Example 1

Predicting the Selling Price of a House

Picture the Scenario

You are saving to buy a home, and you wonder how much it will cost. The House Selling Prices OR data file on the text CD has observations on 200 recent home sales in Corvallis, Oregon. Table 13.1 shows data for two homes.

Table 13.1 Selling Prices and Related Factors for a Sample of Home Sales

House	Selling Price	House Size (sq. ft)	Number of Bedrooms	Number of Bathrooms	Lot Size (sq. ft)	Year Built	Garage (Y/N)
1	$232,500	1679	3	1.5	10,019	1976	Y
2	$158,000	1292	1	1	217,800	1958	N

Variables listed are the selling price (in dollars), house size (in square feet), number of bedrooms, number of bathrooms, the lot size (in square feet), and whether or not the house has a garage. Table 13.2 reports the mean and standard deviation of these variables for all 200 home sales.

Table 13.2 Descriptive Statistics for Sales of 200 Homes

	Selling Price	House Size (sq. ft)	Number of Bedrooms	Number of Bathrooms	Lot Size (sq. ft)	Age
Mean	$267,466	2551	3.08	2.03	23,217	34.75
Standard deviation	$115,808	1238	1.10	0.77	47,637	24.44

Questions to Explore

In your community, if you know the values of such variables,

- How can you predict a home's selling price?
- How can you describe the association between selling price and the other variables?
- How can you make inferences about the population based on the sample data?

Thinking Ahead

You can find a regression equation to predict selling price by treating one of the other variables as the explanatory variable. However, since there are *several* explanatory variables, you might make better predictions by using *all* of them at once. That's the idea behind **multiple regression**. It uses *more than one* explanatory variable to predict a response variable. We'll learn how to apply multiple regression to a variety of analyses in Examples 2, 3, 8, 9, and 10 of this chapter.

Example 2

Predicting Selling Price Using House Size and Number of Bedrooms

Picture the Scenario

For the house selling price data described in Example 1, MINITAB reports the results in Table 13.3 for a multiple regression analysis with selling price as the response variable and with house size and number of bedrooms as explanatory variables.

Table 13.3 Regression of Selling Price on House Size and Bedrooms

The regression equation is price = 60,102 + 63.0 house_size + 15,170 bedrooms

Predictor	Coef	SE Coef	T	P
Constant	60102	18623	3.23	0.001
House_size	62.983	4.753	13.25	0.000
Bedrooms	15170	5330	2.85	0.005

Questions to Explore

a. State the prediction equation.

b. The first home listed in Table 13.1 has house size = 1679 square feet, three bedrooms, and selling price $232,500. Find its predicted selling price and the residual (prediction error). Interpret the residual.

Think It Through

a. The response variable is y = selling price. Let x_1 = house size and x_2 = the number of bedrooms. From Table 13.3, the prediction equation is

$$\hat{y} = 60{,}102 + 63.0x_1 + 15{,}170x_2.$$

b. For $x_1 = 1679$ and $x_2 = 3$, the predicted selling price is

$\hat{y} = 60{,}102 + 63.0(1679) + 15{,}170(3) = 211{,}389$, that is, $211,389.

The residual is the prediction error,

$$y - \hat{y} = \$232{,}500 - \$211{,}389 = \$21{,}111.$$

This result tells us that the actual selling price was $21,111 higher than predicted.

Insight

The coefficients of house size and number of bedrooms are positive. As these variables increase, the predicted selling price increases, as we would expect.

Try Exercise 13.1

Example 3

Predicting House Selling Prices

Picture the Scenario

For the 200 observations on y = selling price in thousands of dollars, using, x_1 = house size in thousands of square feet and x_2 = number of bedrooms, Table 13.5 shows the ANOVA (analysis of variance) table that MINITAB reports for the multiple regression model.

Table 13.5 ANOVA Table and R^2 for Predicting House Selling Price (in thousands of dollars) Using House Size (in thousands of square feet) and Number of Bedrooms

R-Sq = 52.4%

Analysis of Variance

Source	DF	SS
Regression	2	1399524
Residual Error	197	1269345
Total	199	2668870

Questions to Explore

a. Show how to use the sums of squares in the ANOVA table to find R^2 for this multiple regression model. Interpret.

b. Find and interpret the multiple correlation.

