Chapter 15

case study

Introduction

Section 15.1 Prediction and Inference for Logistic Regression

Logistic Regression

The Risks of Space Exploration

On a chilly Florida morning in January 1986, Space Shuttle *Challenger* was ready for liftoff. In spite of engineers' concerns about the cold weather, NASA managers authorized the launch. *Challenger* exploded shortly after liftoff, and all seven crew members were killed. Following the disaster, scientists and statisticians helped analyze what went wrong. They determined that the failure of O-rings that secure joints in the shuttle's booster rockets was to blame. Could this catastrophe have been prevented by a closer look at the data?

The table below shows the launch temperature (°F) and whether or not any O-rings at the joints failed on each of the 24 previous shuttle flights.[1] Of the 23 flights for which data were available, there were 7 instances of O-ring failure. Do you notice anything special about these flights?

Flight	Temp. (°F)	Failure?	Flight	Temp. (°F)	Failure?	Flight	Temp. (°F)	Failure?
STS-1	66	No	STS-9	70	No	STS 51-B	75	No
STS-2	70	Yes	STS 41-B	57	Yes	STS 51-G	70	No
STS-3	69	No	STS 41-C	63	Yes	STS 51-F	81	No
STS-4	80	Unknown	STS 41-D	70	Yes	STS 51-I	76	No
STS-5	68	No	STS 41-G	78	No	STS 51-J	79	No
STS-6	67	No	STS 51-A	67	No	STS 61-A	75	Yes
STS-7	72	No	STS 51-C	53	Yes	STS 61-B	76	No
STS-8	73	No	STS 51-D	67	No	STS 61-C	58	Yes

Figure 15.1 shows a modified scatterplot of the data. The graph plots the quantitative explanatory variable "launch temperature" on the horizontal axis. The categorical response variable "O-ring failure?" is converted to a numerical value (1 = Yes, 0 = No) and plotted on the vertical axis. From the graph, we can see that most of the O-ring failures occurred on shuttle flights that launched in cooler temperatures.

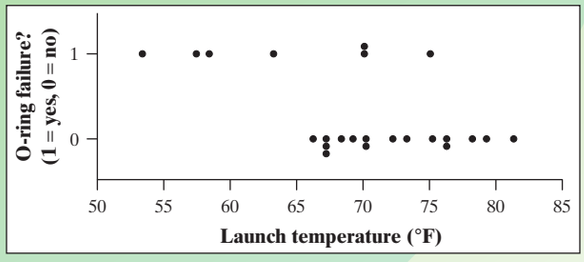


FIGURE 15.1 A modified scatterplot of the data on launch temperature and O-ring failure for 23 space shuttle flights.

The temperature when *Challenger* launched was 31°F. Could NASA scientists have used these data to predict the probability of O-ring failure and call off the launch? By the end of this chapter, you will have developed the tools to help answer this question.

Introduction

The simple and multiple linear regression methods we studied in Chapters 12 and 14 are used to model the relationship between a quantitative response variable and one or more explanatory variables. What if we want to predict the values of a *categorical* variable? In this chapter, we describe methods that are used when the response variable has only two possible values.

- Can a child's body mass index (BMI) at age 4 help predict whether she or he will be obese at age 12?
- To what extent does gender predict whether or not a college student will be a binge drinker?
- How does the concentration of an insecticide relate to whether or not an insect is killed?

Our response variable has only two values: yes or no, success or failure. If we let the two values be 1 (success) and 0 (failure), the mean is the proportion of ones, $p = P(\text{success})$. With n independent observations, we have a *binomial setting* (page 383). What's new here is that we have data on an explanatory variable x. We study how p depends on x.

For example, suppose that we are studying whether a patient lives ($y = 1$) or dies ($y = 0$) after being admitted to a hospital. Here, p is the probability that a patient lives, and possible explanatory variables include (a) whether the patient is in good condition or in poor condition, (b) the type of medical problem that the patient has, and (c) the age of the patient. Note that the explanatory variables can be either categorical or quantitative. *Logistic regression* is a statistical method for describing these kinds of relationships.[2] This chapter describes the basic ideas of logistic regression.

15.1

Prediction and Inference for Logistic Regression

In Section 15.1, you'll learn about:

- Binomial distributions and odds
- The logistic regression model
- Inference for logistic regression
- Putting it all together: Logistic regression
- Multiple logistic regression

To develop the logistic regression model, we start with settings involving one categorical explanatory variable. This makes prediction and inference for logistic regression simpler.

Binomial Distributions and Odds

In Chapter 6, we studied binomial distributions. In Chapters 8 and 9, we learned how to do statistical inference for the proportion p of successes in a binomial setting. We will need some of these ideas in this chapter, and so we start with a brief review.

EXAMPLE

College Students and Binge Drinking

A binomial setting?

Alcohol abuse has been described by college presidents as the number one problem on campus, and it is a significant cause of death in young adults. How common is it? A survey of 10,904 randomly selected U.S. college students collected information on drinking behavior and alcohol-related problems.[3] Researchers defined "frequent binge drinking" as having five or more drinks in a row three or more times in the past two weeks. The researchers wanted to estimate the proportion of students who are frequent binge drinkers. According to this definition, 2486 students were classified as frequent binge drinkers.

In the language of Chapter 6, the chance process here consists of randomly selecting a U.S. college student and determining whether or not she or he is a binge drinker. To confirm that this is a binomial setting, we check the BINS.

- **Binary?** "Success" = frequent binge drinker. "Failure" = not a frequent binge drinker.
- **Independent?** Because we are sampling without replacement, we check the *10% condition*. There are definitely at least $10(10{,}904) = 109{,}040$ U.S. college students.
- **Number?** There are $n = 10{,}904$ trials.
- **Success?** Suppose that p is the actual proportion of frequent binge drinkers in the entire population of U.S. college students. Then the probability of getting a "success" on each trial is p.

Define X = the number of frequent binge drinkers in the sample. Since X counts the number of successes, it is a binomial random variable with $n = 10{,}904$ and unknown p. Based on the researchers' random sample, our point estimate of p is

$$\hat{p} = \frac{2486}{10{,}904} = 0.228.$$

Odds

Logistic regression works 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 other outcome. The resulting odds are

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

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

EXAMPLE

College Students and Binge Drinking

Finding the odds

PROBLEM: Refer to the previous example. Find the odds that a student is a binge drinker.

SOLUTION: The proportion of frequent binge drinkers in the sample is $\hat{p} = 0.228$, so the proportion of students who are not frequent binge drinkers is

$$1 - \hat{p} = 1 - 0.228 = 0.772$$

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

$$\text{odds} = \frac{\hat{p}}{1-\hat{p}} = \frac{0.228}{0.772} = 0.295$$

For Practice *Try Exercise 1*

When people speak about odds, they often round to integers or fractions. Because 0.295 is close to 1/3, we could say that the odds that a college student is a frequent binge drinker are about 1 to 3. In a similar way, we could describe the odds that a college student is *not* a frequent binge drinker as about 3 to 1.

Odds for Two Samples Using the methods of Section 10.1, we could compare the proportions of frequent binge drinkers among men and women college students using a confidence interval. The researchers found that the sample proportion for the men was $\hat{p}_M = 0.252$ (25.2%) and that the sample proportion for the women was $\hat{p}_W = 0.209$ (20.9%). The difference is $\hat{p}_M - \hat{p}_W = 0.043$. We can summarize this result by saying. "The proportion of frequent binge drinkers in the sample is 4.3 percentage points higher among men than among women." A 95% confidence interval for $p_M - p_W$ is (0.026, 0.060). This interval suggests that the proportion of frequent binge drinkers is between 2.6 and 6.0 percentage points higher among male than among female U.S. college students.

