mikeledray/Shutterstock

CHAPTER

29

Multiple Regression*

When a scatterplot shows a linear relationship between a quantitative explanatory variable x and a quantitative response variable y, we fit a regression line to the data to describe the relationship. We can also use the line to predict the value of y for a given value of x. For example, Chapter 5 uses regression lines to describe relationships between

- Fat gain y and nonexercise activity x.
- The brain activity y of women when their partner has a painful experience and their score x on a test measuring empathy.
- The number y of new adults that join a colony of birds and the percent x of adult birds that return from the previous year.

In all these cases, other explanatory variables might improve our understanding of the response y and help us to better predict y:

- Fat gain y depends on nonexercise activity x_1, time spent daily in exercise activity x_2, and sex x_3.
- A woman's brain activity y when her partner has a painful experience may depend on her score x_1 on a test of empathy and also on her score x_2 on a test of emotional attachment to her partner.
- The number y of new adults in a bird colony depends on the percent x_1 of returning adults and also on the species x_2 of birds we study.

In this chapter, we cover...

*The original version of this chapter was written by Professor Bradley Hartlaub of Kenyon College.

simple linear regression
multiple regression

We will now call regression with just one explanatory variable **simple linear regression** to remind us that this is a special case. This chapter introduces the more general case of **multiple regression**, which allows several explanatory variables to combine in explaining a response variable. The material we discuss will help you understand and interpret the results of a multiple regression analysis. However, there are many issues that we do not discuss that are important to understand if you plan to carry out a multiple regression analysis yourself. Thus, we recommend you take a more advanced course on multiple regression if you plan to use these methods.

29.1 Adding a Categorical Variable in Regression

In Chapter 4, we learned how to add a categorical variable to a scatterplot by using different colors or plot symbols to indicate the different values of the categorical variable. Consider a simple case: the categorical variable (call it x_2) takes just two values. We want to explore the effect of both a quantitative variable (call it x_1) and x_2 on a response y. Here is an example.

EXAMPLE 29.1 Gas Mileage

MPG

STATE: The gas mileage of motor vehicles depends on many factors. One of these is the size of the engine. Engine size is reported as engine displacement, which is the swept volume of all the pistons inside the cylinders of the engine in a single movement from top dead center to bottom dead center. One would expect larger engine sizes to be associated with lower gas mileages. To explore this, we examine data from a random sample of 48 vehicles from model year 2016.[1]

Table 29.1 shows the combined city and highway gas mileage in miles per gallon (MPG), the engine displacement in liters, and the type of vehicle (car or truck) for

TABLE 29.1 Gas mileage, engine displacement, and vehicle type for 48 vehicles from model year 2016

MPG	Displacement	Type	MPG	Displacement	Type	MPG	Displacement	Type	MPG	Displacement	Type
20	3.0	Car	21	3.0	Car	17	4.0	Truck	17	6.2	Truck
18	3.8	Car	21	4.0	Car	19	5.3	Truck	16	6.2	Truck
24	3.4	Car	23	3.0	Car	17	5.7	Truck	16	3.5	Truck
31	1.5	Car	22	2.0	Car	22	2.5	Truck	23	3.0	Truck
19	5.0	Car	32	1.8	Car	22	2.0	Truck	14	5.7	Truck
15	4.7	Car	29	2.5	Car	20	3.6	Truck	24	3.0	Truck
16	5.4	Car	32	1.8	Car	20	3.6	Truck			
28	2.4	Car	22	3.0	Car	24	2.0	Truck			
26	2.0	Car	18	5.0	Car	25	2.0	Truck			
21	3.0	Car	18	4.8	Car	22	3.5	Truck			
15	6.0	Car	26	2.0	Car	23	2.0	Truck			
30	2.0	Car	20	3.7	Car	29	1.8	Truck			
29	1.6	Car	27	1.6	Car	21	2.4	Truck			
26	1.6	Car	28	2.0	Car	26	2.4	Truck			

the 48 vehicles. Trucks (which include SUVs) differ in a variety of ways from cars and are often characterized as less fuel efficient. Because trucks tend to have larger engines in order haul heavy loads, this difference in gas mileage may simply be due to engine displacement. Do the data provide any evidence that this is not the case? If so, this suggests that there are features of trucks, other than engine displacement, that are responsible for the difference in gas mileage performance.

PLAN: Make a scatterplot to display the relationship MPG y and engine displacement x_1. Use different colors for the two vehicle types. (Type is a categorical variable x_2 that takes two values.) If both vehicle types show linear patterns, fit two separate least-squares regression lines to describe them.

SOLVE: Figure 29.1 shows a scatterplot with two different plotting symbols, one for cars and one for trucks. Both vehicle types show a linear pattern with larger engine displacement associated with lower MPG. Because the points corresponding to cars and trucks are interspersed, it is difficult to tell if the pattern is different for both. If we look closely, we notice that among the vehicles with the smallest engine displacements (below 3), those with the highest MPG are almost all cars, and the majority of those with the lowest MPG (below 17.5) are trucks. However, if we look at vehicles with the largest engine displacements (above 5), the pattern is less clear, especially because there is only a single car with such a large engine displacement. Closer examination suggests that there may be differences in gas mileages of cars and trucks that cannot simply be explained by engine displacement, but it is also possible that these observed differences are simply due to chance. Would we see these same differences in another sample? We will soon learn how to formally estimate parameters and make inferences for two regression lines. This will help us assess whether the differences in the regression lines can be attributed simply to chance.

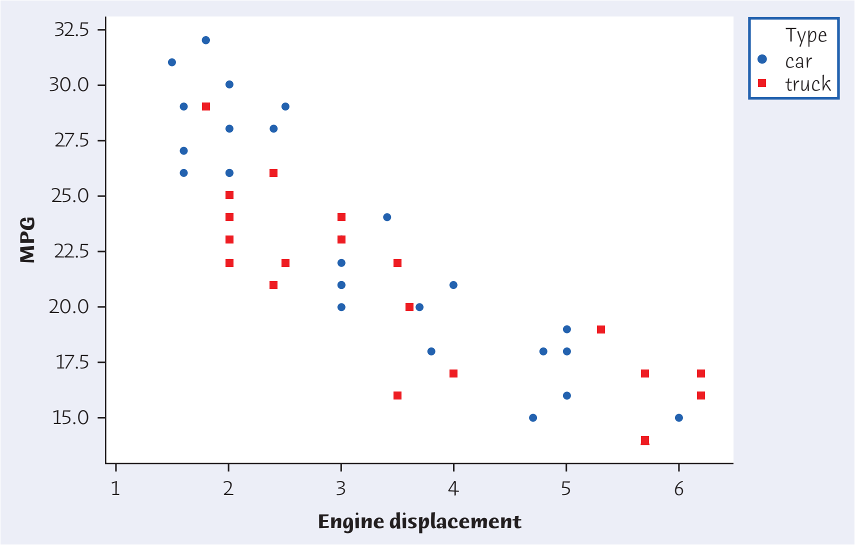


FIGURE 29.1 A scatterplot of MPG for a sample of 48 vehicles from model year 2016, for Example 29.1.

Because we see that the relationship between gas mileage and engine displacement may be different for cars and trucks, we would like to have a single regression model that allows us to explore this insight. To do this, introduce a second explanatory variable x_2 for "vehicle type." Unfortunately, vehicle type is not numerical. To solve this problem, we use values 0 and 1 to distinguish the two vehicle types. Now we have an *indicator variable*

$$x_2 = 0 \text{ for "car"}$$

$$x_2 = 1 \text{ for "truck"}$$

Indicator Variable

An **indicator variable** places individuals into one of two categories, usually coded by the two values 0 and 1.

Indicator variables are commonly used to indicate sex (0 = male, 1 = female), condition of patient (0 = good, 1 = poor), status of order (0 = undelivered, 1 = delivered), and many other characteristics for individuals.

The conditions for inference in simple linear regression (Chapter 26, text page 603) describe the relationship between the explanatory variable x and the mean response μ_y in the population by a *population regression line* $\mu_y = \beta_0 + \beta_1 x$. The switch in notation from $\mu_y = \alpha + \beta x$ allows an easier extension to other models. Suppose we add a second explanatory variable, so that our regression model for the population becomes

$$\mu_y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

The other conditions for inference are the same as in the simple linear regression setting: for any fixed values of the explanatory variables, y varies about its mean according to a Normal distribution with unknown standard deviation σ that is the same for all values of x_1 and x_2. We will look in detail at conditions for inference in multiple regression later on.

EXAMPLE 29.2 Interpreting a Multiple Regression Model

Multiple regression models are no longer simple straight lines, so we must think a bit harder to interpret what they say. Consider our model

$$\mu_y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

in which y is the gas mileage, x_1 is the engine displacement, and x_2 is an indicator variable for vehicle type. For cars, $x_2 = 0$ and the model becomes

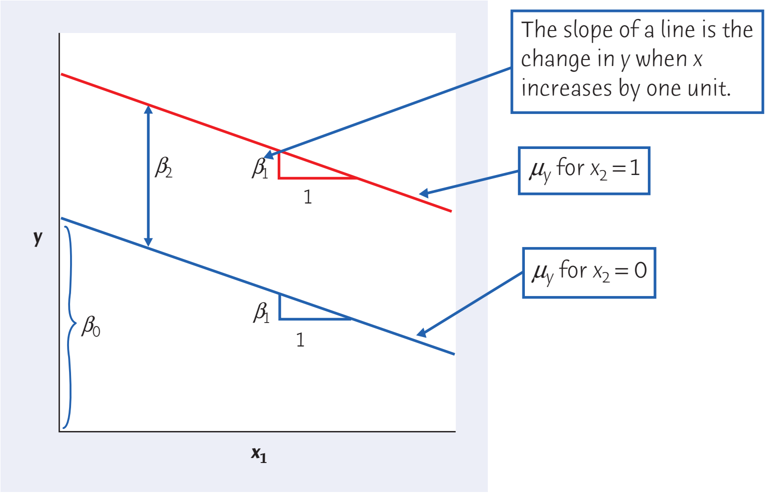
$$\mu_y = \beta_0 + \beta_1 x_1$$

For trucks, $x_2 = 1$ and the model is

$$\begin{aligned}\mu_y &= \beta_0 + \beta_1 x_1 + \beta_2 \\ &= (\beta_0 + \beta_2) + \beta_1 x_1\end{aligned}$$

Look carefully: the slope that describes how the mean gas mileage changes as the engine displacement x_1 varies is β_1 in both models. The intercepts differ: β_0 for cars and $\beta_0 + \beta_2$ for trucks. So β_2 represents a fixed change between cars and trucks. Figure 29.2 is a graph of this model with all three β's identified. By adding the indicator variable x_2 for vehicle type, we have produced a regression model for two *parallel* straight lines.

You will sometimes see indicator variables referred to as *dummy variables*. We have demonstrated how indicator (dummy) variables can be used to represent a categorical variable with two categories. More advanced books on multiple regression discuss how multiple indicator (dummy) variables can be used to represent categorical variables with more than two categories.



The slope of a line is the change in y when x increases by one unit.

FIGURE 29.2 Multiple regression model with two parallel straight lines, for Example 29.2.

Macmillan Learning Online Resources

- The Whiteboard video, *An Overview of Multiple Regession*, discusses the goals of multiple regression and their importance.

APPLY YOUR KNOWLEDGE

29.1 Bird Colonies. Suppose (this is too simple to be realistic) that the number y of new birds that join a colony this year has the same straight-line relationship with the percent x_1 of returning birds in colonies of two different bird species. An indicator variable shows which species we observe: $x_2 = 0$ for one and $x_2 = 1$ for the other. Write a population regression model that describes this setting. Explain in words what each β in your model means.

29.2 How Fast Do Icicles Grow? We have data on the growth of icicles starting at length 10 centimeters (cm) and at length 20 cm. An icicle grows at the same rate, 0.15 cm per minute, starting from either length. Give a population regression model that describes how mean length changes with time x_1 and starting length x_2. Use numbers, not symbols, for the β's in your model.

29.2 Estimating Parameters

How shall we estimate the β's in the model $\mu_y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$? Because we hope to predict y, we want to make the errors in the y direction small. We can't call this the vertical distance from the points to a line as we did for a simple linear regression model because we now have two lines. But we still concentrate on the prediction of y and, therefore, on the deviations between the observed responses y and the responses predicted by the regression model.

The method of least squares estimates the β's in the model by choosing the values that minimize the sum of the squared deviations in the y direction,

$$\Sigma(\text{observed } y - \text{predicted } y)^2 = \Sigma(y - \hat{y})^2$$

Call the values of the β's that do this b's. The least least-squares regression model $\hat{y} = b_0 + b_1 x_1 + b_2 x_2$ estimates the population regression model $\mu_y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$.

The remaining parameter is the standard deviation σ, which describes the variability of the response y about the mean given by the population regression model. Recall that the **residuals** are the differences between the observed responses y and the responses $\hat{y}$ predicted by the least-squares model. Because the residuals estimate the "left-over variation" about the regression model, the standard deviation s of the residuals is used to estimate σ. The value of s is also referred to as the *regression standard error*.

residuals

Statistics In Your World

Why Some Men Earn More

Research based on data from the U.S. Bureau of Labor Statistics and the U.S. Census Bureau suggests that women earn 80 cents for every dollar men earn. Although the literature is full of clear and convincing cases of discrimination based on height, weight, race, sex, and religion, new studies suggest that our choices explain a considerable amount of the variation in wages. Earning more often means that you are willing to accept longer commuting times, safety risks, frequent travel, long hours, and other responsibilities that take away from your time at home with family and friends. When choosing between time and money, make sure that you are happy with your choice!