Think It Through

a. From the sum of squares (SS) column, the total sum of squares is $\Sigma(y - \bar{y})^2 = 2{,}668{,}870$. The residual sum of squares from using the multiple regression equation to predict y is $\Sigma(y - \hat{y})^2 = 1{,}269{,}345$. The value of R^2 is

$$R^2 = \frac{\Sigma(y - \bar{y})^2 - \Sigma(y - \hat{y})^2}{\Sigma(y - \bar{y})^2} = \frac{2{,}668{,}870 - 1{,}269{,}345}{2{,}668{,}870} = 0.524.$$

Using house size and number of bedrooms together to predict selling price reduces the prediction error by 52%, relative to using $\bar{y}$ alone to predict selling price. The R^2 statistic appears (in percentage form) in Table 13.5 under the heading "R-sq."

b. The multiple correlation between selling price and the two explanatory variables is $R = \sqrt{R^2} = \sqrt{0.524} = 0.72$. This equals the correlation for the 200 homes between the observed selling prices and the predicted selling prices from multiple regression. There's a moderately strong association between the observed and the predicted selling prices. In summary, house size and number of bedrooms are very helpful in predicting selling prices.

Insight

The bivariate r^2 value for predicting selling price is 0.51 with house size as the predictor. The multiple regression model has $R^2 = 0.52$, only a marginally larger value to using only house size as a predictor. We appear to be just as well off with the bivariate model using house size as the predictor as compared to using the multiple model. An advantage of using only the bivariate model is easier interpretation of the coefficients.

Try Exercise 13.11

Example 4

Female Athletes' Weight

Picture the Scenario

The College Athletes data set on the text CD comes from a study of 64 University of Georgia female athletes who participated in Division I sports. The study measured several physical characteristics, including total body

weight in pounds (TBW), height in inches (HGT), the percent of body fat (%BF), and age. Table 13.7 shows the ANOVA table for the regression of weight on height, % body fat, and age.

Table 13.7 ANOVA Table for Multiple Regression Analysis of Athlete Weights

```
Analysis of Variance
Source          DF       SS       MS       F      P
Regression       3    12407.9   4136.0   40.48  0.000
Residual Error  60     6131.0    102.2
Total           63    18539.0
S = 10.1086      R-Sq = 66.9%
```

Question to Explore

For female athletes at particular values of height, percent of body fat, and age, estimate the standard deviation of their weights.

Think It Through

The SS column tells us that the residual sum of squares is 6131.0. There were $n = 64$ observations and 4 parameters in the regression model, so the DF column reports $df = n - 4 = 60$ opposite the residual SS. The mean square error is

$$s^2 = (\text{residual SS})/df = 6131.0/60 = 102.2.$$

It appears in the mean square (MS) column, in the row labeled "Residual Error." The residual standard deviation is $s = \sqrt{102.2} = 10.1$, identified as S. For athletes with certain fixed values of height, percent body fat, and age, the weights vary with a standard deviation of about 10 pounds.

Insight

If the conditional distributions of weight are approximately bell shaped, about 95% of the weight values fall within about $2s = 20$ pounds of the true regression equation. More precisely, software can report *prediction intervals* within which a response outcome has a certain chance of falling. For instance, at $x_1 = 66$, $x_2 = 18$, and $x_3 = 20$, which are close to the mean predictor values, software reports $\hat{y} = 133.9$ and a 95% prediction interval of 133.9 ± 20.4.

Try Exercise 13.21

Example 5

Athletes' Weight

Picture the Scenario

For the 64 female college athletes, the ANOVA table for the multiple regression predicting y = weight using x_1 = height, x_2 = percent body fat, and x_3 = age shows:

```
Source          DF       SS       MS       F      P
Regression       3    12407.9   4136.0   40.48  0.000
Residual Error  60     6131.0    102.2
```

Questions to Explore

a. State and interpret the null hypothesis tested in this table.

b. From the F table, which F value would have a P-value of 0.05 for these data?

c. Report the observed test statistic and P-value. Interpret the P-value, and make a decision for a 0.05 significance level.