Another way to analyze these data is to use logistic regression. The explanatory variable is gender, a categorical variable. To use this variable in a regression model (logistic or otherwise), we need to use a numeric code. As we saw in Chapter 14, the usual way to do this is with an *indicator variable*. For the binge-drinking example we will use an indicator of whether or not the student is a man:

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

The response variable indicates whether the person is a frequent binge drinker. As before, we let p represent the actual proportion of frequent binge drinkers. For use in a logistic regression, we begin by converting to odds. For men,

$$\text{odds} = \frac{\hat{p}}{1-\hat{p}} = \frac{0.252}{1-0.252} = 0.337$$

Similarly, for women we have

$$\text{odds} = \frac{\hat{p}}{1-\hat{p}} = \frac{0.209}{1-0.209} = 0.264$$

As you'll see shortly, we still need to perform one more transformation before we can do logistic regression.

CHECK YOUR UNDERSTANDING

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 100 women and 140 men who participated in the study. Commercial A was selected by 45 women and by 80 men.

1. Find the odds of selecting Commercial A for the men.
2. Find the odds of selecting Commercial A for the women.

The Logistic Regression Model

In simple linear regression, we modeled the mean μ_y of the response variable y as a linear function of the explanatory variable x: $\mu_y = \beta_0 + \beta_1 x$. For logistic regression, the response variable is categorical, taking only values 0 (failure) or 1 (success). The population mean response at any x-value is just the true proportion p of successes. We are interested in finding a model to predict the probability p of a successful outcome using the explanatory variable x.

It might help to consider a specific example. Suppose that we want to predict the probability p that a child will be obese at age 12 based on his or her body mass index at age 4. For each possible value of x = BMI, the logistic regression model should give the true proportion p of 12-year-olds who are obese. We could try to relate p and x through the equation $p = \beta_0 + \beta_1 x$. Unfortunately, this is not a good model. As long as $\beta_1 \neq 0$, extreme values of x will give values of $\beta_0 + \beta_1 x$ that are inconsistent with the fact that $0 \leq p \leq 1$. For the childhood obesity example, data from one study involving 1042 randomly selected 4-year-olds gives the equation[4]

$$\text{predicted probability} = -1 + 0.0769x$$

Figure 15.2 is a scatterplot of the actual data from the study (0 = not obese, 1 = obese) with this least-squares regression line added. Based on this model, a child with BMI $x = 27$ at age 4 would have predicted probability $-1 + 0.0769(27) = 1.076$ of being obese at age 12!

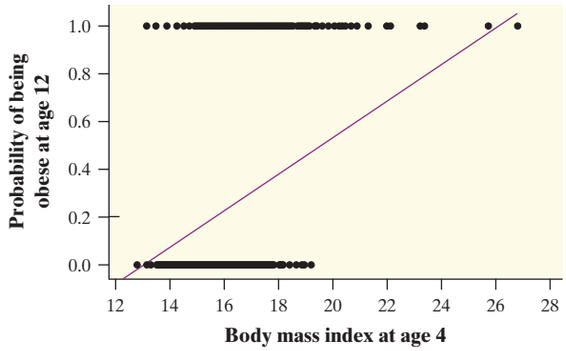


FIGURE 15.2 Scatterplot and least-squares regression line using the data from the childhood obesity study. For 4-year-old children with BMIs less than 13 or greater than 26, the linear model gives unreasonable predictions for their probability of being obese at age 12.

The logistic regression solution to this difficulty is to transform the odds $\frac{p}{1-p}$ using the natural logarithm ln. We use the term **log odds** for this transformation. Instead of modeling p, we model the log odds as a linear function of the explanatory variable:

Log odds

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

Of course, our goal is to predict the probability p of a successful outcome for specific values of x. Using algebra, we can solve the above equation for p:

$$\frac{p}{1-p} = e^{\beta_0+\beta_1 x} \qquad \text{Using the definition of "logarithm"}$$

$$p = e^{\beta_0+\beta_1 x} - pe^{\beta_0+\beta_1 x} \qquad \text{Multiplying both sides of the equation by } (1-p)$$

$$p + pe^{\beta_0+\beta_1 x} = e^{\beta_0+\beta_1 x} \qquad \text{Adding } pe^{\beta_0+\beta_1 x} \text{ to both sides}$$

$$p(1 + e^{\beta_0+\beta_1 x}) = e^{\beta_0+\beta_1 x} \qquad \text{Factoring out } p \text{ on the left side}$$

$$p = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}} \qquad \text{Dividing both sides by } 1 + e^{\beta_0+\beta_1 x}$$

Figure 15.3 graphs the logistic regression model for some different values of β_0 and β_1. In each case, the relationship between p and x displays a clear S-curve shape. Notice that $0 \le p \le 1$ for all values of β_0 and β_1.

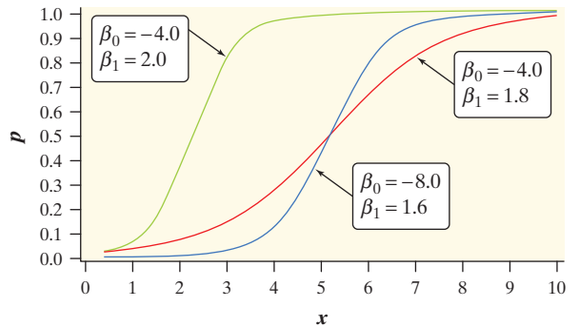


FIGURE 15.3 Plot of p versus x for different logistic regression models. Note the clear S-curve shape!

We are now ready to build the logistic regression model.

Logistic Regression Model

The **logistic regression model** relating the proportion p of successful outcomes in a binomial setting to the explanatory variable x is given by

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

The parameters of the logistic model are β_0 and β_1.

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. As the following example shows, because there are only two values for x, we can write both equations.

EXAMPLE

College Students and Binge Drinking

The logistic regression model

The explanatory variable in the binge-drinking study is gender, which we have coded using an indicator variable with values $x = 1$ for men and $x = 0$ for women. Think of the process of randomly selecting a student and recording the values of x and whether or not the student is a frequent binge drinker. The logistic regression model says that the probability (p) that this student is a frequent binge drinker depends upon the student's gender ($x = 1$ or $x = 0$). So there are two possible values for p: p_{men} and p_{women}. For men, $x = 1$ and

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

For women, $x = 0$ and

$$\ln\left(\frac{p_{\text{women}}}{1 - p_{\text{women}}}\right) = \beta_0 + \beta_1(0) = \beta_0$$

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

EXAMPLE

College Students and Binge Drinking

Interpreting the model

In the gender and binge-drinking study, the log odds for men is

$$\ln\left(\frac{\hat{p}_{\text{men}}}{1 - \hat{p}_{\text{men}}}\right) = \ln\left(\frac{0.252}{1 - 0.252}\right) = \ln(0.337) = -1.09$$

and the log odds for women is

$$\ln\left(\frac{\hat{p}_{\text{women}}}{1 - \hat{p}_{\text{women}}}\right) = \ln\left(\frac{0.209}{1 - 0.209}\right) = \ln(0.264) = -1.33$$

