrees with the output up to rounding error.

**17.15** (a) It is not multiplied by 2. When the explanatory variable changes by 1, the odds are increased by a factor of  $e^2$  or 7.389 times. (b) It is missing the log; the intercept is equal to the *log* odds of an event when  $x = 0$ . (c) The odds of an event are the probability of the event divided by 1 minus the probability of the event.

**17.17** (a) 0.7170. (b) 0.7968. (c) 0.8383.

**17.19** Those who order alcoholic drinks are 12.5% more likely (or 1.125 times as likely) to leave a high tip than those who don’t order alcohol. Senior adults are about 25.8% less likely (or 0.742 times as likely) to leave a high tip than those who aren’t senior. Those who speak English as a second

language are about 26.4% less likely (or 0.736 times as likely) to leave a high tip than their counterparts. Those who are French-speaking Canadians are about 21.6% less likely (or 0.784 times as likely) to leave a high tip than those who aren't French-speaking Canadians.

- 17.21** (a) If the confidence interval for the odds ratio includes the value 1, the variable is not significant in a logistic regression. (b) Because the Reader age, Model sex, and Women's magazines intervals all do not contain 1, they are all significant. The Men's magazine interval contains 1 and is not significant. (c) Interpreting only significant effects: When the reader age is mature adults, the model clothing is 1.27 to 2.16 times more likely to be not sexual. When the model sex is male, the model clothing is 2.74 to 5.01 times more likely to be not sexual. When the intended readership is women, the model clothing is 1.11 to 2.23 times more likely to be not sexual. The odds ratios are often much easier to interpret than the fitted coefficients.

- 17.23** (a) Females are 0.87 times as likely (13% less likely) to use the website as males. Daily Internet users are 4.16 times as likely to use the website as their counterparts. Older-aged people are less likely to use the website than younger-aged people. Those from larger communities are less likely to use the website than those from smaller communities. Those with different incomes and/or educations are about equally likely to use the website because they aren't significantly different from 1. (b) Sex: -0.1393, Daily Internet use: 1.4255, Age: -0.2107, Income: 0.01, Size: -0.1625, Education: -0.0305, Intercept: -0.4155. (c) 0.6537.

- 17.25** (a) 0.8022. odds = 4.0556. (b) 0.6881. odds = 2.2059. (c) odds ratio = 1.8385. The high-tech companies are 1.8385 times more likely to offer incentive stock options to key employees than the non-high-tech companies.

- 17.27** (a) (-0.047, 1.265). (b) (0.954, 3.543). (c) Because the interval in part (b) includes 1, there

is no significant difference in the proportions of high-tech and non-high-tech companies that offer stock options to key employees.

- 17.29** (a) 0.4974. (b) odds = 0.9897. (c) 0.5026. (d) odds = 1.0104. (e) They are reciprocals.

- 17.31** (a)  $\log(\text{odds}) = -3.5017 + 1.0369x$ . (b) The binomial distribution assumes that each employee's termination is independent from one another's and the probability of being terminated is the same for each employee. Certainly the latter is not true because an individual's performance is likely different and largely determines whether or not they are terminated. (c) odds = 2.82, with 95% the confidence interval is (1.644, 4.840). Because the interval does not contain 1, the results are significant at the 5% level. Employees over 40 are 2.82 times more likely to be terminated than those under 40. (d) We could use the additional variables in the logistic regression model to account for their effects before assessing if age has an effect.

- 17.33**  $\log(\text{odds}) = -0.0282 + 0.6393x$ .  $X^2 = 48.34$ ,  $P\text{-value} < 0.0001$ . The odds ratio estimate is 1.8952; that is, those who have completed college are 1.8952 times more likely to use the Internet for travel arrangements than those who have not completed college.

- 17.35** (a)  $\log(\text{odds}) = -3.5017 + 1.0369x$ .  $X^2 = 14.17$ ,  $P\text{-value} = 0.0002$ . The model is significant. For a person over 40: 0.085. For a person under 40: 0.030. (b) Answers will vary.

- 17.37** The estimated model becomes  $\log(\text{odds}) = -0.4447 - 1.1436x_{\text{men}} - 0.8791x_{\text{general}}$ . Now both men ( $X^2 = 69.7000$ ,  $P\text{-value} < 0.0001$ ) and general ( $X^2 = 29.1872$ ,  $P\text{-value} < 0.0001$ ) terms are significant.

- 17.39** (a) For Division I: -1.5720. For Division II: -1.3248. For Division III: -1.1304. (b) The plot shows that  $\log(\text{odds})$  of gambling increases as Division increases. (c) Because the relationship is quite linear, we could use a regression analysis.