Think It Through

a. Since there are three explanatory variables, the null hypothesis is H_0: $\beta_1 = \beta_2 = \beta_3 = 0$. It states that weight is independent of height, percent body fat, and age.

b. In the DF column, the ANOVA table shows $df_1 = 3$ and $df_2 = 60$. The F table or software indicates that the F value with right-tail probability of 0.05 is 2.76. (See also Table 13.8.)

c. From the ANOVA table, the observed F test statistic value is 40.5. Since this is well above 2.76, the P-value is less than 0.05. The ANOVA table reports P-value = 0.000. If H_0 were true, it would be extremely unusual to get such a large F test statistic. We can reject H_0 at the 0.05 significance level. In summary, we conclude that at least one predictor has an effect on weight.

Insight

The F test tells us that *at least one* explanatory variable has an effect. The following section discusses how to follow up from the F test to investigate which explanatory variables have a statistically significant effect on predicting y.

Try Exercise 13.25

Example 6

What Helps Predict a Female Athlete's Weight?

Picture the Scenario

The College Athletes data set (Examples 4 and 5) measured several physical characteristics, including total body weight in pounds (TBW), height in inches (HGT), the percent of body fat (%BF), and age. Table 13.9 shows summary statistics for these variables.

Table 13.9 Summary Statistics for Study of Female College Athletes

Variables are TBW = total body weight, HGT = height, %BF = percent body fat, and AGE.

Variable	Mean	StDev	Minimum	Q1	Median	Q3	Maximum
TBW	133.0	17.2	96.0	119.2	131.5	143.8	185.0
HGT	65.5	3.5	56.0	63.0	65.0	68.2	75.0
%BF	18.4	4.1	11.2	15.2	18.5	21.5	27.6
AGE	20.0	1.98	17.0	18.0	20.0	22.0	23.0

Table 13.10 shows results of fitting a multiple regression model for predicting weight using the other variables. The predictive power is good, with $R^2 = 0.669$.

Table 13.10 Multiple Regression Analysis for Predicting Weight

Predictors are HGT = height, %BF = body fat, and age of subject.

Predictor	Coef	SE Coef	T	P
Constant	−97.69	28.79	−3.39	0.001
HGT	3.4285	0.3679	9.32	0.000
%BF	1.3643	0.3126	4.36	0.000
AGE	−0.9601	0.6483	−1.48	0.144

R-Sq = 66.9%

Questions to Explore

a. Interpret the effect of age on weight in the multiple regression equation.

b. In the population, does age help you to predict weight if you already know height and percent body fat? Show all steps of a significance test, and interpret.

Think It Through

a. Let $\hat{y}$ = predicted weight, x_1 = height, x_2 = %body fat, and x_3 = age. Then

$$\hat{y} = -97.7 + 3.43x_1 + 1.36x_2 - 0.96x_3.$$

The slope coefficient of age is −0.96. The sample effect of age on weight is negative, which may seem surprising, but in practical terms it is small: For athletes having fixed values of x_1 and x_2, the predicted weight decreases by only 0.96 pounds for a one-year increase in age, and the ages vary only from 17 to 23.

b. If $\beta_3 = 0$, then x_3 = age has *no* effect on weight in the population, controlling for height and body fat. The hypothesis that age does *not* help us better predict weight, if we already know height and body fat, is H_0: $\beta_3 = 0$. Here are the steps:

1. **Assumptions:** The 64 female athletes were a convenience sample, not a random sample. Although the goal was to make inferences about all female college athletes, inferences are tentative. We'll discuss the other assumptions and learn how to check them in Section 13.4.
2. **Hypotheses:** The null hypothesis is H_0: $\beta_3 = 0$. Since there's no prior prediction about whether the effect of age is positive or negative (for fixed values of x_1 and x_2), we use the two-sided H_a: $\beta_3 \neq 0$.
3. **Test statistic:** Table 13.10 reports a slope estimate of −0.960 for age and a standard error of $se = 0.648$. It also reports the t test statistic of

$$t = (b_3 - 0)/se = -0.960/0.648 = -1.48.$$

Since the sample size equals $n = 64$ and the regression equation has four parameters, the degrees of freedom are $df = n - 4 = 60$.

4. **P-value:** Table 13.10 reports P-value = 0.14. This is the two-tailed probability of a t statistic below −1.48 or above 1.48, if H_0 were true.
5. **Conclusion:** The P-value of 0.14 does not give much evidence against the null hypothesis that $\beta_3 = 0$. At common significance levels, such as 0.05, we cannot reject H_0. Age does not significantly predict weight if we already know height and percentage of body fat. These conclusions are tentative because the sample of 64 female athletes was selected using a convenience sample rather than a random sample.