The logistic regression model for men is

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

and for women it is

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

To find the estimates of β_0 and β_1, we match the male and female model equations with the corresponding equations based on the data. We see that the estimate of the intercept β_0 is simply the log odds for the women:

$$b_0=-1.33$$

The estimate of the slope β_1 is the difference between the log odds for the men and the log odds for the women:

$$b_1=-1.09-(-1.33)=0.24$$

So the fitted logistic regression model is

$$\widehat{\ln(\text{odds})}=-1.33+0.24x$$

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 scale of the log odds, so interpretation of the results in terms of the regression slope is difficult. Usually, we apply a transformation to help us:

$$b_1=0.24=\ln(\text{odds}_{\text{men}})-\ln(\text{odds}_{\text{women}})$$

$$0.24=\ln\left(\frac{\text{odds}_{\text{men}}}{\text{odds}_{\text{women}}}\right) \qquad \text{using } \ln\frac{a}{b}=\ln a-\ln b$$

$$\frac{\text{odds}_{\text{men}}}{\text{odds}_{\text{women}}}=e^{0.24}=1.27 \qquad \text{using } \log_e y=x\Rightarrow y=e^x$$

Odds ratio

The transformation $e^{0.24}$ undoes the logarithm and transforms the logistic regression slope into an **odds ratio,** in this case, the ratio of the odds that a man is a frequent binge drinker to the odds that a woman is a frequent binge drinker. In other words, we can multiply the odds for women by the odds ratio to obtain the odds for men:

$$\text{odds}_{\text{men}}=1.27(\text{odds}_{\text{women}})$$

In this case, the odds for men are 1.27 times the odds for women.

THINK ABOUT IT

What would happen if we coded the explanatory variable the other way?
Notice that we chose the coding for the indicator variable so that the regression slope is positive. This choice gives an odds ratio that is greater than 1. Had we coded women as 1 and men as 0, the signs of the parameters would be reversed. The fitted equation would be

$$\widehat{\ln(\text{odds})}=1.33-0.24x$$

and the odds ratio would be $e^{-0.24}=0.787$. The odds for women are 78.7% of the odds for men. You can check that this is just 1/1.27.

CHECK YOUR UNDERSTANDING

Refer to the previous Check Your Understanding (page 7).

1. Find the logistic regression equation for predicting the probability p that a person in the study will choose Commercial A based on gender ($x = 1$ for male and $x = 0$ for female). Show your work.
2. Compute and interpret the odds ratio.

Inference for Logistic Regression

Statistical inference for logistic regression is much like inference for simple linear regression. The big difference is the binomial setting for logistic 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 significance tests.

Confidence Intervals and Significance Tests for Logistic Regression Parameters

Suppose that the **logistic regression model** relating the proportion p of successful outcomes in a binomial setting to the explanatory variable x is given by

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

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

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

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

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

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

- To perform a **significance test** of $H_0: \beta_1 = 0$ versus $H_a: \beta_1 \neq 0$, compute the test statistic

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

The P-value is the area in the tails of the standard Normal distribution beyond z and $-z$.

The z test statistic is sometimes called a *Wald statistic*. Output from some statistical software reports the test statistic as $\chi^2 = \left(\frac{b_1}{\mathrm{SE}_{b_1}}\right)^2$. In that case, the P-value is obtained from the chi-square distribution with 1 degree of freedom.

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

THINK ABOUT IT

What does the odds ratio represent for a logistic regression model with a nonindicator, quantitative explanatory variable? We usually think of the slope as the predicted change in y for each one-unit increase in x. Let's use the logistic regression model to look at the predicted change in log odds if the value of the explanatory variable changes from x to $x + 1$.

$$\ln(\text{odds}_{x+1}) = b_0 + b_1(x+1)$$
$$\ln(\text{odds}_x) = b_0 + b_1 x$$
$$\ln(\text{odds}_{x+1}) - \ln(\text{odds}_x) = b_0 + b_1(x+1) - (b_0 + b_1 x) = b_1$$
$$\ln\left(\frac{\text{odds}_{x+1}}{\text{odds}_x}\right) = b_1$$
$$\frac{\text{odds}_{x+1}}{\text{odds}_x} = e^{b_1}$$

Now we can see that the odds ratio tells us the predicted change in the odds for a one-unit increase in the value of the explanatory variable. This fact should prove helpful for interpreting the odds ratio in a given setting.

EXAMPLE

College Students and Binge Drinking

Interpreting computer regression output

Figure 15.4 gives the output from Minitab for the binge-drinking example. The parameter estimates are given as $b_0 = -1.33063$ and $b_1 = 0.242518$, the same as we calculated directly earlier but with more significant digits. The standard errors are 0.0294 and 0.0471. Recall that we checked the BINS conditions for performing inference earlier.

Session

Predictor	Coef	SE Coef	Z	P	Odds Ratio	95% CI Lower	95% CI Upper
Constant	-1.33063	0.0294374	-45.20	0.000			
X	0.242518	0.0470986	5.15	0.000	1.27	1.16	1.40

FIGURE 15.4 Logistic regression output from Minitab for the binge-drinking data.

A 95% confidence interval for the slope is

$$b_1 \pm z^* \text{SE}_{b_1} = 0.242518 \pm 1.96(0.0470986)$$
$$= 0.2425 \pm 0.0923 = (0.1502, 0.3348)$$

We are 95% confident that the interval from 0.1502 to 0.3348 captures the true slope β_1 in the logistic regression model

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

The output provides a point estimate for the odds ratio (1.27) and the corresponding 95% confidence interval (1.16, 1.40). This is easy to compute from the interval for the slope:

$$(e^{b_1 - z^* \mathrm{SE}_{b_1}}, e^{b_1 + z^* \mathrm{SE}_{b_1}}) = (e^{0.1502}, e^{0.3348}) = (1.16, 1.40)$$

For this problem, we would report, "College men are significantly more likely to be frequent binge drinkers than college women (odds ratio = 1.27, 95% CI = 1.16 to 1.40)."

In applications like these, it is standard to use a 95% confidence level. 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.

CHECK YOUR UNDERSTANDING

The figure below shows Minitab output from a logistic regression analysis of the energy-drink study data from the previous two Check Your Understandings (pages 7 and 11). Recall that the model is intended to predict the probability p that a person in the study will choose Commercial A based on gender ($x = 1$ for male and $x = 0$ for female).

1. Check the conditions for performing inference about the logistic regression model.
2. Construct and interpret a 95% confidence interval for the slope.
3. Show how the confidence interval for the odds ratio in the computer output was computed. Then interpret this interval in context.
4. What conclusion would you draw from a test of H_0: the odds ratio is 1 versus H_a: the odds ratio is not 1? Justify your answer.

Session

Logistic Regression Table

Predictor	Coef	SE Coef	Z	P	Odds Ratio	95% CI Lower	95% CI Upper
Constant	-0.200671	0.201008	-1.00	0.318			
X	0.488353	0.263763	1.85	0.064	1.63	0.97	2.73

Putting It All Together: Logistic Regression

Logistic regression with an explanatory variable having two values is a very important special case. Now we look at settings where the explanatory variable

is quantitative. The following example is typical of many applications of logistic regression. This designed experiment has five different values for the explanatory variable.