Regression Standard Error

The regression standard error for the multiple regression model $\hat{y} = b_0 + b_1x_1 + b_2x_2$ is

$$s = \sqrt{\frac{1}{n-3}\Sigma \text{ residual}^2}$$

$$= \sqrt{\frac{1}{n-3}\Sigma(y-\hat{y})^2}$$

Use s to estimate the standard deviation σ of the responses about the mean given by the population regression model.

Notice that instead of dividing by $(n - 2)$, the number of observations less 2, as we did for the simple linear regression model in Chapter 25, we are now dividing by $(n - 3)$, the number of observations less 3. Because we are estimating three β parameters in our population regression model, the degrees of freedom must reflect this change. In general, the **degrees of freedom** for the regression standard error will be the number of data points minus the number of parameters in the population regression model.

degrees of freedom

EXAMPLE 29.3 Gas Mileage, continued

DATA MPG2

Example 29.2 introduced the regression model

$$\mu_y = \beta_0 + \beta_1x_1 + \beta_2x_2$$

for predicting MPG y from engine displacement x_1 and vehicle type x_2. Statistical software gives the least-squares estimate of this model as

$$\hat{y} = 32.077 - 2.830x_1 - 1.264x_2$$

By substituting the two values of the indicator variable into this estimated regression equation, we can obtain a least-squares line for each vehicle type. The predicted MPGs are

$$\hat{y} = 30.813 - 2.830x_1 \text{ for trucks } (x_2 = 1)$$

and

$$\hat{y} = 32.077 - 2.830x_1 \text{ for cars } (x_2 = 0)$$

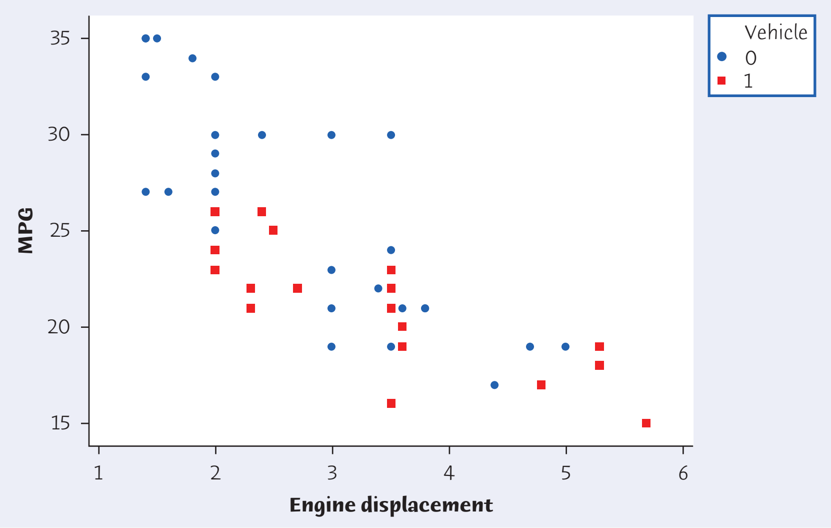
These two least-squares lines have the same slope and thus are parallel. The estimate $b_2 = -1.264$ tells us that using this regression model we would predict trucks, on average, to get a 1.264 lower MPG than cars for all values of engine displacement. The regression standard error $s = 2.599$ indicates the size of the "typical" error. We would expect approximately 95% of the reporting percents of all vehicles of the same engine size and vehicle type (the same values of x_1 and x_2) to be within $2 \times 2.599 = 5.198$ of the mean predicted reporting percent for that engine size and vehicle type.

APPLY YOUR KNOWLEDGE

29.3 Gas Mileage, continued. Table 29.2 provides data on gas mileage, engine displacement, whether the vehicle is turbocharged, and vehicle type from a second random sample of 48 vehicles from model year 2016. Descriptive statistics and a scatterplot for gas mileages for cars and trucks from Table 29.2 accompany this exercise. MPG3

TABLE 29.2 Gas mileage, engine displacement, turbocharged, and vehicle type for second sample of 48 vehicles from model year 2016

MPG	Displacement	Turbocharged	Type	MPG	Displacement	Turbocharged	Type
27	1.4	Yes	Car	20	3.6	No	Truck
22	3.4	No	Car	19	5.3	No	Truck
21	3.8	No	Car	15	5.7	No	Truck
21	3.8	No	Car	23	3.5	No	Truck
23	3.0	Yes	Car	22	2.7	No	Truck
35	1.4	No	Car	24	2.0	Yes	Truck
19	5.0	No	Car	23	2.0	Yes	Truck
19	3.0	Yes	Car	22	3.5	No	Truck
17	4.4	Yes	Car	26	2.4	No	Truck
35	1.4	Yes	Car	26	2.0	No	Truck
33	1.4	Yes	Car	21	3.5	No	Truck
30	2.0	No	Car	19	3.6	Yes	Truck
35	1.5	No	Car	25	2.5	No	Truck
30	3.5	No	Car	22	2.3	Yes	Truck
30	3.0	Yes	Car	18	5.3	No	Truck
30	2.4	No	Car	21	2.3	Yes	Truck
28	2.0	No	Car	16	3.5	Yes	Truck
24	3.5	No	Car	17	4.8	Yes	Truck
33	2.0	No	Car				
23	3.5	No	Car				
27	2.0	Yes	Car				
34	1.8	No	Car				
21	3.0	Yes	Car				
21	3.6	No	Car				
19	3.5	No	Car				
23	2.0	Yes	Car				
19	4.7	Yes	Car				
27	1.6	Yes	Car				
25	2.0	Yes	Car				
29	2.0	Yes	Car				



Descriptive Statistics: MPG-car, Engine Displacement-car, MPG-truck, Engine Displacement-truck

```
Variable                         Mean   StdDev
MPG-car                         26.00    5.60
Engine Displacement-car          2.720   1.059
MPG-truck                       21.056   3.262
Engine Displacement-truck        3.361   1.217
```

Correlations: MPG-car, Engine Displacement-car

```
Pearson correlation of MPG-car and Engine Displacement-car = -0.794
```

Correlations: MPG-truck, Engine Displacement-truck

```
Pearson correlation of MPG-truck and Engine Displacement -truck = -0.80
```

(a) Use the descriptive statistics to compute the least-squares regression line for predicting MPG from engine displacement for cars.

(b) Use the descriptive statistics to compute the least-squares regression line for predicting MPG from engine displacement for trucks.

(c) Interpret the value of the slope for each of your estimated models.

(d) Would you be willing to use the multiple regression model with equal slopes to predict MPG for cars and trucks? Explain why or why not.

29.4 Gas Mileage, continued. In Example 29.3, the indicator variable for vehicle type ($x_2 = 0$ for cars and $x_2 = 1$ for trucks) was used to combine the two separate regression models from Example 29.1 into one multiple regression model. Suppose that instead of x_2, we use an indicator variable x_3 that reverses the two types so that $x_3 = 1$ for cars and $x_3 = 0$ for trucks. The mean MPG is $\mu_y = \beta_0 + \beta_1 x_1 + \beta_2 x_3$, where x_1 is engine displacement (the value on the x axis in Figure 29.1) and x_3 is an indicator variable to identify the vehicle type (different symbols in Figure 29.1). Statistical software now gives the estimated regression model as $\hat{y} = 30.813 - 2.830x_1 + 1.264x_3$.

(a) Substitute the two values of the indicator variable into the estimated regression equation to obtain a least-squares line for each vehicle type.

(b) How do your estimated regression lines in part (a) compare with the estimated regression lines provided for each vehicle type in Example 29.3?

(c) Will the regression standard error change when this new indicator variable is used? Explain.

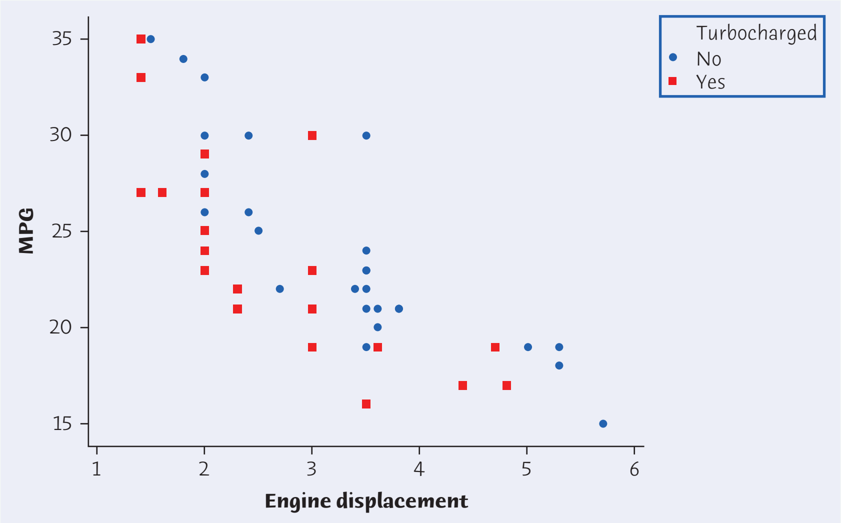
29.5 Gas Mileage, continued. Descriptive statistics and a scatterplot for MPG for nonturbocharged and turbocharged vehicles from Table 29.2 accompany this exercise.

(a) Use the descriptive statistics to compute the least-squares regression line for predicting MPG from engine displacement for nonturbocharged vehicles.

(b) Use the descriptive statistics to compute the least-squares regression line for predicting MPG from engine displacement for turbocharged vehicles.

(c) Interpret the value of the slope for each of your estimated models.

(d) Would you be willing to use the multiple regression model with equal slopes to predict MPG for nonturbocharged and turbocharged vehicles? Explain why or why not.



Descriptive Statistics: MPG -Turbo No, Engine Displacement-Turbo No, MPG-Turbo Yes, Engine Displacement-Turbo Yes

```
Variable          Mean  StdDev
MPG-Turbo No     24.52   5.57
ED-Turbo No      3.211  1.175
MPG-Turbo Yes    23.67   5.25
ED-Turbo Yes     2.638  1.062
```

Correlations: MPG-Turbo No, Engine Displacement-Turbo No

```
Pearson correlation of MPG-Turbo No and ED-Turbo No = -0.865
```

Correlations: MPG-Turbo Yes, Engine Displacement-Turbo Yes

```
Pearson correlation of MPG-Turbo Yes and ED-Turbo Yes = -0.762
```

29.3 Examples of Technology

Table 29.1 provides a compact way to display data in a textbook, but this is not the best way to enter your data into a statistical software package for analysis. The usual format for data files is that each row contains data on one individual, and each column contains the values of one variable.

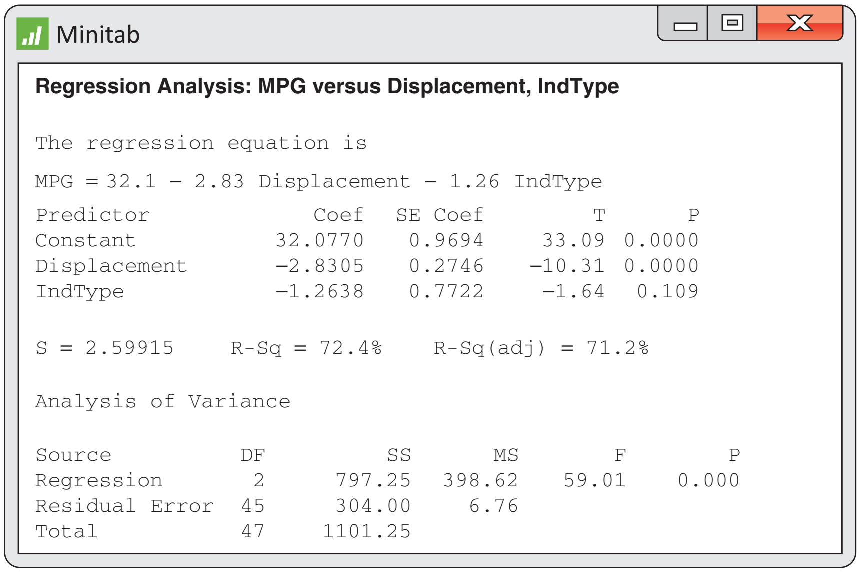
EXAMPLE 29.4 Organizing Data

The multiple regression model in Example 29.3 requires three columns. The 48 MPGs y for cars and trucks appear in a column labeled *MPG*, values of the explanatory variable x_1 make up a column labeled *Displacement*, and values of the indicator variable x_2 make up a column labeled *IndType*. The first five rows of the worksheet are shown here.

Row	MPG	Displacement	IndType
1	20	3.0	0
2	18	3.8	0
3	24	3.4	0
4	31	1.5	0
5	19	5.0	0

To use statistical software, we need only identify the response variable *MPG* and the two explanatory variables *Displacement* and *IndType*. Figure 29.3 shows the regression output from Minitab, CrunchIt!, and JMP. Each package provides parameter estimates, standard errors, t statistics, P-values, the regression standard error, and R^2. Minitab and JMP also provide an analysis of variance table. We will digest this output one piece at a time: first describe the model, then look at the conditions needed for inference, and finally interpret the results of inference.