Insight

By contrast, Table 13.10 shows that $t = 9.3$ for testing the effect of height (H_0: $\beta_1 = 0$) and $t = 4.4$ for testing the effect of %BF (H_0: $\beta_2 = 0$). Both P-values are 0.000. It *does* help to have each of these variables in the model, given the other two.

Try Exercise 13.19

Example 7

What's Plausible for the Effect of Age on Weight?

Picture the Scenario

For the college athletes data, consider the multiple regression analysis of y = weight and predictors x_1 = height, x_2 = %body fat, and x_3 = age.

Question to Explore

Find and interpret a 95% confidence interval for β_3, the effect of age while controlling for height and percent of body fat.

Think It Through

From the previous example, $df = 60$. For $df = 60$, $t_{.025} = 2.00$. From Table 13.10 (shown partly in the margin), the estimate of β_3 is -0.96, with $se = 0.648$. The 95% confidence interval equals

$$b_3 \pm t_{.025}(se) = -0.96 \pm 2.00(0.648),$$
$$\text{or } -0.96 \pm 1.30, \text{ roughly } (-2.3, 0.3).$$

At fixed values of x_1 and x_2, we infer that the population mean of weight changes very little (and may not change at all) for a one-year increase in age.

Insight

The confidence interval contains 0. Age may have *no* effect on weight, once we control for height and percent body fat. This is in agreement with not rejecting H_0: $\beta_3 = 0$ in favor of H_a: $\beta_3 \neq 0$ at the $\alpha = 0.05$ level in the significance test.

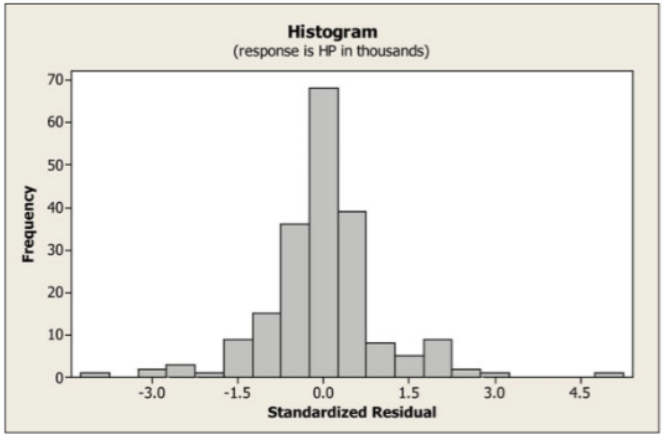
Try Exercise 13.20

Example 8

House Selling Price

Picture the Scenario

For the House Selling Price OR data set (Examples 1–3), Figure 13.4 is a MINITAB histogram of the standardized residuals for the multiple regression model predicting selling price by the house size and the number of bedrooms.



▲ **Figure 13.4** Histogram of Standardized Residuals for Multiple Regression Model Predicting Selling Price. **Question** Give an example of a shape for this histogram that would indicate that a few observations are highly unusual.

Question to Explore

What does Figure 13.4 tell you? Interpret.

Think It Through

The residuals are roughly bell shaped about 0. They fall mostly between about -3 and $+3$. The shape suggests that the conditional distribution of the response variable may have a bit of skew, but no severe nonnormality is indicated.

Insight

When n is small, don't make the mistake of reading too much into such plots. We're mainly looking for dramatic departures from the assumptions and highly unusual observations that stand apart from the others.

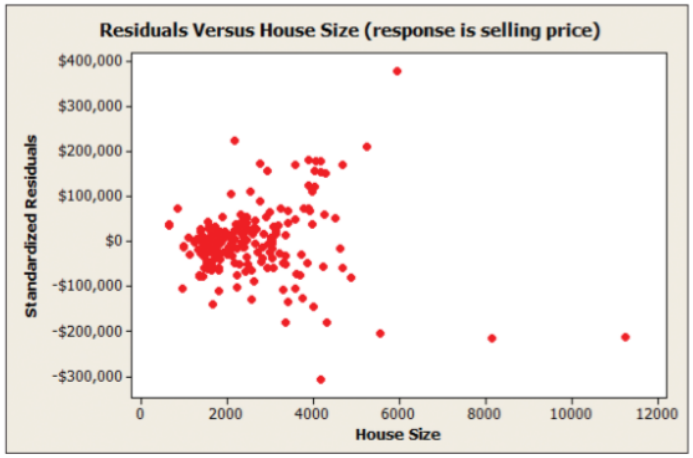
Try Exercise 13.31

Example 9

House Selling Price

Picture the Scenario

For the House Selling Price OR data set, Figure 13.6 is a residual plot for the multiple regression model relating selling price to house size and to number of bedrooms. It plots the standardized residuals against house size.