EXAMPLE *Insecticide for Aphids*

A quantitative explanatory variable

An experiment was designed to examine how well the insecticide rotenone kills aphids called *Macrosiphoniella sanborni* that feed on the chrysanthemum plant.[5] The explanatory variable is the log concentration (in milligrams per liter) of the insecticide. Approximately 50 insects were assigned at random to be exposed to each concentration. Each insect was either killed or not killed by the insecticide. 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	Proportion killed $\hat{p}$,	$\ln\left(\frac{\hat{p}}{1-\hat{p}}\right)$
0.96	50	6	0.120	−1.992
1.33	48	16	0.333	−0.693
1.63	46	24	0.522	0.088
2.04	49	42	0.857	1.791
2.32	50	44	0.880	1.992

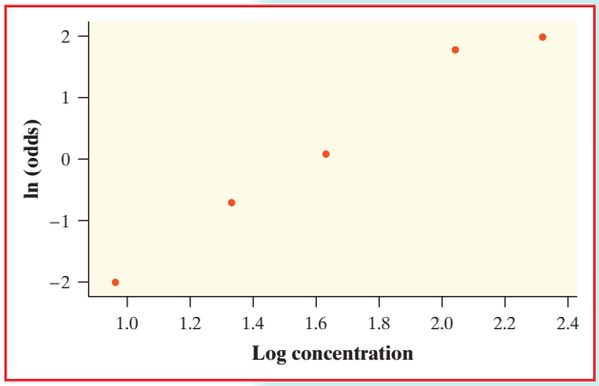


FIGURE 15.5 Plot of log odds of proportion killed versus log concentration for the insecticide data.

A scatterplot of the log odds versus the log concentration of insecticide is shown in Figure 15.5. Note the fairly strong, positive linear relationship between the variables. It makes sense to fit a model of the form

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

where the values of the explanatory variable x are 0.96, 1.33, 1.63, 2.04, and 2.32.

Figure 15.6 gives output from a Minitab logistic regression analysis of the insecticide data. From the output we see that the fitted model is

$$\widehat{\ln(\text{odds})} = -4.892 + 3.109x$$

Session

```
Logistic Regression Table

                                                    Odds     95% CI
Predictor             Coef   SE Coef      Z      P  Ratio  Lower  Upper
Constant          -4.89234  0.642613  -7.61  0.000
log concentration  3.10878  0.387891   8.01  0.000  22.39  10.47  47.90
```

FIGURE 15.6 Logistic regression output from Minitab for the insecticide data.

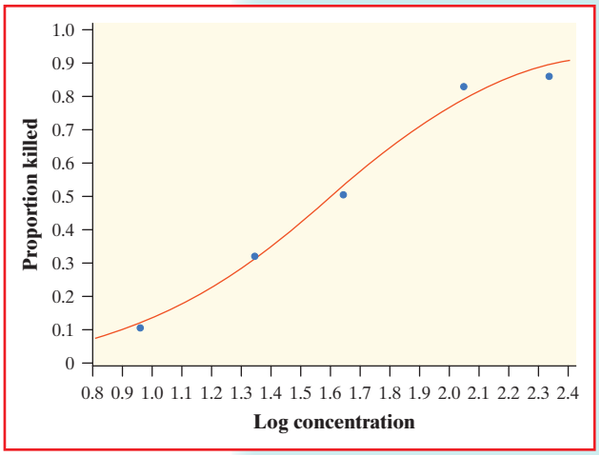


FIGURE 15.7 Plot of the proportion killed versus log concentration, with the logistic regression model, for the insecticide data.

The logistic regression model is shown in Figure 15.7, along with a scatterplot of the original data. We can use this model to predict the proportion of aphids killed when 30 mg/l of rotenone is applied:

$$\widehat{\ln(\text{odds})} = -4.89234 + 3.10878 \log(30) = -0.300$$

$$\widehat{\text{odds}} = e^{-0.300} = 0.741$$

If we substitute $\text{odds} = \dfrac{p}{1-p}$ and solve for p, we get

$$\frac{p}{1-p} = 0.741 \Rightarrow p = 0.741 - 0.741p$$

$$\Rightarrow 1.741p = 0.741 \Rightarrow p = 0.426$$

That is, we predict that a 30 mg concentration of rotenone will kill about 42.6% of aphids.

Figure 15.7 appears to show a strong relationship between the proportion of aphids killed and the concentration of insecticide. But suppose that rotenone has no ability to kill these aphids. What is the chance that we would observe experimental results at least as convincing as the ones in this study? 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 *Insecticide for Aphids*

Inference for the logistic regression model

PROBLEM: Here once again is Minitab computer output for the logistic regression analysis of the previous example. Assume that the conditions for performing inference are met.

Session

```
Logistic Regression Table

                                                    Odds     95% CI
Predictor             Coef   SE Coef      Z      P Ratio  Lower  Upper
Constant          -4.89234  0.642613  -7.61  0.000
log concentration  3.10878  0.387891   8.01  0.000 22.39  10.47  47.90
```

(a) What conclusion would you draw from a test of $H_0: \beta_1 = 0$ versus $H_a: \beta_1 \neq 0$ at the $\alpha = 0.05$ significance level? Justify your answer.

(b) Construct a 95% confidence interval for the true slope β_1. Explain what additional information the interval gives over the test in part (a).

(c) Interpret the odds ratio. Then show how the confidence interval for the odds ratio in the computer output was obtained.

SOLUTION:

(a) With $z = 8.01$ and a P-value of approximately 0, we would reject H_0 and conclude that there is convincing evidence that the true slope β_1 isn't 0.

(b) The 95% confidence interval is

$$b_1 \pm z^* SE_{b_1} = 3.10878 \pm 1.96(0.387891) = 3.10878 \pm 0.76027 = (2.34851, 3.86905)$$

This is an interval of plausible values for the true slope β_1. Note that 0 is not included in the interval, which is consistent with the significance test in part (a).

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 ratio is 1.

(c) An increase of one unit in the log concentration of insecticide (x) is associated with a predicted 22-fold increase in the odds that an aphid will be killed. Here is how the confidence interval for the odds ratio was obtained:

$$(e^{b_1 - z^* SE_{b_1}}, e^{b_1 + z^* SE_{b_1}}) = (e^{2.34851}, e^{3.86905}) = (10.47, 47.90)$$

For Practice *Try Exercise 13*

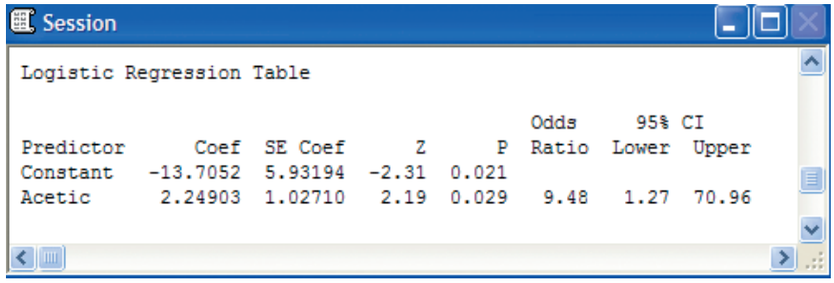
CHECK YOUR UNDERSTANDING

As cheddar cheese matures, a variety of chemical processes take place. The taste of mature cheese is related to the concentration of several chemicals in the final product. In one study, researchers analyzed the chemical compositions of a random sample of 30 pieces of a particular variety of cheddar cheese. Data from the study appear in the table below.