Minitab



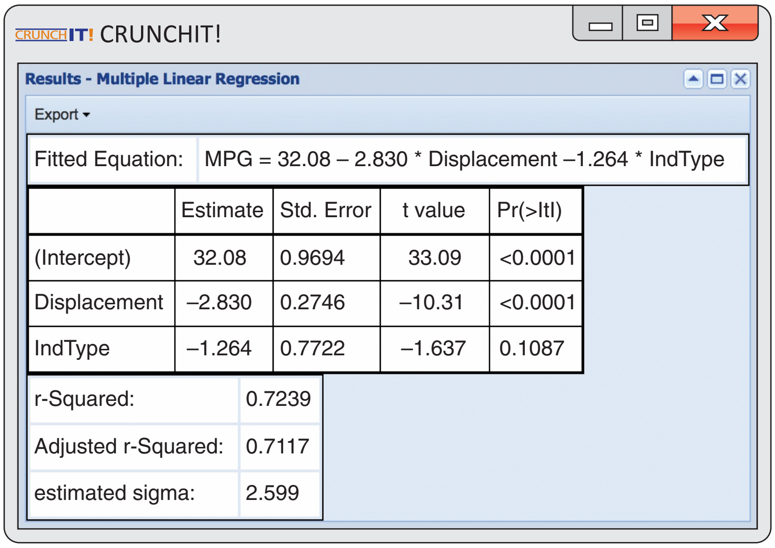
Regression Analysis: MPG versus Displacement, IndType

```
The regression equation is

MPG = 32.1 − 2.83 Displacement − 1.26 IndType

Predictor            Coef   SE Coef        T       P
Constant          32.0770    0.9694    33.09  0.0000
Displacement      −2.8305    0.2746   −10.31  0.0000
IndType           −1.2638    0.7722    −1.64   0.109

S = 2.59915    R-Sq = 72.4%    R-Sq(adj) = 71.2%

Analysis of Variance

Source          DF        SS      MS        F        P
Regression       2    797.25  398.62    59.01    0.000
Residual Error  45    304.00    6.76
Total           47   1101.25
```

FIGURE 29.3 Output from Minitab, CrunchIt!, and JMP for the model with parallel regression lines in Example 29.3.

FIGURE 29.3
(Continued)

CrunchIt!

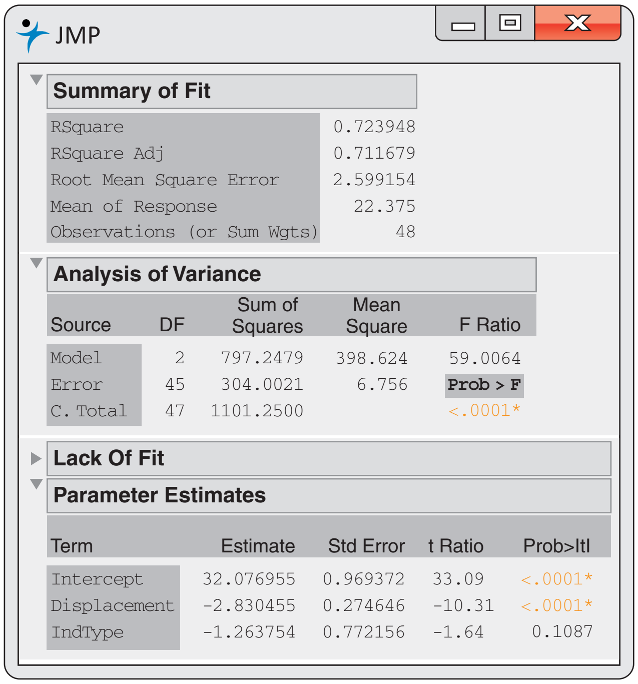


Fitted Equation:	MPG = 32.08 – 2.830 * Displacement –1.264 * IndType

	Estimate	Std. Error	t value	Pr(>ItI)
(Intercept)	32.08	0.9694	33.09	<0.0001
Displacement	–2.830	0.2746	–10.31	<0.0001
IndType	–1.264	0.7722	–1.637	0.1087

r-Squared:	0.7239
Adjusted r-Squared:	0.7117
estimated sigma:	2.599

JMP



```
Summary of Fit
RSquare                        0.723948
RSquare Adj                    0.711679
Root Mean Square Error         2.599154
Mean of Response                 22.375
Observations (or Sum Wgts)           48

Analysis of Variance
                   Sum of       Mean
Source     DF     Squares     Square     F Ratio
Model       2    797.2479    398.624     59.0064
Error      45    304.0021      6.756     Prob > F
C. Total   47   1101.2500                <.0001*

Lack Of Fit
Parameter Estimates
Term            Estimate    Std Error   t Ratio   Prob>ItI
Intercept       32.076955   0.969372    33.09     <.0001*
Displacement    -2.830455   0.274646    -10.31    <.0001*
IndType         -1.263754   0.772156    -1.64      0.1087
```

EXAMPLE 29.5 Parameter Estimates on Statistical Output

All three outputs give the estimated multiple regression model for predicting *MPG* (after rounding) as $\hat{y} = 32.08 - 2.830x_1 - 1.264x_2$.

Although the labels differ, the regression standard error is provided by all three packages:

Minitab:	S = 2.59915
CrunchIt!:	Sigma = 2.599
JMP	Root mean square error = 2.599154

For simple linear regression models, the square of the correlation coefficient (r^2) between y and x measures the proportion of variation in the response variable that is explained by using the explanatory variable. For our multiple regression model with parallel regression lines, we do not have one correlation coefficient. However, by squaring the correlation coefficient between the observed responses y and the predicted responses $\hat{y}$ we obtain the *squared multiple correlation coefficient* R^2.

The analysis of variance table helps us interpret this new statistic. The sum of squares row in the ANOVA table breaks the total variability in the responses into two pieces. One piece summarizes the variability explained by the model, and the other piece summarizes the "leftover" variability, traditionally called "error." That is,

$$\text{Total sum of squares} = \text{Model sum of squares} + \text{Error sum of squares}$$

The value of R^2 is the ratio of the model sum of squares to the total sum of squares, so R^2 tells us what proportion of the variation in the response variable y we explained by using the set of explanatory variables in the multiple regression model.

Squared Multiple Correlation Coefficient

The squared multiple correlation coefficient R^2 is the square of the correlation between the observed responses y and the predicted responses $\hat{y}$. It is also equal to

$$R^2 = \frac{\text{Variability explained by model}}{\text{Total variability in } y} = \frac{\text{Model sum of squares}}{\text{Total sum of squares}}$$

R^2 is almost always given with a regression model to describe the fit of the model to the data.

EXAMPLE 29.6 Using R^2

All three outputs in Figure 29.3 give the value $R^2 = .7239$ (rounded up to 72.4% in Minitab) for our multiple regression model with parallel lines in Example 29.3. That is, the regression model with explanatory variables *Displacement* and *IndType* explains about 72% of the variation in the response variable *MPG*.

APPLY YOUR KNOWLEDGE

29.6 Heights and Weights for Boys and Girls. Suppose you are designing a study to investigate the relationship between height and weight for boys and girls.

(a) Specify a model with parallel regression lines that could be used to predict height separately for boys and for girls. Be sure to identify all variables and describe all parameters in your model.

(b) How many columns in a worksheet would be required to fit this model with statistical software? Describe each column.

29.7 Nestling Mass and Nest Humidity. Researchers investigated the relationship between nestling mass, measured in grams, and nest humidity index, measured as the ratio of total mass of water in the nest divided by nest dry mass, for two different groups of great titmice parents.[2] One group was exposed to fleas during egg laying and the other was not. Exposed parents were coded as 1, and unexposed parents were coded as 0. Use the output below, obtained by fitting a multiple regression model with parallel lines for the two groups of parents, to answer the following questions. DATA NESTL

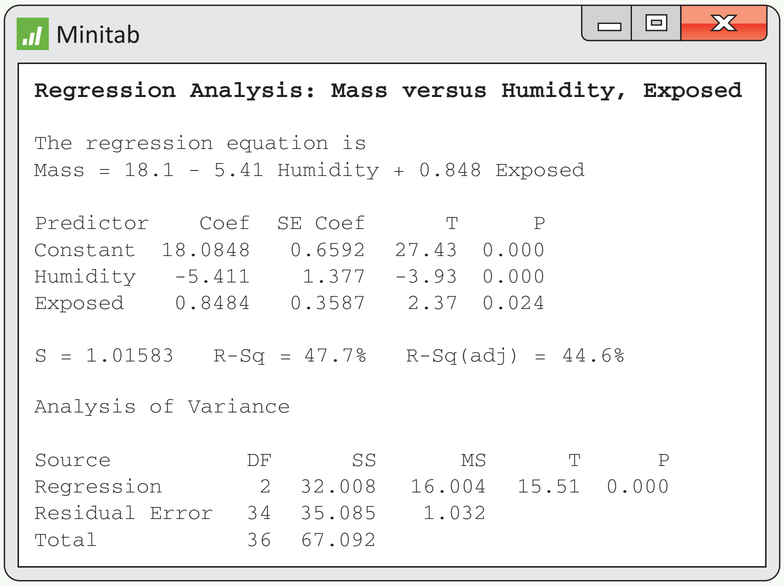
(a) Identify the estimated regression model for predicting nestling mass from nest humidity index for the two groups of great titmice parents.

(b) Based on your model, do you think that nestling mass was higher in nests of birds exposed to fleas during egg laying? Explain.

(c) What is the value of the regression standard error? Interpret this value.

(d) What is the value of the squared multiple correlation coefficient? Interpret this value.

Minitab



```
Regression Analysis: Mass versus Humidity, Exposed

The regression equation is
Mass = 18.1 - 5.41 Humidity + 0.848 Exposed

Predictor     Coef  SE Coef      T      P
Constant   18.0848   0.6592  27.43  0.000
Humidity    -5.411    1.377  -3.93  0.000
Exposed     0.8484   0.3587   2.37  0.024

S = 1.01583   R-Sq = 47.7%   R-Sq(adj) = 44.6%

Analysis of Variance

Source          DF      SS      MS      T      P
Regression       2  32.008  16.004  15.51  0.000
Residual Error  34  35.085   1.032
Total           36  67.092
```

29.4 Inference for Multiple Regression

The output in Figure 29.3 (page 29-10) contains a considerable amount of additional information that deals with statistical inference for our multiple regression model with parallel lines. Before taking our first look at inference for multiple regression, we will check the conditions for inference.

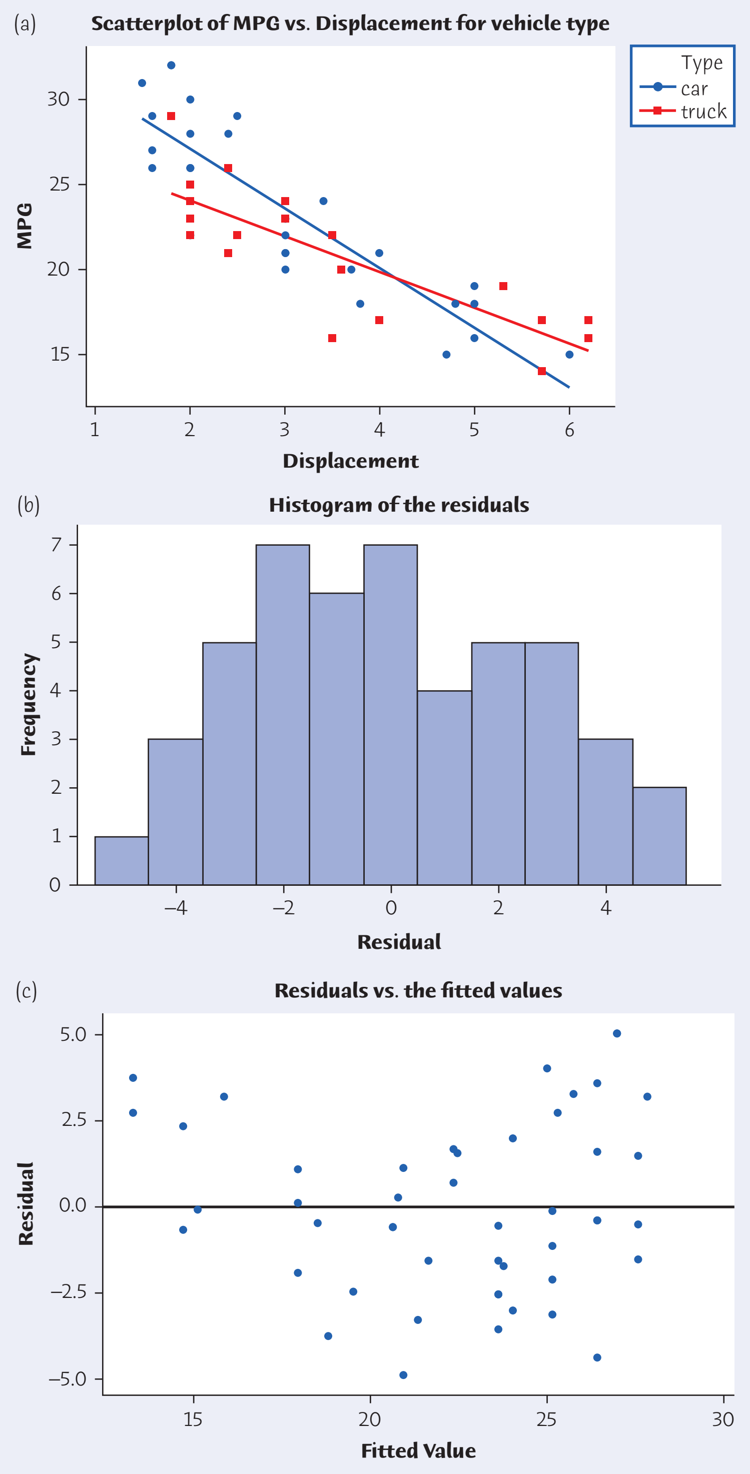
EXAMPLE 29.7 Checking the Conditions

A scatterplot and residual plots for the multiple regression model with parallel lines in Example 29.3 are shown in Figure 29.4. The conditions for inference are linearity, Normality, constant variance, and independence. We will check these conditions one at a time.