▲ **Figure 13.6** Standardized Residuals of Selling Price Plotted Against House Size, for Model With House Size and Number of Bedrooms as Predictors. Questions How does this plot suggest that selling price has more variability at higher house size values, for given number of bedrooms? You don't see number of bedrooms on the plot, so how do its values affect the analysis?

Question to Explore

Does this plot suggest any irregularities with the model?

Think It Through

As house size increases, the variability of the standardized residuals seems to increase. This suggests more variability in selling prices when house size is larger, for a given number of bedrooms. We must be cautious, though, because the few points with large negative residuals for the largest houses and the one point with a large positive residual may catch our eyes more than the others. Generally, it's not a good idea to allow a few points to overly influence your judgment about the shape of a residual pattern. However, there is definite evidence that the variability increases as house size increases. A larger data set would provide more evidence about this.

Insight

Nonconstant variability does not invalidate the use of multiple regression. It would, however, make us cautious about using prediction intervals. We would expect predictions about selling price to have smaller prediction errors when house size is small than when house size is large.

Try Exercise 13.32

Example 10

Including Condition in Regression for House Selling Price

Picture the Scenario

Let's now fit a regression model for y = selling price of home using x_1 = house size and x_2 = condition of the house. Table 13.11 shows MINITAB output.

Table 13.11 Regression Analysis of y = Selling Price Using x_1 = House Size and x_2 = Indicator Variable for Condition (Good, Not Good)

```
The regression equation is

House Price = 96271 + 66.5 House Size + 12927 Condition

Predictor     Coef      SE Coef       T       P

Constant      96271       13465     7.15   0.000

House Size   66.463        4.682    14.20   0.000

Condition     12927       17197     0.75   0.453

S = 81787.4 R-Sq = 50.6%
```

Questions to Explore

a. Find and plot the lines showing how predicted selling price varies as a function of house size, for homes in good condition or not in good condition.

b. Interpret the coefficient of the indicator variable for condition.

Think It Through

a. Table 13.11 reports the prediction equation,

$$\hat{y} = 96{,}271 + 66.5x_1 + 12{,}927x_2.$$

For homes not in good condition, $x_2 = 0$. The prediction equation for y = selling price using x_1 = house size then simplifies to

$$\hat{y} = 96{,}271 + 66.5x_1 + 12{,}927(0) = 96{,}271 + 66.5x_1.$$

For homes in good condition, $x_2 = 1$. The prediction equation then simplifies to

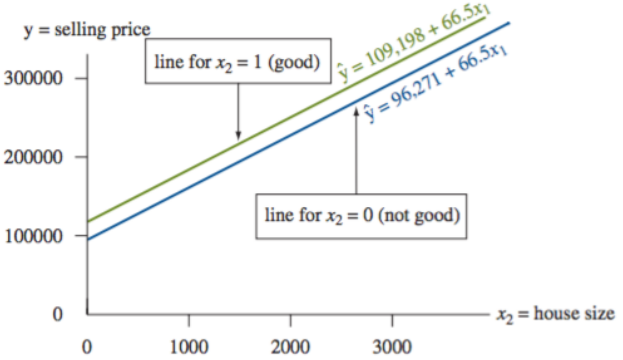
$$\hat{y} = 96{,}271 + 66.5x_1 + 12{,}927(1) = 109{,}198 + 66.5x_1.$$

Both lines have the same slope, 66.5. For homes in good condition or not in good condition, the predicted selling price increases by $66.5 for each square-foot increase in house size. Figure 13.7 plots the two prediction equations. The quantitative explanatory variable, house size, is on the x-axis. The figure portrays a separate line for each category of condition (good, other).

b. At a fixed value of x_1 = house size, the difference between the predicted selling prices for homes in good (1) versus not good (0) condition is

$$(109{,}198 + 66.5x_1) - (96{,}271 + 66.5x_1) = 12{,}927.$$

This is precisely the coefficient of the indicator variable, x_2. For any fixed value of house size, we predict that the selling price is \$12,927 higher for homes that are good versus not in good condition.