The variable Case is used to number the observations from 1 to 30. Values for Taste were obtained by combining the scores from several tasters. Three of the chemicals whose concentrations were measured were acetic acid, hydrogen sulfide, and lactic acid. For acetic acid and hydrogen sulfide, (natural) log transformations were taken. Thus, the possible explanatory variables are the transformed concentrations of acetic acid (Acetic) and hydrogen sulfide (H_2S) and the untransformed concentration of lactic acid (Lactic).[6]

Case	Taste	Acetic	H_2S	Lactic	Case	Taste	Acetic	H_2S	Lactic
01	12.3	4.543	3.135	0.86	16	40.9	6.365	9.588	1.74
02	20.9	5.159	5.043	1.53	17	15.9	4.787	3.912	1.16
03	39.0	5.366	5.438	1.57	18	6.4	5.412	4.700	1.49
04	47.9	5.759	7.496	1.81	19	18.0	5.247	6.174	1.63
05	5.6	4.663	3.807	0.99	20	38.9	5.438	9.064	1.99
06	25.9	5.697	7.601	1.09	21	14.0	4.564	4.949	1.15
07	37.3	5.892	8.726	1.29	22	15.2	5.298	5.220	1.33
08	21.9	6.078	7.966	1.78	23	32.0	5.455	9.242	1.44
09	18.1	4.898	3.850	1.29	24	56.7	5.855	10.199	2.01
10	21.0	5.242	4.174	1.58	25	16.8	5.366	3.664	1.31
11	34.9	5.740	6.142	1.68	26	11.6	6.043	3.219	1.46
12	57.2	6.446	7.908	1.90	27	26.5	6.458	6.962	1.72
13	0.7	4.477	2.996	1.06	28	0.7	5.328	3.912	1.25
14	25.9	5.236	4.942	1.30	29	13.4	5.802	6.685	1.08
15	54.9	6.151	6.752	1.52	30	5.5	6.176	4.787	1.25

Researchers decided to classify the cheese as acceptable (TasteOK = 1) if Taste ≥ 37 and unacceptable (TasteOK = 0) if Taste < 37. This is the response variable of interest. The Minitab output below shows the results of a logistic regression analysis for predicting the probability p that the cheese is acceptable using Acetic as the explanatory variable x.



1. Give the equation of the logistic regression model.
2. For the first piece of cheese tested, Acetic = 4.543. Use your model from Question 1 to predict the probability that a piece of cheese with this transformed acetic acid concentration has an acceptable taste. Show your work.

Assume that the conditions for performing inference are met in Questions 3 through 5.

3. What conclusion would you draw from a test of $H_0: \beta_1 = 0$ versus $H_a: \beta_1 \neq 0$ at the $\alpha = 0.05$ significance level? Justify your answer.
4. Construct a 95% confidence interval for the true slope β_1. Explain what additional information the interval gives over the test in Question 3.
5. Interpret the odds ratio. Then show how the confidence interval for the odds ratio in the computer output was obtained.

Multiple Logistic Regression

Multiple logistic regression

The cheese data set in the preceding Check Your Understanding includes three possible explanatory variables: Acetic, H_2S, and Lactic. You examined the model where Acetic was used to predict the odds that the cheese has an acceptable taste. Do the other explanatory variables contain additional information that will give us a better prediction? We use **multiple logistic regression** to answer this question. Generating the computer output is easy, just as it was when we generalized simple linear regression with one explanatory variable to multiple linear regression with more than one explanatory variable in Chapter 14. The statistical concepts are similar, but the computations are more complex. Here is the example.

EXAMPLE

Tasty Cheese

A multiple logistic regression model

We want to predict the odds that the cheese has an acceptable taste. The explanatory variables are Acetic, H_2S, and Lactic. Figure 15.8 (on the next page) gives Minitab output for this analysis.

The multiple logistic regression model is

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1(\text{Acetic}) + \beta_2(H_2S) + \beta_3(\text{Lactic})$$

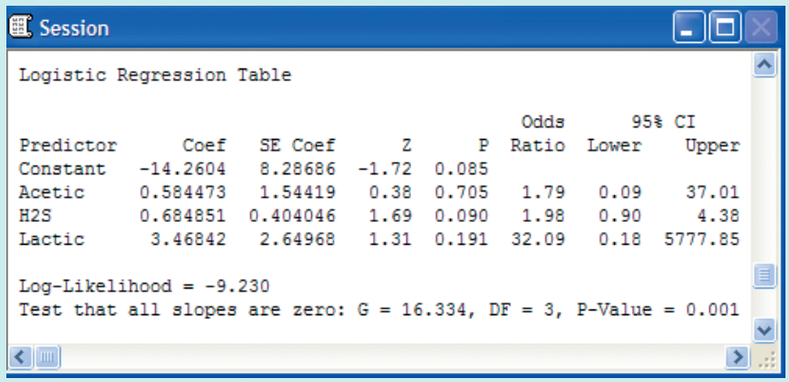


FIGURE 15.8 Logistic regression output from Minitab for the cheese data with Acetic, H_2S, and Lactic as the explanatory variables.

From the computer output, the fitted model is

$$\widehat{\ln(\text{odds})} = -14.26 + 0.584(\text{Acetic}) + 0.685(H_2S) + 3.468(\text{Lactic})$$

For Case 6, which had Acetic = 5.697, H_2S = 7.601, and Lactic = 1.09, this model gives

$$\begin{aligned}\widehat{\ln(\text{odds})} &= -14.26 + 0.584(5.697) + 0.685(7.601) + 3.468(1.09)\\ &= -1.946\\ \widehat{(\text{odds})} &= e^{-1.946} = 0.143\end{aligned}$$

If we substitute odds $= \dfrac{p}{1-p}$ and solve for p, we get

$$\frac{p}{1-p} = 0.143 \Rightarrow p = 0.143 - 0.143p \Rightarrow 1.143p = 0.143 \Rightarrow p = 0.125$$

That is, we predict that a piece of cheddar cheese with this chemical composition has probability 0.125 of having an acceptable taste.

When performing inference about a multiple linear regression model, we first examine the null hypothesis that all the regression coefficients for the explanatory variables are zero. We do the same for multiple logistic regression. A test of $H_0: \beta_1 = \beta_2 = \beta_3 = 0$ versus H_a: at least one of these β's is not 0 uses a chi-square distribution with 3 degrees of freedom. The appropriate test statistic is called G and is shown in the last line of the Minitab output.

In this case, G = 16.33 and the *P*-value is 0.001. We have sufficient evidence to reject H_0 and conclude that one or more of the explanatory variables are useful for predicting the odds that the cheese has an acceptable taste. We now examine the coefficients for each variable and the tests that each of these is 0. The *P*-values are 0.71, 0.09, and 0.19. None of the null hypotheses—$H_0: \beta_1 = 0$, $H_0: \beta_2 = 0$, and $H_0: \beta_3 = 0$—can be rejected at the 0.05 significance level.

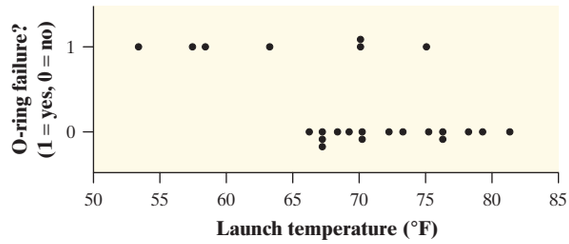
Our initial multiple logistic regression analysis told us that the explanatory variables contain information that is useful for predicting whether or not the

cheese has an acceptable taste. 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.

case closed

The Risks of Space Exploration

In the chapter-opening Case Study (page 3), we presented data on the launch temperature and whether or not O-rings failed for each of the 24 shuttle flights before the *Challenger* disaster. Here again is a modified scatterplot of the data.



Here is Minitab output from a logistic regression analysis.