LINEAR TREND: The scatterplot in Figure 29.4(a) shows a linear pattern for the two vehicle types, but the pattern suggests that the lines are not parallel. So the model, which assumes parallel lines, may not be reasonable. The residual plot in Figure 29.4(c) shows a nonlinear pattern, with a high proportion of negative residuals for fitted values between about 18 and 23. Again, this suggests that the model may not be reasonable.

NORMALITY: The histogram of the residuals in Figure 29.4(b) indicates that the residuals are symmetric about zero and approximately Normal.

FIGURE 29.4
Scatterplot, histogram, and residual plot to check the conditions for inference in the model with parallel regression lines in Example 29.3.



CONSTANT VARIANCE: The residual plot in Figure 29.4(c) is not a perfectly unstructured horizontal band of points. However, the overall pattern does suggest that the variability in the residuals is roughly constant, with perhaps slightly less variability for smaller fitted values. In general, however, this residual plot does not provide compelling evidence against the model's condition that a single σ describes the scatter about the car line and the truck line.

INDEPENDENCE: Because 48 vehicles were randomly selected, it is reasonable to assume that the MPGs are independent. Patterns in residual plots, such as in

Figure 29.4(c), could occur if observations with similar responses are correlated. However, they can also occur because we have fit the wrong model for describing the relation between the response and explanatory variables. That appears to be the case here, so we see no compelling evidence that the assumption of independence is violated. Also, we can rely on the fact that multiple regression models are robust to slight departures from the conditions and proceed with inference for this model.

To this point we have concentrated on understanding the model, estimating parameters, and verifying the conditions for inference that are part of a regression model. Inference in multiple regression begins with tests that help us decide if a model adequately fits the data and choose between several possible models.

The first inference for a multiple regression model examines the overall model. The ANOVA table summarizes the breakdown of the variability in the response variable. There is one row for each of the three sources of variation: Model, Error, and Total. Each source of variation has a number of degrees of freedom associated with it. These degrees of freedom are listed in a column. Another column provides a sum of squares for the three components. The sums of squares are divided by the degrees of freedom within each row to form a column for the mean sum of squares. Finally, the mean sum of squares for the model is divided by the mean sum of squares for error to form the *F statistic* for the overall model. This *F statistic* is used to find out if all of the regression coefficients, except the intercept, are equal to zero.

The *F* Statistic for the Regression Model

The analysis of variance *F* statistic for testing the null hypotheses that all of the regression coefficients (β's), except β_0, are equal to zero has the form

$$F = \frac{\text{Variation due to model}}{\text{Variation due to error}} = \frac{\text{Model mean square}}{\text{Error mean square}}$$

F will be large if most of the variation in the response variable (as measured by the total variation) can be explained by the variation predicted by the regression model (as measured by the model variation).

EXAMPLE 29.8 The F Statistic

The regression model for the mean *MPG* is $\mu_y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$, where x_1 is labeled as *Displacement* and x_2 is labeled as *IndType* on the output in Figure 29.3 (page 29-10). The null and alternative hypotheses for the overall *F* test are

$$H_0\text{: } \beta_1 = \beta_2 = 0 \quad (\text{that is } \mu_y = \beta_0)$$
$$H_a\text{: at least one of } \beta_1 \text{ and } \beta_2 \text{ is not } 0$$

null model

The null hypothesis H_0 specifies a model, called the **null model**, where the response variable y is a constant (its mean) plus random variation. In other words, the null model says that x_1 and x_2 together do not help predict y.

The value of the *F* statistic reported in the ANOVA table in Figure 29.3 is $F =$ 59.01. You should check that this value is the mean square for the model divided by the mean square for error. The *P*-value is obtained from an *F* distribution with 2 numerator and 45 denominator degrees of freedom. Minitab reports a *P*-value of 0.000, that is, zero to 3 decimal places. Because the *P*-value is less than any reasonable significance level, say $\alpha = 0.01$, we reject the null hypothesis and conclude that at least one of the *x*'s helps explain the variation in the reporting percent *y*.

Rejecting the null hypothesis with the *F* statistic tells us that at least one of our β parameters is not equal to zero, but it doesn't tell us which parameters are not equal to zero. We turn to individual tests for each parameter to answer that question.

Individual *t* Tests for Coefficients

To test the null hypothesis that one of the β's in a specific regression model is zero, compute the *t* statistic

$$t = \frac{\text{Parameter estimate}}{\text{Standard error of estimate}} = \frac{b}{SE_b}$$

If the conditions for inference are met, the *t* distribution with $(n - 3)$ degrees of freedom can be used to compute confidence intervals and conduct hypothesis tests for β_0, β_1, and β_2.

EXAMPLE 29.9 Individual *t* Tests

The output in Figure 29.3 (page 29-10) provides parameter estimates and standard errors for the coefficients β_0, β_1, and β_2. The individual *t* statistic for x_1 (*Group*) tests the hypotheses

$$H_0: \beta_1 = 0 \quad (\text{that is } \mu_y = \beta_0 + \beta_2 x_2)$$
$$H_a: \beta_1 \neq 0$$

We explicitly state the model in the null hypothesis because the bare statement H_0: $\beta_1 = 0$ can be misleading. The hypothesis of interest is that *in this model* the coefficient of x_1 is 0. If the same x_1 is used in a different model with different explanatory variables, the hypothesis H_0: $\beta_1 = 0$ has a different meaning even though we would write it the same way.

Using the CrunchIt! output, we see that the test statistic is (with round-off)

$$t = \frac{-2.830}{0.2746} = -10.31$$

The *P*-value is the area under a *t* distribution curve with $48 - 3 = 45$ degrees of freedom below -10.31 or above 10.31. Because this value is very small, CrunchIt! simply reports that the *P*-value is <0.0001. Look back at the hypotheses to interpret this result: we have good evidence that x_1 (*Displacement*) helps explain *MPG y even after we allow vehicle type (IndType) x_2 to explain MPG.*

The test statistics for the other two coefficients are

$$t = \frac{32.08}{0.9694} = 33.09 \text{ for } \beta_0$$

$$t = \frac{-1.264}{0.7722} = -1.637 \text{ for } \beta_2$$

The *P*-values are again obtained using the *t* distribution with 45 degrees of freedom. The *P*-value for β_0 is so small that it is reported by CrunchIt! as being <0.0001. There is good evidence that the constant term β_0 is not 0. The *P*-value for β_2 is 0.1087, which would not be considered statistically significant. Thus, *IndType* x_2 does not add to our ability to explain *MPG* after we take *Displacement* x_1 into account.

Example 29.9 is not completely straightforward. The overall *F* test tells us that a regression model that includes both explanatory variables together helps explain the variation in the response, *MPG*. The individual *t* tests indicate that one explanatory variable, *Displacement*, significantly improves the explanation for the variation in the response, *MPG*, after adjusting for the effect of the other explanatory variable, *IndType*. However, the explanatory variable *IndType* does not significantly improve the explanation for the variation in *MPG* once we have adjusted for the effect of the explanatory variable *Displacement*. An important subtlety is what *IndType* tells us in our model. *IndType* tells us whether a model that assumes *parallel* regression lines explains the response *MPG*. The *t* test therefore tells us that a model that assumes two parallel lines, one for cars and the other for trucks, does not significantly improve the explanation for the variation in *MPG* compared to a model that assumes a common line using the explanatory variable *Displacement*.

Interpreting the results of individual *t* tests can get very tricky, and we will return to other challenging situations later. We end our discussion of the model with parallel regression lines with an example that applies the four-step process.

EXAMPLE 29.10 Metabolic Rate and Body Mass in Caterpillars

BMASSLG

STATE: Scientists have long been interested in the question of how body mass (BM) determines physiological characteristics such as metabolic rate (MR). Recent experimental and theoretical research has confirmed the general relationship

$$MR = \alpha(BM)^{\beta}$$

between basal metabolic rate and body mass.[3] However, there is still considerable debate on whether the scaling exponent is $\beta = 2/3$ or $3/4$.

A group of researchers investigated the relationship between metabolic rate and body mass for tobacco hornworm caterpillars (*Manduca sexta*). These caterpillars were chosen because they maintain their shape throughout the five stages of larval development, and the size of the tracheal system increases at each molt. A subset of the metabolic rates and body masses, after applying the logarithm transformation, is shown in Table 29.3 for caterpillars at the fourth and fifth stages of development.[4] Does the general relationship between metabolic rate and body mass hold for tobacco hornworm caterpillars? Is the relationship the same for the two different stages?

TABLE 29.3 **Body masses and metabolic rates, after applying the logarithm transformation, for caterpillars in the fourth and fifth stages of development**

Log of Body Mass	Log of Metabolic rate	Stage	Stage Indicator
−0.56864	0.90780	4	0
−0.21753	1.24695	4	0
0.05881	1.51624	4	0
0.03342	1.42951	4	0
0.29336	1.56236	5	1
0.65562	1.92571	5	1
0.84757	1.83893	5	1
0.97658	2.03313	5	1

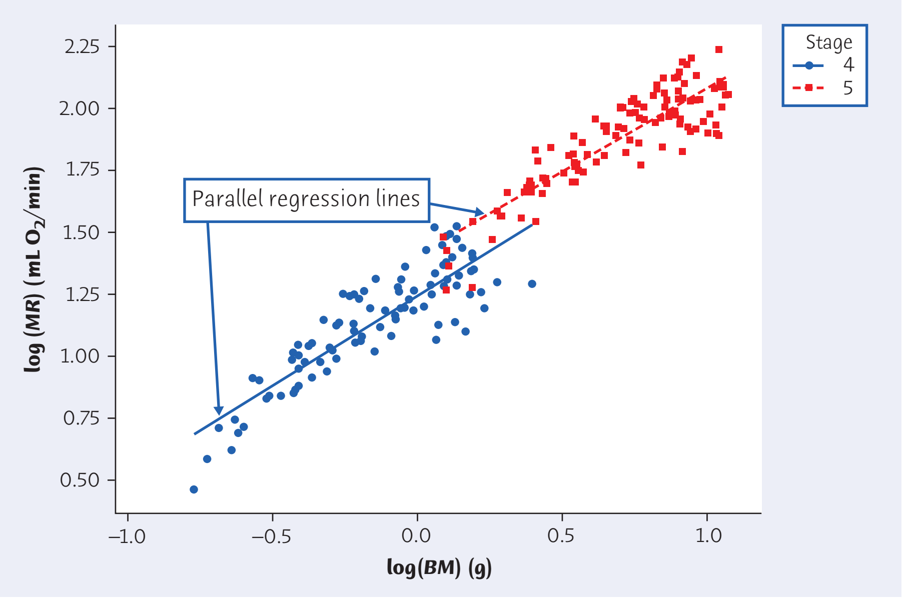


FIGURE 29.5 Scatterplot for the predicted model using parallel regression lines, for Example 29.10.

PLAN: To investigate the relationship between MR and BM, transform the data using logarithms so that the linear model

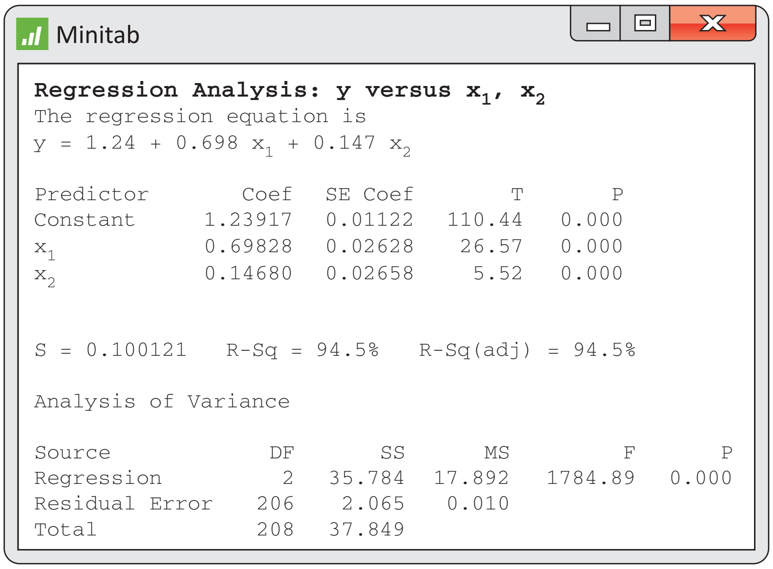
$$\mu_{log(MR)} = log(\alpha) + \beta\, log(BM)$$

can be fitted. Because a simple linear regression model can be used to address the first research question, we will leave the details for a review exercise (see Exercise 29.8). To check if the linear relationship is the same for both stages, we will fit a model with parallel regression lines.

SOLVE: Figure 29.5 shows a scatterplot of the transformed metabolic rate, measured in microliters of oxygen per minute (μ/min), against the transformed body mass measured in grams (g). The parallel regression lines on the plot, one for Stage 4 and one for Stage 5, illustrate the predicted model. The overall patterns for each of the two stages appear to be very similar. However, the measurements for Stage 5 (red points on the plot) are shifted up and to the right of those for Stage 4 (blue points on the plot).