▲ **Figure 13.7** Plot of Equation Relating $\hat{y}$ = Predicted Selling Price to x_1 = House Size, According to x_2 = Condition (1 = Good, 0 = Not Good). **Question** Why are the lines parallel?

Insight

Since the two lines have the same slope, they are parallel. The line for homes in good condition is above the other line because its y-intercept is larger. This means that for any fixed value of house size, the predicted selling price is higher for homes in good condition. The P-value of 0.453 for the test for the coefficient of the indicator variable suggests that this difference is not statistically significant.

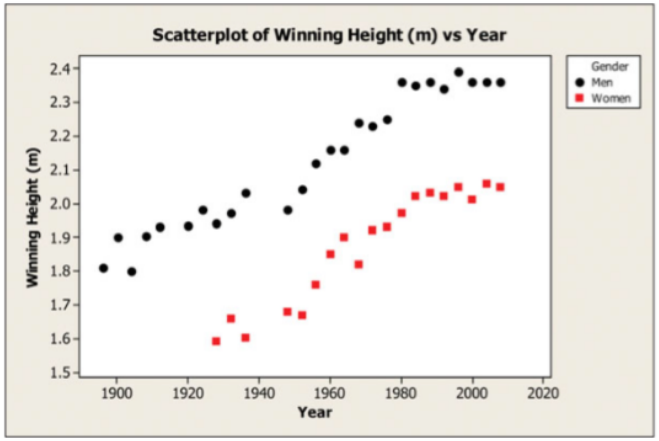
Try Exercise 13.40

Example 11

Comparing Winning High Jumps for Men and Women

Picture the Scenario

Men have competed in the high jump in the Olympics since 1896 and women since 1928. Figure 13.9 shows how the winning high jump in the Olympics has changed over time for men and women. The High Jump data file on the text CD contains the winning heights for each year. A multiple regression analysis of y = winning height (in meters) as a function of x_1 = number of years since 1928 (when women first participated in the high jump) and x_2 = gender (1 = male, 0 = female) gives $\hat{y} = 1.63 + 0.0057x_1 + 0.35x_2$.



▲ **Figure 13.9** Scatterplot for the Winning High Jumps (in Meters) in the Olympic Games from 1896–2008.

Questions to Explore

a. Interpret the coefficient of year and the coefficient of gender in the equation.

b. To allow interaction, we can fit equations separately to the data for males and the data for females. We then get $\hat{y} = 1.98 + 0.0055x_1$ for males and $\hat{y} = 1.60 + 0.0065x_1$ for females. Describe the interaction by comparing slopes.

c. Describe the interaction allowed in part b by comparing predicted winning high jumps for males and females in 1928 and in 2008.

Think It Through

a. For the prediction equation $\hat{y} = 1.63 + 0.0057x_1 + 0.35x_2$, the coefficient of year is 0.0057. For each gender, the predicted winning high jump increases by 0.0057 meters per year. This seems small, but over a hundred years it projects to an increase of $100(0.0057) = 0.57$ meters, about 27 inches. The model does not allow interaction, as it assumes that the slope of 0.0057 is the same for each gender. The coefficient of gender is 0.35. In a given year, the predicted winning high jump for men is 0.35 meters higher than for women. Because this model does not allow interaction, the predicted difference between men and women is the same each year.

b. The slope of 0.0065 for females is higher than the slope of 0.0055 for males. So the predicted winning high jump increases a bit more for females than for males over this time period.

a difference of 0.30 meters. The predicted difference between the winning high jumps of males and females decreased a bit between 1928 and 2008.

Insight

When we allow interaction, the estimated slope is a bit higher for females than for males. This is what caused the difference in predicted winning high jumps to be less in 2008 than in 1928. However, the slopes were not dramatically different, as Figure 13.9 shows that the points go up at similar rates for the two genders. The sample degree of interaction was not strong.