Session

```
Logistic Regression Table

                                                  Odds     95% CI
Predictor         Coef   SE Coef      Z      P  Ratio  Lower  Upper
Constant       15.0429   7.37862   2.04  0.041
Temperature  -0.232163  0.108236  -2.14  0.032   0.79   0.64   0.98

Log-Likelihood = -10.158
Test that all slopes are zero: G = 7.952, DF = 1, P-Value = 0.005
```

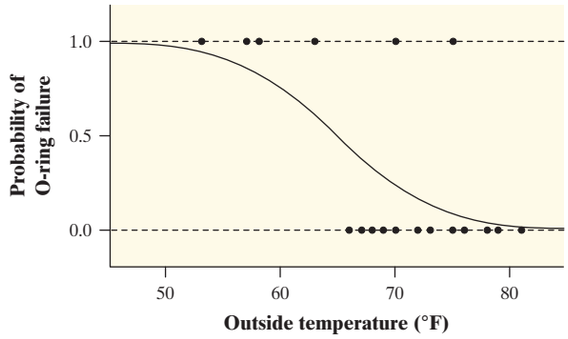
The fitted logistic regression model is

$$\widehat{\ln\left(\frac{p}{1-p}\right)} = 15.0429 - 0.232163x$$

where p is the probability of having failed O-rings and x is the launch temperature (°F).

A scatterplot with the logistic regression model added is shown on the next page. The predicted odds ratio is $e^{-0.232163} = 0.793$. An increase of one degree

Fahrenheit in the launch temperature (x) is predicted to decrease the odds of O-ring failure by a factor of about 0.8.



On the morning when *Challenger* lifted off, the temperature was 31°F. Our model gives

$$\widehat{\ln(\text{odds})} = 15.0429 - 0.232163(31) = 7.846$$

$$\widehat{(\text{odds})} = e^{7.846} = 2555.49$$

If we substitute odds $= \dfrac{p}{1-p}$ and solve for p, we get

$$\frac{p}{1-p} = 2555.49 \Rightarrow p = 2555.49 - 2555.49p \Rightarrow 2556.49p = 2555.49 \Rightarrow p = 0.9996$$

That is, we predict the probability of O-ring failure on a shuttle flight that launches when the temperature is 31°F to be 0.9996.

Why did NASA scientists allow the *Challenger* launch to proceed? Because they analyzed only data from previous shuttle flights that had O-ring failures. By omitting the launches with no O-ring failures from the data set, the scientists' model for making predictions was flawed.

SECTION 15.1 Summary

- Let $\hat{p}$ be the sample proportion of successful outcomes in a binomial setting. The **odds** of a success are $\dfrac{\hat{p}}{1-\hat{p}}$, the ratio of the proportion of times a success occurs to the proportion of times a failure occurs.
- The **logistic regression model** relates the **log odds** to the explanatory variable x

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

  The parameters of the logistic model are β_0 and β_1.
- The **odds ratio** is e^{β_1}, where β_1 is the slope in the logistic regression model. The odds ratio tells us the predicted change in the odds for a one-unit increase in the value of the explanatory variable.
- A **level C confidence interval for the slope β_1** is

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

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

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

  In both of these expressions, z^* is the critical value for the standard Normal curve with area C between $-z^*$ and z^*.
- To perform a **significance test** of $H_0: \beta_1 = 0$ versus $H_a: \beta_1 \neq 0$, compute the test statistic

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

  The P-value is the area in the tails of the standard Normal distribution beyond z and $-z$.
- In **multiple logistic regression,** the response variable has two possible values, as in logistic regression, but there can be several explanatory variables.

SECTION 15.1 Exercises

pg 6

1. **Blood pressure and heart disease** There is much evidence that high blood pressure is associated with increased risk of death from heart disease. A major study of this association examined random samples of 3338 men with high blood pressure and 2676 men with low blood pressure. During the period of the study, 55 of the men with high blood pressure and 21 of the men with low blood pressure died from heart disease.

   (a) Find the odds of dying from heart disease among men with high blood pressure. Show your work.

   (b) Find the odds of dying from heart disease among men with low blood pressure. Show your work.

2. **Textbooks and gender bias** To what extent does vocabulary illustrate gender bias in textbooks? A study of this question analyzed a random sample of sentences from 10 textbooks. One part of the study examined the use of the words "girl," "boy," "man," and "woman." We will call the first two words juvenile and the last two adult. Here are data from one of the texts:[7]

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

   (a) Find the odds that a female reference is juvenile. Show your work.

   (b) Find the odds that a male reference is juvenile. Show your work.

3. **Blood pressure and heart disease** Refer to Exercise 1.

   (a) Compute the log odds of dying from heart disease for men with high blood pressure and for men with low blood pressure.

   (b) Find the logistic regression equation for predicting the probability p that a man will die from heart disease ($x = 1$ for high blood pressure and $x = 0$ for low blood pressure). Show your work.

4. **Textbooks and gender bias** Refer to Exercise 2.

   (a) Compute the log odds that a reference is juvenile for female and for male words.

   (b) Find the logistic regression equation for predicting the probability p that a reference is juvenile ($x = 1$ for female words and $x = 0$ for male words). Show your work.

5. **Blood pressure and heart disease** Refer to Exercises 1 and 3. Compute and interpret the odds ratio (high to low blood pressure) for getting heart disease.

6. **Textbooks and gender bias** Refer to Exercises 2 and 4. Compute and interpret the odds ratio (female words to male words) of a juvenile reference.

7. **Franchises** Many popular businesses are franchises—think of McDonald's. The owner of a local franchise benefits from the brand recognition, national advertising, and detailed guidelines provided by the franchise chain. In return, he or she pays fees to the franchise firm and agrees to follow its policies. The relationship between the local owner and the franchise firm is spelled out in a detailed contract. One clause that the contract may or may not contain is the entrepreneur's right to an exclusive territory. This means that the new outlet will be the only representative of the franchise in a specified territory and will not have to compete with other outlets of the same chain. How does the presence of an exclusive-territory clause in the contract relate to the survival of the business?

   A study designed to address this question collected data from a random sample of 170 new franchise firms. Two categorical variables were measured for each franchisor. First, the franchisor was classified as successful or not based on whether or not it was still offering franchises as of a certain date. Second, the contract each franchisor offered to franchisees was classified according to whether or not it contained an exclusive-territory clause.[8] Here are the data:

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

   (a) Find the logistic regression model for predicting the probability p that a franchise firm is successful based on whether the firm offers its franchisees an exclusive-territory contract. Show your work ($x = 1$ for Yes and $x = 0$ for No).

   (b) Compute and interpret the odds ratio (exclusive territory to not).

8. **No sweat!** Following complaints about the working conditions in some clothing factories in the United States and abroad, a joint government and industry commission recommended that companies that monitor and enforce proper standards be allowed to display a "No Sweat" label on their products. Does

the presence of these labels influence consumer behavior?

A survey of a random sample of U.S. residents aged 18 or older asked a series of questions about how likely they would be to purchase a garment under various conditions. For some conditions, it was stated that the garment had a "No Sweat" label. For others, there was no mention of such a label. On the basis of the responses, each person was classified as a "label user" or a "label nonuser."[9] Here are the data for comparing women and men:

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

(a) Find the logistic regression model for predicting the probability p that a person is a label user based on gender. Show your work ($x = 1$ for Women and $x = 0$ for Men).

(b) Compute and interpret the odds ratio (women to men).

9. **Blood pressure and heart disease** The Minitab output below shows the results of a logistic regression analysis of the heart disease study data from Exercise 1. Recall that the model is intended to predict the probability p that a man will die from heart disease ($x = 1$ for high blood pressure and $x = 0$ for low blood pressure).