The Minitab output was obtained by regressing the response variable (the logarithm of metabolic rate) on two predictor variables, x_1 (the logarithm of body mass)

Minitab



```
Regression Analysis: y versus x1, x2
The regression equation is
y = 1.24 + 0.698 x1 + 0.147 x2

Predictor        Coef   SE Coef       T      P
Constant      1.23917   0.01122  110.44  0.000
x1            0.69828   0.02628   26.57  0.000
x2            0.14680   0.02658    5.52  0.000


S = 0.100121   R-Sq = 94.5%   R-Sq(adj) = 94.5%

Analysis of Variance

Source            DF      SS      MS        F      P
Regression         2  35.784  17.892  1784.89  0.000
Residual Error   206   2.065   0.010
Total            208  37.849
```

and an indicator variable x_2, which is 1 for Stage 5 and 0 for Stage 4. Our multiple regression model is $\mu_y = \beta_0 + \beta_1 x_1 + \beta_1 x_2$.

The estimated multiple regression model is

$$\hat{y} = 1.24 + 0.698x_1 + 0.147x_2$$

Substituting the values of 0 and 1 for x_2, we obtain the parallel regression lines

$$\hat{y} = 1.24 + 0.698x_1, \text{ for Stage 4 } (x_2 = 0)$$
$$\hat{y} = 1.387 + 0.698x_1, \text{ for Stage 5 } (x_2 = 1)$$

To check the conditions for inference we notice that the scatterplot in Figure 29.5 seems to show a parallel linear pattern, so the model makes sense. The residual plots in Figure 29.6 are used to check the other conditions. The histogram in Figure 29.6(b)

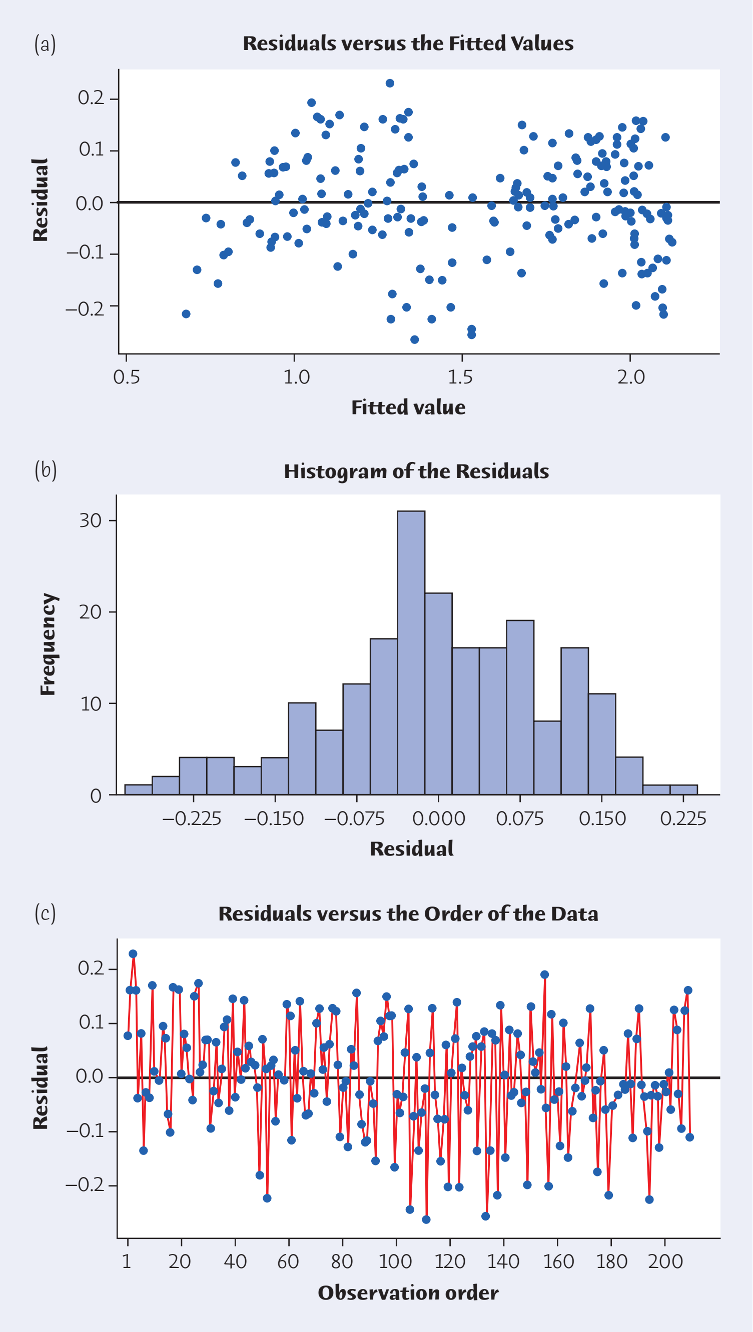


FIGURE 29.6 Residual plots for the model with parallel regression lines in Example 29.10.

indicates that the residuals are approximately symmetric about zero, so the Normality condition is satisfied. The plot of the residuals versus the fitted values in Figure 29.6(a) shows some trends that concerned the researchers. In particular, it appears that a model with some curvature might do a slightly better job because the residuals were always negative for the lowest body mass measurements within each stage. They were also slightly concerned about the constant-variance assumption. The plot of the residuals versus the data order in Figure 29.6(c) shows no systematic change of spread about the model. However, there is a slight curvilinear or "U-shaped" pattern, perhaps suggesting some structure in the process with respect to order.

Because the researchers were interested in comparing their results for caterpillars with the general relationship used by other scientists for a variety of other animals and insects, they decided to proceed with statistical inference for the model parameters. The overall F statistic $F = 1784.89$ and corresponding P-value $P = 0.000$ clearly indicate that at least one of the parameters in the model is not equal to zero. Because the t statistics 110.44, 26.57, and 5.52 all have reported P-values of zero, we conclude that all three parameters β_0, β_1, and β_2 are significantly different from zero.

CONCLUDE: The researchers were pleased that they were able to explain 94.5% of the variation in the logarithm of the metabolic rates by using a regression model with two parallel lines, one for each stage. The general form of the linear relationship is the same for both stages, with overall slope $b_1 = 0.698$. The major difference in the relationship for the two stages is indicated by an upward shift in the line for the larger caterpillars, which is estimated by $b_2 = 0.147$.

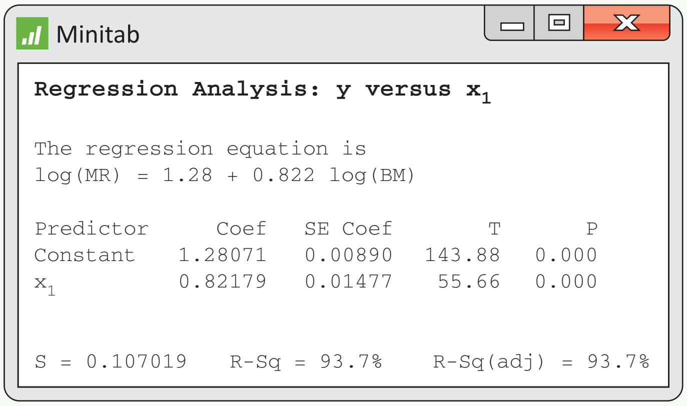
APPLY YOUR KNOWLEDGE

Grady Harrison/Alamy

29.8 Metabolic Rate and Body Mass for Caterpillars. Does the general relationship between metabolic rate and body mass described in Example 29.10 hold for tobacco hornworm caterpillars? The Minitab output (see below) was obtained by regressing the response variable $y = \log(\text{MR})$ on $x_1 = \log(\text{BM})$ for the data.

(a) Use the regression equation from the Minitab output to estimate and in the general relationship $\text{MR} = \alpha(\text{BM})\beta$, which is the same as $y = \log \alpha x_1$. The predicted model is $\hat{y} = a + bx_1$, so that a estimates $\log(\alpha)$ and b estimates β in the original model.

Minitab

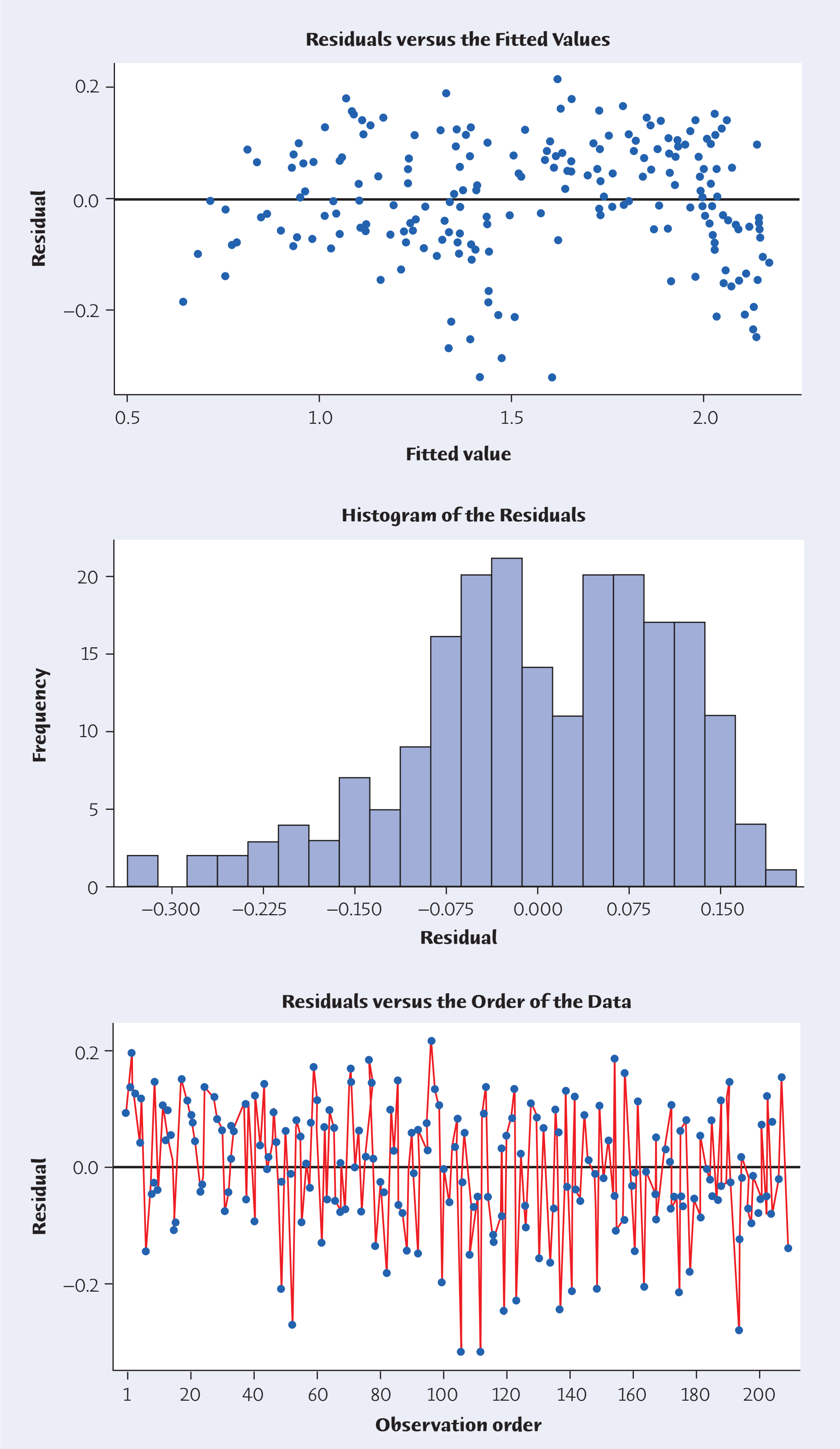


```
Regression Analysis: y versus x1

The regression equation is
log(MR) = 1.28 + 0.822 log(BM)

Predictor      Coef    SE Coef        T       P
Constant    1.28071    0.00890   143.88   0.000
x1          0.82179    0.01477    55.66   0.000


S = 0.107019    R-Sq = 93.7%    R-Sq(adj) = 93.7%
```

(b) Residual plots for the linear regression model $\mu_y = \alpha + \beta x_1$ are shown in the accompanying charts. Do you think that the conditions for inference are satisfied?

(c) Identify the percent of variation in y that is explained by using linear regression with the explanatory variable x_1.

(d) Even if you noticed some departures from the conditions for inference, the researchers were interested in making inferences because this model is well known in the field and has been used for a variety of different insects and animals. Find a 95% confidence interval for the slope parameter β.



Residuals versus the Fitted Values

Histogram of the Residuals

Residuals versus the Order of the Data

(e) Are the values $\beta = 2/3$ and $\beta = 3/4$ contained in your confidence interval?

(f) Use appropriate values from the Minitab output to test the claim that $\beta = 2/3$.

(g) Use appropriate values from the Minitab output to test the claim that $\beta = 3/4$.

29.9 Metabolic Rate and Body Mass for Caterpillars. Use the output provided in Example 29.10 (page 29-17) to answer the questions below.

(a) Find a 95% confidence interval for the slope parameter β for caterpillars during Stage 4.

(b) If you were asked to report a confidence interval for the slope parameter β for caterpillars during Stage 5, would you report the same interval that you calculated in part (a)? Explain why or why not.