Try Exercises 13.46 and 13.47

Example 12

Travel Credit Cards

Picture the Scenario

An Italian study with 100 randomly selected Italian adults considered factors associated with whether a person has at least one travel credit card. Table 13.12 shows results for the first 15 people on this response variable and on the person's annual income, in thousands of euros.[2] The complete data set is in the Credit Card and Income data file on the text CD. Let x = annual income and let y = whether the person has a travel credit card (1 = yes, 0 = no).

Table 13.12 Annual Income (in thousands of euros) and Whether Person Has a Travel Credit Card

The response y equals 1 if a person has a travel credit card and equals 0 otherwise. The complete data set is on the text CD.

Income	y	Income	y	Income	y
12	0	14	1	15	0
13	0	14	0	15	1
14	1	14	0	15	0
14	0	14	0	15	0
14	0	14	0	15	0

Source: Data from R. Piccarreta, Bocconi University, Milan (personal communication).

Questions to Explore

Table 13.13 shows what software provides for conducting a logistic regression analysis.

Table 13.13 Results of Logistic Regression for Italian Credit Card Data

Predictor	Coef	SE Coef	Z	P
Constant	−3.5180	0.71034	−4.95	0.000
income	0.1054	0.02616	4.03	0.000

a. State the prediction equation for the probability of owning a travel credit card, and explain whether annual income has a positive or a negative effect.

b. Find the estimated probability of having a travel credit card at the lowest and highest annual income levels in the sample, which were $x = 12$ and $x = 65$.

[2]The data were originally recorded in Italian lira but have been changed to euros and adjusted for inflation.

Think It Through

a. Substituting the α and β estimates from Table 13.13 into the logistic regression model formula, $p = e^{\alpha+\beta x}/(1 + e^{\alpha+\beta x})$, we get the equation for the estimated probability $\hat{p}$ of having a travel credit card,

$$\hat{p} = \frac{e^{-3.52+0.105x}}{1 + e^{-3.52+0.105x}}.$$

Because the estimate 0.105 of β (the coefficient of x) is positive, this sample suggests that annual income has a positive effect: The estimated probability of having a travel credit card is higher at higher levels of annual income.

b. For subjects with income $x = 12$ thousand euros, the estimated probability of having a travel credit card equals

$$\hat{p} = \frac{e^{-3.52+0.105(12)}}{1 + e^{-3.52+0.105(12)}} = \frac{e^{-2.26}}{1 + e^{-2.26}} = \frac{0.104}{1.104} = 0.09.$$

For $x = 65$, the highest income level in this sample, you can check that the estimated probability equals 0.97.

Insight

There is a strong effect. The estimated probability of having a travel credit card changes from 0.09 to 0.97 (nearly 1.0) as annual income changes over its range.

Using software, we could also fit the straight-line regression model. Its prediction equation is $\hat{p} = -0.159 + 0.0188x$. However, its $\hat{p}$ predictions are quite different at the low end and at the high end of the annual income scale. At $x = 65$, for instance, it provides the prediction $\hat{p} = 1.06$. This is a poor prediction because we know that a proportion must fall between 0 and 1.

Try Exercise 13.50

Example 13

Effect of Income on Credit Card Use

Picture the Scenario

For the Italian travel credit card study, Example 12 found that the equation

$$\hat{p} = \frac{e^{-3.52+0.105x}}{1 + e^{-3.52+0.105x}}$$

estimates the probability p of having such a credit card as a function of annual income x.

Questions to Explore

Interpret this equation by (a) finding the income value at which the estimated probability of having a credit card equals 0.50 and (b) finding the approximate rate of change in the probability at that income value.

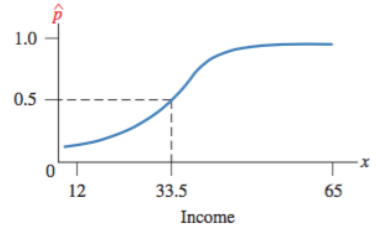
Think It Through

a. The estimate of α is -3.52 and the estimate of β is 0.105. Substituting the estimates into the expression $-\alpha/\beta$ for the value of x at $p = 0.50$, we get $x = (3.52)/(0.105) = 33.5$. The estimated probability of having a travel credit card equals 0.50 when annual income equals €33,500.

b. A line drawn tangent to the logistic regression curve at the point where $p = 0.50$ has slope equal to $\beta/4$. The estimate of this slope is $0.105/4 = 0.026$. For each increase of €1000 in annual income near the income value of €33,500, the estimated probability of having a travel credit card increases by approximately 0.026.