Session

Logistic Regression Table

Predictor	Coef	SE Coef	Z	P	Odds Ratio	95% CI Lower	95% CI Upper
Constant	-4.83968	0.219078	-22.09	0.000			
High blood pressure?	0.750498	0.257840	2.91	0.004	2.12	1.28	3.51

(a) Check the conditions for performing inference about the logistic regression model.

(b) Construct and interpret a 95% confidence interval for the slope parameter β_1.

(c) Show how the confidence interval for the odds ratio in the computer output was computed. Then interpret this interval in context.

(d) What conclusion would you draw from a test of H_0: the odds ratio is 1 versus H_a: the odds ratio is not 1? Justify your answer.

10. **Textbooks and gender bias** The Minitab output at top right shows the results of a logistic regression analysis of the gender-bias study data from Exercise 2. Recall that the model is intended to predict the probability p that a reference is juvenile ($x = 1$ for female words and $x = 0$ for male words).

Session Brush Mode

Logistic Regression Table

Predictor	Coef	SE Coef	Z	P	Odds Ratio	95% CI Lower	95% CI Upper
Constant	-0.430783	0.178131	-2.42	0.016			
Word gender	1.81708	0.368641	4.93	0.000	6.15	2.99	12.67

(a) Check the conditions for performing inference about the logistic regression model.

(b) Construct and interpret a 95% confidence interval for the slope parameter β_1.

(c) Show how the confidence interval for the odds ratio in the computer output was computed. Then interpret this interval in context.

(d) What conclusion would you draw from a test of H_0: the odds ratio is 1 versus H_a: the odds ratio is not 1? Justify your answer.

11. **Franchises** The Minitab output below shows the results of a logistic regression analysis of the franchise success study from Exercise 7. Recall that the model is intended to predict the probability p that a franchise firm is successful based on whether or not the firm offers an exclusive-territory contract.

Session

Logistic Regression Table

Predictor	Coef	SE Coef	Z	P	Odds Ratio	95% CI Lower	95% CI Upper
Constant	0.143101	0.378932	0.38	0.706			
Exclusive territory?	1.01267	0.426920	2.37	0.018	2.75	1.19	6.36

(a) Check the conditions for performing inference about the logistic regression model.

(b) Construct and interpret a 95% confidence interval for the slope.

(c) Show how the confidence interval for the odds ratio in the computer output was computed. Then interpret this interval in context.

(d) What conclusion would you draw from a test of H_0: the odds ratio is 1 versus H_a: the odds ratio is not 1? Justify your answer.

12. **No sweat!** The Minitab output below shows the results of a logistic regression analysis of the survey about No Sweat labels from Exercise 8. Recall that the model is intended to predict the probability p that a person is a label user based on gender.

Session

Logistic Regression Table

Predictor	Coef	SE Coef	Z	P	Odds Ratio	95% CI Lower	95% CI Upper
Constant	-2.11581	0.203719	-10.39	0.000			
Gender	0.807905	0.248327	3.25	0.001	2.24	1.38	3.65

(a) Check the conditions for performing inference about the logistic regression model.

(b) Construct and interpret a 95% confidence interval for the slope.

(c) Show how the confidence interval for the odds ratio in the computer output was computed. Then interpret this interval in context.

(d) What conclusion would you draw from a test of H_0 : the odds ratio is 1 versus H_a : the odds ratio is not 1? Justify your answer.

Exercises 13 through 16 refer to the following setting. In the Check Your Understanding on page 16, we analyzed data from an observational study on the chemical composition and taste of randomly selected pieces of a particular variety of cheddar cheese. In that analysis, we used the natural logarithm of the acetic acid concentration as the explanatory variable.

13. **Cheese!** The Minitab output below shows the results of a logistic regression analysis for predicting the probability p that the cheese has an acceptable taste using H_2S, the natural logarithm of the hydrogen sulfide concentration, as the explanatory variable.

pg 15

Session

Logistic Regression Table

Predictor	Coef	SE Coef	Z	P	Odds Ratio	95% CI Lower	95% CI Upper
Constant	-7.27894	2.52165	-2.89	0.004			
H2S	0.939916	0.344331	2.73	0.006	2.56	1.30	5.03

(a) Give the equation of the logistic regression model.

(b) For the first piece of cheese tested, $H_2S = 3.135$. Use your model from part (a) to predict the probability that a piece of cheese with this transformed hydrogen sulfide concentration has an acceptable taste. Show your work.

14. **Cheese!** The Minitab output below shows the results of a logistic regression analysis for predicting the probability p that the cheese has an acceptable taste using Lactic, the concentration of lactic acid, as the explanatory variable.

Session

Logistic Regression Table

Predictor	Coef	SE Coef	Z	P	Odds Ratio	95% CI Lower	95% CI Upper
Constant	-10.7799	3.97542	-2.71	0.007			
Lactic	6.33193	2.45317	2.58	0.010	562.24	4.59	68880.95

(a) Give the equation of the logistic regression model.

(b) For the first piece of cheese tested, Lactic = 0.86. Use your model from part (a) to predict the probability that a piece of cheese with this lactic acid concentration has an acceptable taste. Show your work.

15. **Cheese!** Refer to Exercise 13. Assume that the conditions for performing inference are met.

(a) What conclusion would you draw from a test of $H_0: \beta_1 = 0$ versus $H_a: \beta_1 \neq 0$ at the $\alpha = 0.05$ significance level? Justify your answer.

(b) Construct a 95% confidence interval for the true slope β_1. Explain what additional information the interval gives over the test in part (a).

(c) Interpret the odds ratio. Then show how the confidence interval for the odds ratio in the computer output was obtained.

16. **Cheese!** Refer to Exercise 14. Assume that the conditions for performing inference are met.

(a) What conclusion would you draw from a test of $H_0: \beta_1 = 0$ versus $H_a: \beta_1 \neq 0$ at the $\alpha = 0.05$ significance level? Justify your answer.

(b) Construct a 95% confidence interval for the true slope β_1. Explain what additional information the interval gives over the test in part (a).

(c) Interpret the odds ratio. Then show how the confidence interval for the odds ratio in the computer output was obtained.

17. **Childhood obesity** Suppose we want to predict the probability p that a child will be obese at age 12 based on his or her body mass index x at age 4. Data from one study involving 1042 randomly selected 4-year-olds gives the fitted logistic regression model

$$\widehat{\ln\left(\frac{p}{1-p}\right)} = -16.71 + 0.989x$$

(a) Find and interpret the odds ratio.

(b) Use this model to predict the probability that a 4-year-old child whose BMI is 28 will be obese at age 12. Show your method clearly.

18. **The dangers of space exploration** Suppose we want to predict the probability p that a space shuttle flight will have damaged O-rings based on the launch temperature x (in °F). In the Case Closed, we gave the fitted logistic regression model as

$$\widehat{\ln\left(\frac{p}{1-p}\right)} = 15.0429 - 0.232163x$$

(a) Find and interpret the odds ratio.

(b) Use this model to predict the probability of damaged O-rings on a shuttle flight that launches when the temperature is 45°F. Show your method clearly.

Exercises 19 through 21 refer to the following setting. Do high school grades or SAT scores predict college GPA

more accurately? To find out, researchers collected data from a random sample of 224 students at a large university.[10] Among the explanatory variables recorded at the time the students enrolled in the university were gender, average high school grades in mathematics (HSM), science (HSS), and English (HSE), and students' scores on the SAT Math and Critical Reading tests. In Chapter 14, page 37, we used multiple regression for our analysis. (You should review the description of the study.) This time, we'll use logistic regression. We define an indicator variable, HIGPA, to be 1 if the GPA is 3.0 or better and 0 otherwise.