(c) Are the values $\beta = 2/3$ and $3/4$ contained in your confidence interval from part (a)?

(d) How does your confidence interval in part (a) compare with the confidence interval you computed in part (d) of Exercise 29.8?

(e) Use appropriate values from the output to test the claim that $\beta = 2/3$.

(f) Use appropriate values from the output to test the claim that $\beta = 3/4$.

29.10 MPG. Use the output in Figure 29.3 (page 29-10) to answer the following questions.

(a) Is the value of the regression standard error the same on all three sets of output? Interpret this value.

(b) The value of the squared multiple correlation coefficient is reported as 72.4% by Minitab, 0.7239 by CrunchIt!, and 0.723948 by JMP. Interpret the value of R^2 for this model.

(c) Is the value of the estimate of β_1, the coefficient for the explanatory variable *Displacement*, the same for all three sets of output?

(d) Give a 98% confidence interval for the value of the parameter β_1. (*Hint:* Remember the general form for t confidence intervals.)

(e) Is there a significant difference in the intercepts for the two regression models (the one for cars only and the one for trucks only)?

29.5 Interaction

interaction

Examples with two parallel linear patterns for two values of an indicator variable are rather rare. It's more common to see two linear patterns that are not parallel, as appeared to be the case for Example 29.7 (page 29-13). To write a regression model for this setting, we need an idea that is new and important: **interaction** between two explanatory variables. Interaction between variables x_1 and x_2 appears as a product term x_1x_2 in the model. The product term means that *the relationship between the mean response and one explanatory variable x_1 changes when we change the value of the other explanatory variable x_2*. Here is an example.

EXAMPLE 29.11 Revisiting State SAT Scores

STATE: In Example 4.4 (text page 106), you discovered that states with a higher percent of high school graduates taking the SAT (rather than the ACT) tend to have lower mean SAT scores. You saw that states fall into two distinct clusters, one for states with 36% or more of high school graduates taking the SAT and the other for states with fewer than 20% of high school graduates taking the SAT. Is a model with two regression lines helpful in predicting the SAT Math score for the two clusters of states?

MATHSAT2

PLAN: Fit and evaluate a model with two regression lines for predicting SAT Math score.

Let's see how adding an interaction term allows two lines that are not parallel. Consider the model

$$\mu_y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$$

in which y is the SAT Math score, x_1 is the percent of high school students taking the SAT, x_2 is an indicator variable that is 1 if the percent of high school graduates taking the SAT is less than 20% and 0 otherwise, and $x_1 x_2$ is the interaction term. For states with 36% or more of students taking the SAT, $x_2 = 0$ and the model becomes

$$\mu_y = \beta_0 + \beta_1 x_1$$

For states with less than 20% of the students taking the SAT, $x_2 = 1$ and the model is

$$\begin{aligned}\mu_y &= \beta_0 + \beta_1 x_1 + \beta_2 + \beta_3 x_1 \\ &= (\beta_0 + \beta_2) + (\beta_1 + \beta_3) x_1\end{aligned}$$

A careful look allows us to interpret all four parameters: β_0 and β_1 are the intercept and slope for states with 36% or more of students taking the SAT. The parameters β_2 and β_3 indicate the fixed change in the intercept and slope, respectively, for states with less than 20% of students taking the SAT. Be careful not to interpret β_2 as the intercept and β_3 as the slope for states with a low percent of students taking the SAT. The indicator variable allows us to change the intercept as we did before, and the new interaction term allows us to change the slope.

29.6 A Model with Two Regression Lines

We have n observations on an explanatory variable x_1, an indicator variable x_2 coded as 0 for some individuals and as 1 for other individuals, and a response variable y. The mean response μ_y is a linear function of the four parameters β_0, β_1, β_2, and β_3:

$$\mu_y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$$

EXAMPLE 29.12 Revisiting State SAT Scores, continued

SOLVE: Figure 29.7 shows the two regression lines, one for each cluster, for predicting the mean SAT Math score for each state. The fitted model, as shown by the two regression lines in this case, appears to provide a good visual summary for the two clusters.

Figure 29.8 provides the regression output from Minitab. By substituting 0 and 1 for the indicator variable x_2, we can easily obtain the two estimated regression lines. The estimated regression lines are $\hat{y} = 561.02 - 0.867x_1$ for states with at least 36% of high school graduates taking the SAT and

$$\begin{aligned}\hat{y} &= (561.02 + 41.69) - (0.867 + 2.751)x_1 \\ &= 602.71 - 3.618x_1\end{aligned}$$

for states with less than 20% of high school graduates taking the SAT.

The overall F statistic 76.15 and corresponding P-value in the ANOVA table clearly indicate that at least one of the regression coefficients is significantly different from

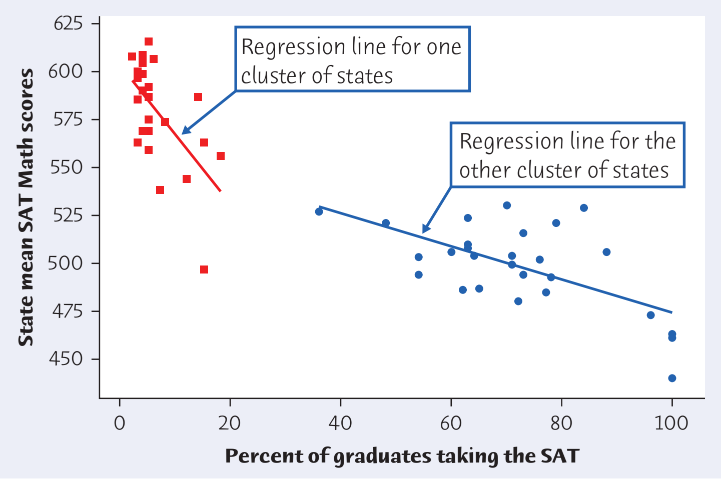
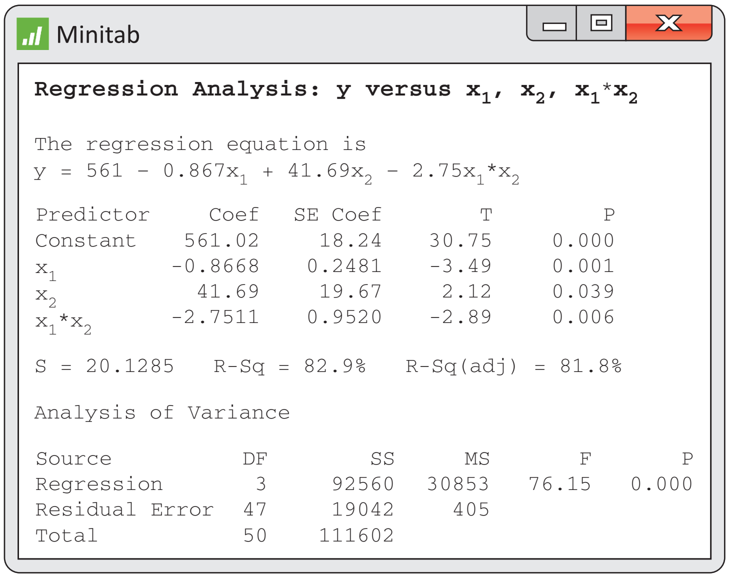


FIGURE 29.7 Model with two regression lines for predicting mean SAT Math score in each state based on the percent of high school graduates who take the SAT, for Example 29.12.

Minitab



```
Regression Analysis: y versus x1, x2, x1*x2

The regression equation is
y = 561 - 0.867x1 + 41.69x2 - 2.75x1*x2

Predictor      Coef   SE Coef       T       P
Constant     561.02     18.24   30.75   0.000
x1          -0.8668    0.2481   -3.49   0.001
x2            41.69     19.67    2.12   0.039
x1*x2       -2.7511    0.9520   -2.89   0.006

S = 20.1285   R-Sq = 82.9%   R-Sq(adj) = 81.8%

Analysis of Variance

Source           DF       SS      MS       F      P
Regression        3    92560   30853   76.15  0.000
Residual Error   47    19042     405
Total            50   111602
```

FIGURE 29.8 Output from Minitab for the model with two regression lines in Example 29.12.

zero. Thus, at least one of the two explanatory variables or the interaction of both is helpful in predicting the state mean SAT Math scores.

Looking at the individual t tests for the coefficients, we notice that all are significantly different from zero at $\alpha = 0.05$, so there is a clear ACT/SAT state difference.

Residual plots (not shown) indicate no major problems with the Normality or constant-variance assumptions.

CONCLUDE: The model with two regression lines, one for each cluster, explains approximately 82.9% of the variation in the mean SAT Math scores. This model provides a better fit than the simple linear regression model that predicts mean SAT Math score from just the percent of high school graduates who take the SAT.

Even though we developed models without interaction first, it is best in practice to consider models with interaction terms before going to the more restrictive model with parallel regression lines. If you begin your model fitting with the more restrictive model with parallel regression lines, then you are basically assuming that there is no interaction. We won't discuss model selection formally, but deciding which model to use is an important skill.

EXAMPLE 29.13 Choosing a Model

Let's compare two separate models for predicting SAT Math score y using the explanatory variables x_1, x_2, and x_1x_2 described in Example 29.11.

- Model 1: A simple linear regression model that ignores the two clusters of states.
- Model 2: The two-line model from Example 29.12.

The predicted response $\hat{y}$, regression standard error s, and squared multiple correlation coefficient R^2 for the three models are

Model 1: $\hat{y} = 586.05 - 1.207x_1$ $s = 21.48$ $R^2 = 0.797$

Model 2: $\hat{y} = 561.02 - 0.867x_1 + 41.69x_2 - 2.751x_1x_2$ $s = 20.13$ $R^2 = 0.829$

We have already seen the fitted lines for Model 2 in Figure 29.7. The fitted line for Model 1 appears in Figure 29.9. The blue line shows the simple linear regression model. Comparing Models 1 and 2, we find that Model 2 has the smaller s and the larger R^2, although the differences are not large. We conclude that the model with two separate regression lines provides a somewhat better fit than the simple linear regression model.

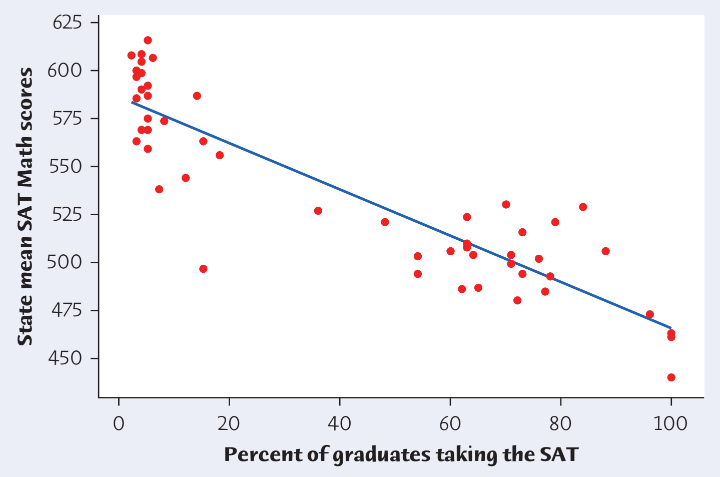


FIGURE 29.9 Scatterplot with two different models for predicting mean SAT Math score in each state based on the percent of high school graduates who take the SAT, for Example 29.13.

APPLY YOUR KNOWLEDGE

29.11 Bird Colonies. Suppose that the number y of new birds that join a colony this year has a straight-line relationship with the percent x_1 of returning birds in colonies of two different bird species. An indicator variable shows which species we observe: $x_2 = 0$ for one and $x_2 = 1$ for the other. Write a population regression model that allows different linear models for the two different bird species. Explain in words what each β in your model means.

29.12 How Fast Do Icicles Grow? We have data on the growth of icicles starting at length 10 centimeters (cm) and at length 20 cm. Suppose icicles that start at 10 cm grow at a rate of 0.15 cm per minute and icicles that start at 20 cm grow at a rate of 0.16 cm per minute. Give a regression model that describes how mean length changes with time x_1 and starting length x_2. Use numbers, not symbols, for the β's in your model.

29.13 Touring Battlefields. Suppose that buses complete tours at an average rate of 20 miles per hour and that self-guided cars complete tours at an average rate of 28 miles per hour. Give a regression model that describes how mean time to complete a tour changes with distance x_1 and mode of transportation x_2. To be realistic, we want the mean time to complete the tour to be zero for both modes of transportation when the distance $x_1 = 0$. Use numbers, not symbols, for the β's in your model.

29.14 Revisiting State SAT Scores. We have examined the relationship between SAT Math scores and the percent of high school graduates who take the SAT. We could also fit a model with two regression lines, one for each cluster, for predicting SAT Writing score. Use software to answer the following questions. WSAT

(a) What is the estimated regression line for predicting mean SAT Writing score for states with more than half of high school graduates taking the SAT?

(b) What is the estimated regression line for predicting mean SAT Writing score for states with at most half of high school graduates taking the SAT?

(c) Interpret the squared multiple correlation.

(d) A t distribution was used to compute the P-values provided after each t-value in the table. How many degrees of freedom does that t distribution have?

(e) Identify the value you would use to estimate the standard deviation σ.

(f) Create a scatterplot containing the estimated regression lines for each cluster.

(g) Plot the residuals against the fitted values. Does this plot indicate any serious problems with the conditions for inference?