Insight

Figure 13.12 shows the estimated logistic regression curve, highlighting what we've learned in Examples 12 and 13: The estimated probability increases from 0.09 to 0.97 between the minimum and maximum income values, and it equals 0.50 at an annual income of €33,500.



▲ **Figure 13.12** Logistic Regression Curve Relating Estimated Probability $\hat{p}$ of Having a Travel Credit Card to Annual Income x.

Try Exercise 13.54

Example 14

Estimating Proportion of Students Who've Used Marijuana

Picture the Scenario

Table 13.14 is a three-variable contingency table from a Wright State University survey asking senior high-school students near Dayton, Ohio, whether they had ever used alcohol, cigarettes, or marijuana. We'll treat marijuana use as the response variable and cigarette use and alcohol use as explanatory variables.

Table 13.14 Alcohol, Cigarette, and Marijuana Use for High School Seniors

		Marijuana Use	
Alcohol Use	**Cigarette Use**	**Yes**	**No**
Yes	Yes	911	538
	No	44	456
No	Yes	3	43
	No	2	279

Source: Data from Professor Harry Khamis, Wright State University (personal communication).

Questions to Explore

Let y indicate marijuana use, coded 1 = yes, 0 = no. Let x_1 be an indicator variable for alcohol use (1 = yes, 0 = no), and let x_2 be an indicator variable for cigarette use (1 = yes, 0 = no). Table 13.15 shows MINITAB output for a logistic regression model.

Table 13.15 MINITAB Output for Estimating the Probability of Marijuana Use Based on Alcohol Use and Cigarette Use

Predictor	Coef	SE Coef	Z	P
Constant	−5.30904	0.475190	−11.17	0.000
alcohol	2.98601	0.464671	6.43	0.000
cigarettes	2.84789	0.163839	17.38	0.000

a. Report the prediction equation, and interpret.

b. Find the estimated probability $\hat{p}$ of having used marijuana (i) for students who have not used alcohol or cigarettes and (ii) for students who have used both alcohol and cigarettes.

The coefficient of alcohol use (x_1) is positive (2.99). The indicator variable x_1 equals 1 for those who've used alcohol. Thus, the alcohol users have a higher estimated probability of using marijuana, controlling for whether they used cigarettes. Likewise, the coefficient of cigarette use (x_2) is positive (2.85), so the cigarette users have a higher estimated probability of using marijuana, controlling for whether they used alcohol. Table 13.15 tells us that for each predictor, the test statistic is large. In other words, the estimated effect is a large number of standard errors from 0. The P-values are both 0.000, so there is strong evidence that the corresponding population effects are positive also.

b. For those who have not used alcohol or cigarettes, $x_1 = x_2 = 0$. For them, the estimated probability of marijuana use is

$$\hat{p} = \frac{e^{-5.31+2.99(0)+2.85(0)}}{1 + e^{-5.31+2.99(0)+2.85(0)}} = \frac{e^{-5.31}}{1 + e^{-5.31}} = \frac{0.0049}{1.0049} = 0.005.$$

For those who have used alcohol and cigarettes, $x_1 = x_2 = 1$. For them, the estimated probability of marijuana use is

$$\hat{p} = \frac{e^{-5.31+2.99(1)+2.85(1)}}{1 + e^{-5.31+2.99(1)+2.85(1)}} = 0.629.$$

In summary, the probability that students have tried marijuana seems highly related on whether they've used alcohol and cigarettes.

Insight

Likewise, you can find the estimated probability of using marijuana for those who have used alcohol but not cigarettes (let $x_1 = 1$ and $x_2 = 0$) and for those who have not used alcohol but have used cigarettes. Table 13.16 summarizes results. We see that marijuana use is unlikely unless a student has used both alcohol and cigarettes.

Table 13.16 Estimated Probability of Marijuana Use, by Alcohol Use and Cigarette Use, Based on Logistic Regression Model

The sample proportions of marijuana use are shown in parentheses for the four cases.

	Cigarette Use	
Alcohol Use	**Yes**	**No**
Yes	0.629 (0.629)	0.089 (0.088)
No	0.079 (0.065)	0.005 (0.007)

Try Exercises 13.56 and 13.57