19. **High school grades and GPA** Minitab output from a logistic regression analysis for predicting the probability of getting a high college GPA using high school math grade as the explanatory variable is shown below.

Session

```
Logistic Regression Table

                                                 Odds     95% CI
Predictor       Coef   SE Coef      Z      P  Ratio  Lower  Upper
Constant   -5.06382   1.00706  -5.03  0.000
hsm         0.531046  0.113865   4.66  0.000   1.70   1.36   2.13
```

(a) Is a student who earned B's in high school math (HSM = 7) likely to earn a high GPA in college? Give appropriate evidence to support your answer.

(b) How much more likely is a student who earned B+ grades in high school math (HSM = 8) to earn a high GPA in college than a student who earned B's? Give appropriate evidence to support your answer.

20. **Gender and GPA** The Minitab output below shows the results of a logistic regression analysis for predicting the probability of getting a high college GPA based on gender ($x = 1$ for female and 0 for male). Write a brief report that summarizes the results.

Session

```
Logistic Regression Table

                                                 Odds     95% CI
Predictor       Coef   SE Coef      Z      P  Ratio  Lower  Upper
Constant   -0.551497  0.172446  -3.20  0.001
indgender   0.0608737 0.288927   0.21  0.833   1.06   0.60   1.87
```

21. **Grades and GPA** The Minitab output at top right shows the results of a logistic regression analysis using high school math (HSM), science (HSS), and English (HSE) grades as the explanatory variables. One student in the study had HSM = 7, HSS = 10, and HSE = 9. What does the model predict for the probability that this student's college GPA is 3.0 or higher? Show your work.

Session

```
Logistic Regression Table

                                                 Odds     95% CI
Predictor       Coef   SE Coef      Z      P  Ratio  Lower  Upper
Constant    -6.05283   1.15623  -5.23  0.000
hsm          0.370974  0.130223   2.85  0.004   1.45   1.12   1.87
hss          0.248929  0.127530   1.95  0.051   1.28   1.00   1.65
hse         0.0360515  0.125315   0.29  0.774   1.04   0.81   1.33

Log-Likelihood = -130.846
Test that all slopes are zero: G = 33.648, DF = 3, P-Value = 0.000
```

22. **Cheese!** Refer to the cheddar cheese study described in the Check Your Understanding on page 16. The Minitab output below displays the results of a logistic regression analysis using Acetic and H_2S as the explanatory variables. For Case 6, Acetic = 5.697, $H_2S = 7.601$, and Lactic = 1.09. What does the model predict for the probability that the cheese has an acceptable taste? Show your work.

Session

```
Logistic Regression Table

                                                 Odds     95% CI
Predictor       Coef   SE Coef      Z      P  Ratio  Lower  Upper
Constant    -12.8470   7.86772  -1.63  0.102
Acetic       1.09631   1.38201   0.79  0.428   2.99   0.20  44.93
H2S         0.830262  0.367313   2.26  0.024   2.29   1.12   4.71


Log-Likelihood = -10.284
Test that all slopes are zero: G = 14.226, DF = 2, P-Value = 0.001
```

23. **Grades and GPA** Refer to Exercise 21. Assume that the conditions for inference are met.

(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).

24. **Cheese!** Refer to Exercise 22. Assume that the conditions for inference are met.

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

(b) Give the coefficient for Acetic with a 95% confidence interval. Do the same for H_2S.

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

25. **Cheese!** Use statistical software to fit a logistic regression model for predicting the probability that the cheese has an acceptable taste using Acetic and Lactic as the explanatory variables.

(a) For Case 6, Acetic = 5.697, $H_2S = 7.601$, and Lactic = 1.09. What does the model predict for the

probability that the cheese has an acceptable taste? Show your work.

(b) Assume that the conditions for inference are met. Summarize what the computer output tells you about the coefficients in the regression model.

26. **Cheese!** Use statistical software to fit a logistic regression model for predicting the probability that the cheese has an acceptable taste using H_2S and Lactic as the explanatory variables.

(a) For Case 6, Acetic = 5.697, H_2S = 7.601, and Lactic = 1.09. What does the model predict for the probability that the cheese has an acceptable taste? Show your work.

(b) Assume that the conditions for inference are met. Summarize what the computer output tells you about the coefficients in the regression model.

27. **Getting into college** For their final project, a group of AP Statistics students collected data from a sample of applicants to a selective university. They used computer software to fit a multiple logistic regression model for predicting the probability that a student is admitted to the university based on his or her combined SAT Math and Critical Reading scores and cumulative GPA. The resulting model was $\widehat{\ln(\text{odds})} = -17.4897 + 0.00547(\text{SAT}) + 3.3335(\text{GPA})$. One applicant had a combined SAT score of 1450 and a GPA of 3.95. What does the model predict for this student's probability of being admitted? Show your work.

28. **Simpson's paradox*** Here is an example of Simpson's paradox, *the reversal of the direction of 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 before surgery:

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

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

(a) Use logistic regression to model the odds of death using 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).

*This exercise involves the optional material on Simpson's paradox from Chapter 1, pages 20 and 21.

SECTION 15.1 Notes and Data Sources

1. Data from the Report of the Presidential Commission on the Space Shuttle Challenger Accident, 1986.
2. 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, 2002; and D. W. Hosmer and S. Lemeshow, *Applied Logistic Regression*, 2nd ed., Wiley, 2000.
3. Results of this survey are reported in Henry Wechsler et al., "Health and behavioral consequences of binge drinking in college," *Journal of the American Medical Association*, 272 (1994), pp. 1672–1677.
4. The study is described in Philip R. Nader et al., "Identifying risk for obesity in early childhood," *Pediatrics*, 118, No. 3 (September 2006), pp. e594–e601. We obtained the data as part of the *Statistically Speaking* video series, a joint project of Coast Learning Systems and W. H. Freeman and Company.
5. 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 text, the category is described as "apparently dead, moribund, or so badly affected as to be unable to walk more than a few steps." This is an early example of the need to make careful judgments when defining variables to be used in a statistical analysis. An insect that is "unable to walk more than a few steps" is unlikely to eat very much of a chrysanthemum plant!
6. These data are based on experiments performed by G. T. Lloyd and E. H. Ramshaw of the CSIRO Division of Food Research, Victoria, Australia. Some results of the statistical analyses of these data are given in G. P. McCabe, L. McCabe, and A. Miller, "Analysis of taste and chemical composition of cheddar cheese, 1982–83 experiments," CSIRO Division of Mathematics and Statistics Consulting Report VT85/6; and in I. Barlow et al., "Correlations and changes in flavour and chemical parameters of cheddar cheeses during maturation," *Australian Journal of Dairy Technology*, 44 (1989), pp. 7–18.
7. From Monica Macaulay and Colleen Brice, "Don't touch my projectile: gender bias and stereotyping in syntactic examples," *Language*, 73, No. 4 (1997), pp. 798–825.
8. P. Azoulay and S. Shane, "Entrepreneurs, contracts, and the failure of young firms," *Management Science*, 47 (2001), pp. 337–358.
9. Marsha A. Dickson, "Utility of no sweat labels for apparel customers: profiling label users and predicting their purchases," *Journal of Consumer Affairs*, 35 (2001), pp. 96–119.
10. Results of the study are reported in P. F. Campbell and G. P. McCabe, "Predicting the success of freshmen in a computer science major," *Communications of the ACM*, 27 (1984), pp. 1108–1113.