(h) Use a visual display to check the Normality condition for the residuals. Do you think the residuals follow a Normal distribution?

29.15 World Record Running Times. The accompanying table shows the progress of world record times (in seconds) for the 10,000-meter run for both men and women. RECORD

Men				Women	
Record Year	Time (seconds)	Record Year	Time (seconds)	Record Year	Time (seconds)
1912	1880.8	1963	1695.6	1967	2286.4
1921	1840.2	1965	1659.3	1970	2130.5
1924	1835.4	1972	1658.4	1975	2100.4
1924	1823.2	1973	1650.8	1975	2041.4
1924	1806.2	1977	1650.5	1977	1995.1
1937	1805.6	1978	1642.4	1979	1972.5
1938	1802.0	1984	1633.8	1981	1950.8
1939	1792.6	1989	1628.2	1981	1937.2
1944	1775.4	1993	1627.9	1982	1895.2
1949	1768.2	1993	1627.9	1983	1895.0
1949	1767.2	1994	1612.2	1983	1887.6
1949	1761.2	1995	1603.5	1984	1873.8
1950	1742.6	1996	1598.1	1985	1859.4
1953	1741.6	1997	1591.3	1986	1813.7
1954	1734.2	1997	1587.8	1993	1771.8
1956	1722.8	1998	1582.7	2016	1757.5
1956	1710.4	2004	1580.4		
1960	1698.8	2005	1577.5		
1962	1698.2				

(a) Make a scatterplot of world record time against year, using separate symbols for men and women. Describe the pattern for each gender. Then compare the progress of men and women.

(b) Fit the model with two regression lines, one for women and one for men, and identify the estimated regression lines.

(c) Women began running this long distance later than men, so we might expect their improvement to be more rapid. Moreover, it is often said that men have little advantage over women in distance running as opposed to sprints, where muscular strength plays a greater role. Do the data appear to support these claims?

29.16 MPG Revisited. In Example 29.9 (page 29-16), we found that type of motor vehicle was not statistically significant for predicting *MPG* in a model that already included *Displacement*. However, we noted that the model in Example 29.9 assumed parallel regression lines for cars and trucks, and Figure 29.4(a) suggested that the regression lines are not parallel. MPG2

(a) Use software to fit a model with two regression lines, one for cars and one for trucks. What is the overall F statistic, standard error s, and R^2? How do s and R^2 compare with the values for the model that assumes parallel regression lines? (See Examples 29.5 and 29.6.)

(b) Use an individual t test to determine if there is a significant interaction effect in a model that includes *Dispersion* and *IndType*. What do you conclude?

29.17 Heights and Weights for Boys and Girls. Suppose that you are designing a study to investigate the relationship between height and weight for boys and girls. Specify a model with two regression lines that could be used to predict height separately for boys and for girls. Be sure to identify all variables and describe all parameters in your model.

29.7 The General Multiple Linear Regression Model

We have seen in a simple but useful case how adding another explanatory variable can fit patterns more complex than the single straight line of simple linear regression. Our examples to this point included two explanatory variables: a quantitative variable x_1 and an indicator variable x_2. Some of our models added an interaction term x_1x_2. Now we want to allow any number of explanatory variables, each of which can be either quantitative or an indicator variable. Here is a statement of the general model that includes the conditions needed for inference.

The Multiple Linear Regression Model

We have observations on n individuals. Each observation consists of values of p explanatory variables $x_1, x_2, \ldots, x_p$ and a response variable y. Our goal is to study or predict the behavior of y given the values of the explanatory variables.

- For any set of fixed values of the explanatory variables, the response y varies according to a **Normal distribution**. Repeated responses y are **independent** of each other.

- The mean response μ_y has a **linear relationship** given by the **population regression model**

$$\mu_y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

  The β's are unknown parameters.
- The **standard deviation** of y (call it σ) is the same for all values of the explanatory variables. The value of σ is unknown.

This model has $p + 2$ parameters that we must estimate from data: the $p + 1$ coefficients $\beta_0, \beta_1, \ldots, \beta_p$ and the standard deviation.

This is *multiple regression* because there is more than one explanatory variable. Some of the x's in the model may be interaction terms, products of two explanatory variables. Others may be squares or higher powers of quantitative explanatory variables. So the model can describe quite general relationships.[5] The main restriction is that the model is *linear regression* because each term is a constant multiple βx. Here are some examples that illustrate the flexibility of multiple regression models.

EXAMPLE 29.14 Two Interacting Explanatory Variables

Suppose we have n observations on two explanatory variables x_1 and x_2 and a response variable y. Our goal is predict the behavior of y for given values of x_1 and x_2. The mean response is given by

$$\mu_y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$$

Because there are two explanatory variables x_1 and x_2, we can graph the relationship of y with x_1 and x_2 in three dimensions. Figure 29.10 shows y vertically above a plane in which x_1 and x_2 take their values. The result is a surface in space. Figure 29.10(a) shows the easiest extension of our simple linear regression model from Chapter 25. Instead of fitting a line to the data, we are now fitting a plane. This figure shows the plane $\mu_y = x_1 + x_2$. The plane is a population model, and when we collect data on our explanatory variables, we will see vertical deviations from the points to the plane. The goal of least-squares regression is to minimize the vertical distances from the points to the plane.

Figure 29.10(b) adds a slight twist. The twist is created by the interaction term in the model. The mean response in Figure 29.10(b) is $\mu_y = 2x_1 + 2x_2 + 10x_1x_2$. The coefficients in front of the explanatory variables indicate part of the effect of a one-unit change on the mean response for each one-unit change in one of the explanatory variables. But the interpretation of the effect of a one-unit change in the mean response for one variable also depends on the other variable. For example, if $x_2 = 1$, the mean response increases by 12 ($\mu_y = 2 + 12x_1$) for a one-unit increase in x_1. However, when $x_2 = 2$, the mean response increases by 22 ($\mu_y = 4 + 22x_1$) for a one-unit increase in x_1. *To interpret the parameters in multiple regression models, we think about the impact of one variable on the mean response while all of the other variables are held fixed.*

Another way to think about possible multiple regression models for two explanatory variables is to take a piece of paper and hold it as shown in Figure 29.10(a). Now begin moving the corners of the paper to get different surfaces. You see that a wide variety of surfaces are possible with only two explanatory variables.

Another possible response surface is shown in Figure 29.10(c). A quick inspection of this figure reveals some curvature in the mean response. To get a curved response

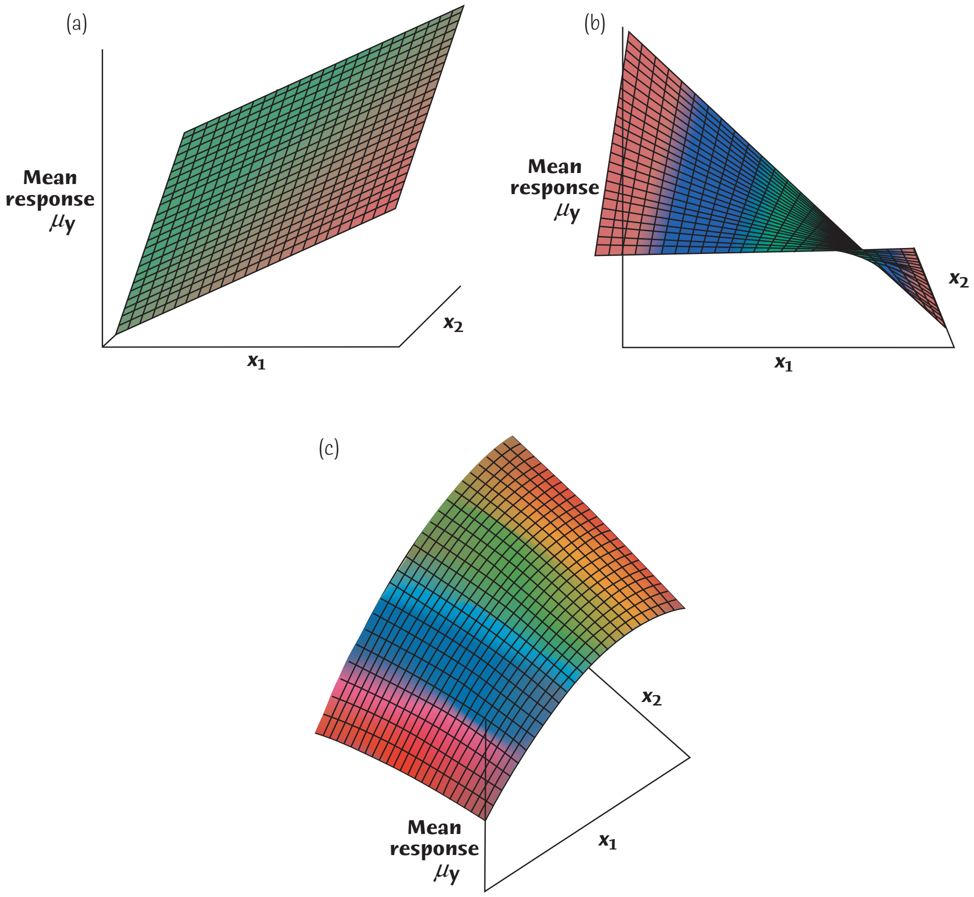


FIGURE 29.10 Some possible surfaces for multiple regression models. (a) Shows the plane $\mu_y = x_1 + x_2$. (b) Shows the surface $\mu_y = 2x_1 + 2x_2 + 10x_1x_2$. (c) Shows the surface $\mu_y = 2000 - 20x_1^2 - 2x_1 - 3x_2^2 + 5x_2 + 10x_1x_2$.

surface, add terms for the squares or higher powers of the explanatory variables. The mean response in Figure 29.10(c) is $\mu_y = 2000 - 20x_1^2 - 2x_1 - 3x_2^2 + 5x_2 + 10x_1x_2$. This model has two linear terms, two quadratic terms, and one interaction term. Models of this form are known as second-order polynomial regression models.

Software fits the model just as before, estimating the β's by the least-squares method and estimating σ by the regression standard error based on the residuals. Nothing essential is new, though you will notice different degrees of freedom depending on the number of terms in the model.

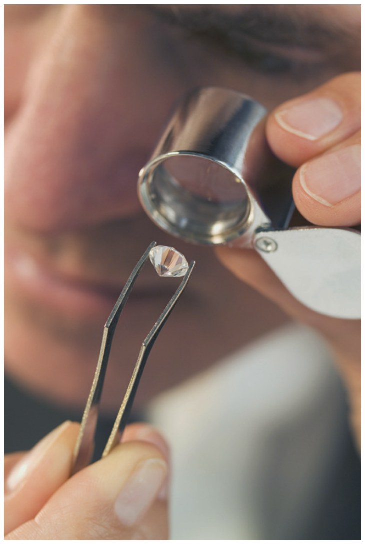
EXAMPLE 29.15 Quadratic Regression

DIAMOND

If there is a quadratic relationship between a quantitative variable y and another quantitative variable x_1, the mean response is given by

$$\mu_y = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2$$

A young couple are shopping for a diamond, so they are interested in learning more about how these gems are priced. They have heard about the 4 Cs: carat, color, cut, and clarity. Is there is a relationship between these diamond characteristics and the price? Table 29.4 shows records for the first 10 diamonds in a large database.[6] The complete database contains 351 diamonds. The variables include *Carat, Color, Clarity,* the *Depth* of the cut, the price per carat *Price/Ct*, and the *Total Price.* Because the young couple are primarily interested in the price of a diamond, they decide to begin by examining the relationship between *Total Price* and *Carat*. Figure 29.11 shows a scatterplot of



© Tom Grill/Corbis

TABLE 29.4 Subset of diamond database

Carat	Color	Clarity	Depth	Price/Carat	Total Price
1.08	E	VS1	68.6	$6693.3	$7228.8
0.31	F	VVS1	61.9	3159.0	979.3
0.31	H	VS1	62.1	1755.0	544.1
0.32	F	VVS1	60.8	3159.0	1010.9
0.33	D	IF	60.8	4758.8	1570.4
0.33	G	VVS1	61.5	2895.8	955.6
0.35	F	VS1	62.5	2457.0	860.0
0.35	F	VS1	62.3	2457.0	860.0
0.37	F	VVS1	61.4	3402.0	1258.7
0.38	D	IF	60.0	5062.5	1923.8

Total Price versus *Carat*, along with the estimated quadratic regression model. Using the quadratic regression model, the couple estimate the mean price of a diamond to be

$$\hat{\mu}_{Price} = -522.7 + 2386Carat + 4498Carat^2$$

The couple are happy because they can explain 92.6% of the variation in the total price of the diamonds in the database using this quadratic regression model. However, they are concerned because they used explanatory variables that are not independent. An explanatory variable and its square are obviously related to one another. The correlation between *Carat* (x_1) and *Carat*^{2}(x_1^2) is 0.952.

The residual plots in Figure 29.12 give more reasons for the couple to be concerned. The histogram in Figure 29.12(b) shows that the residuals are roughly symmetric about zero, but the Normal distribution may not be appropriate because of the unusually large and small residuals. The scatterplot of the residuals against the fitted values in Figure 29.12(a) indicates that the variance increases as the fitted value increases up to approximately $30,000. Finally, the plot of the residuals against order in Figure 29.12(c) does not reveal any troubling pattern, but it does clearly illustrate several unusually large and small residuals.

Having noticed all of the problems with the residual plots, the couple step back and reconsider their objective. They were interested in learning about the relationship between the total price of a diamond and one particular characteristic, carat. The quadratic regression model clearly provides useful information to them even though they will not use this model to make inferences. You will consider additional models to help the couple learn more about diamond pricing in the chapter exercises.

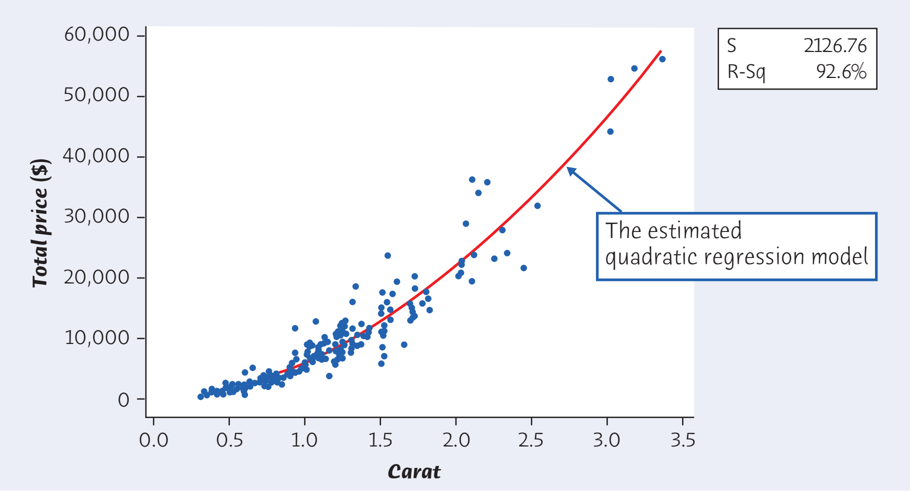


FIGURE 29.11
A scatterplot of *Total Price versus Carat*, for Example 29.15. The estimated quadratic regression model is also shown.

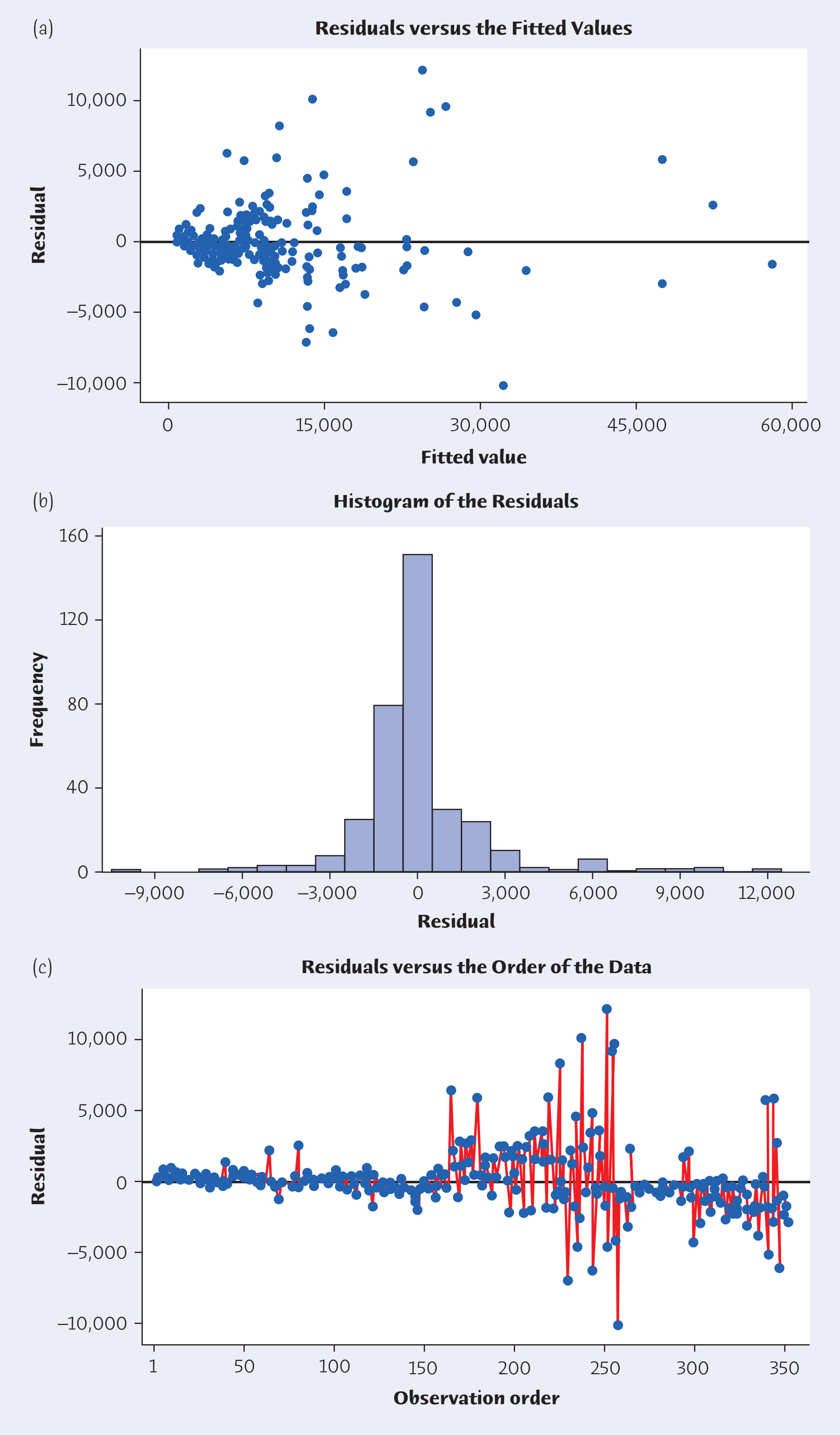


FIGURE 29.12 Residual plots for the quadratic regression model, Example 29.15.

APPLY YOUR KNOWLEDGE

29.18 Nest Humidity and Fleas. In the setting of Exercise 29.7 (page 29-12), researchers showed that the square root of the number of adult fleas y has a quadratic relationship with the nest humidity index x. Specify the population regression model for this situation.

29.19 Diamonds. Specify the population regression model for predicting the total price of a diamond from two interacting variables, *Carat* and *Depth* (see Example 29.15 on page 29-29).

29.20 Radioactive Decay. An experiment was conducted using a Geiger-Mueller tube in a physics lab. Geiger-Mueller tubes respond to gamma rays and to beta particles (electrons). A pulse that corresponds to each detection of a decay product is produced, and these pulses were counted using a computer-based nuclear counting board. Elapsed time (in seconds) and 29–33 counts of pulses for a short-lived unstable isotope of silver are shown in Table 29.5.[7] GAMMA

TABLE 29.5 Counts of pulses over time for an unstable isotope of silver

Seconds	Count	Seconds	Count	Seconds	Count	Seconds	Count
20	4611	330	288	640	86	950	13
30	3727	340	331	650	71	960	24
40	3071	350	298	660	77	970	15
50	2587	360	274	670	64	980	13
60	2141	370	289	680	58	990	21
70	1816	380	253	690	48	1000	23
80	1577	390	235	700	58	1010	16
90	1421	400	220	710	57	1020	17
100	1244	410	216	720	55	1030	19
110	1167	420	219	730	50	1040	14
120	992	430	200	740	54	1050	18
130	927	440	170	750	53	1060	10
140	833	450	185	760	38	1070	13
150	811	460	174	770	35	1080	10
160	767	470	163	780	38	1090	11
170	658	480	178	790	28	1100	21
180	656	490	144	800	34	1110	10
190	651	500	147	810	34	1120	10
200	582	510	154	820	32	1130	12
210	530	520	138	830	30	1140	12
220	516	530	140	840	21	1150	11
230	483	540	121	850	33	1160	8
240	500	550	134	860	19	1170	12
250	508	560	105	870	25	1180	13
260	478	570	108	880	30	1190	11
270	425	580	83	890	22	1200	14
280	441	590	104	900	23	1210	11
290	388	600	95	910	28	1220	10
300	382	610	68	920	28	1230	12
310	365	620	85	930	28	1240	8
320	349	630	83	940	19	1250	11

(a) Create a scatterplot of the counts versus time and describe the pattern.

(b) Because some curvature is apparent in the scatterplot, you might want to consider the quadratic model for predicting counts based on time. Fit the quadratic model and identify the estimated mean response.

(c) Add the estimated mean response to your scatterplot. Would you recommend the use of the quadratic model for predicting radioactive decay in this situation? Explain.

(d) Transform the counts using the natural logarithm and create a scatterplot of the transformed variable versus time.

(e) Fit a simple linear regression model using the natural logarithm of the counts. Provide the estimated regression line, a scatterplot with the estimated regression line, and appropriate residual plots.

(f) Does the simple linear regression model for the transformed counts fit the data better than the quadratic regression model? Explain.

29.8 The Woes of Regression Coefficients

When we start to explore models with several explanatory variables, we quickly meet the big new idea of multiple regression in practice: *the relationship between the response y and any one explanatory variable can change greatly depending on what other explanatory variables are present in the model.* Let's try to understand why this can happen before we illustrate the idea with data.

EXAMPLE 29.16 How Much Change?

Let y denote the total amount of change in a person's pocket or purse. Suppose you are interested in modeling this response variable based on two explanatory variables. The first explanatory variable x_1 is the total number of coins in a person's pocket or purse, and the second explanatory variable x_2 is the total number of pennies, nickels, and dimes. Both of these explanatory variables will be positively correlated with the total amount of change in a person pocket or purse.

Regress y on x_2 alone: we expect the coefficient of x_2 to be positive because the money amount y generally goes up when your pocket has more pennies, nickels, and dimes in it.

Regress y on both x_1 and x_2: for any fixed x_1, larger values of x_2 mean fewer quarters in the overall count of coins x_1, and this means that the money amount y often gets *smaller* as x_2 gets larger. So when we add x_1 to the model, the coefficient of x_2 not only changes but may change sign from positive to negative.

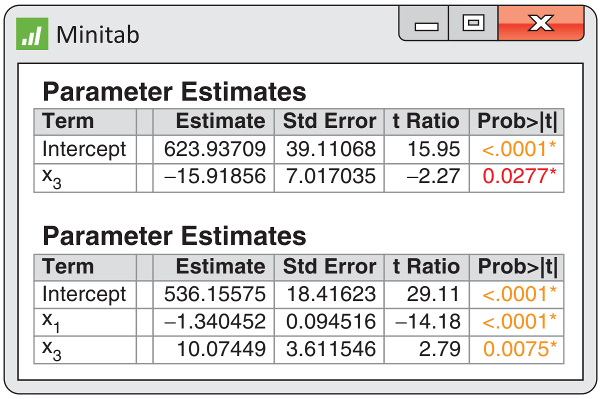
The reason for the behavior in Example 29.16 is that the two explanatory variables x_1 and x_2 are related to each other as well as to the response y. When the explanatory variables are correlated, multiple regression models can produce some very striking and sometimes counterintuitive results, so we must check carefully for correlation among our potential set of explanatory variables.

For an example with data, let's return to the setting described in Example 29.11 (page 29-22), where we are interested in predicting state average SAT Math scores y based on the percent x_1 of graduates in each state who take the SAT.

EXAMPLE 29.17 Predicting SAT Math Scores

Exercise 4.49 (page 126) provides data on average high school teacher salaries and average Mathematics SAT scores for each of the 50 states. The top of Figure 29.13 gives part of the regression output for fitting a simple linear regression for predicting average Mathematics SAT score (y) from average high school teacher salary (x_3, in tens of thousands of dollars). The estimated model is (with rounding) $\hat{y} = 623.94 - 15.92x_3$. The individual t statistic is $t = -2.27$ and corresponding P-value 0.0277 indicate the slope is not equal to 0. The fitted model suggests that for each increase in average teacher salaries of $10,000, predicted average Mathematics SAT scores *decrease* by about 16 points.

When we add the percent taking the exam in each state (x_1), the output at the bottom of Figure 29.13 shows that the individual t statistic for x_3 is $t = 2.79$ and the P-value is 0.0075. The coefficient b_3 for x_3 is statistically significant but is 10.07. The fitted model now suggests that for each increase in average teacher salaries of $10,000, predicted average Mathematics SAT scores *increase* by about 10 points, holding the percentage taking constant. So, depending on the other variables present in the model, predicted average Mathematics SAT scores can either decrease or increase for each $10,000 increase in average teacher salaries!



Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	623.93709	39.11068	15.95	<.0001*
x_3	–15.91856	7.017035	–2.27	0.0277*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	536.15575	18.41623	29.11	<.0001*
x_1	–1.340452	0.094516	–14.18	<.0001*
x_3	10.07449	3.611546	2.79	0.0075*

FIGURE 29.13
Partial regression output for a simple linear regression model and a multiple regression model, for Example 29.17.

APPLY YOUR KNOWLEDGE

29.21 Predicting SAT Writing Scores. We have been developing models for SAT Math scores for two different clusters of states. Use the SAT data to evaluate similar models for SAT Writing scores. WSAT

(a) Find the least-squares line for predicting SAT Writing scores from percent taking the exam.

(b) Plot SAT Writing score versus percent taking the exam, and add the least-squares line to your plot.

(c) Are you happy with the fit of your model? Comment on the value of R^2 and the residual plots.

(d) Fit a model, using indicator variables, with two regression lines. Identify the two lines, parameter estimates, t statistics, and corresponding P-values. Does this model improve the fit?

29.22 Body Fat for Men. You are interested in predicting the amount of body fat on a man y using the explanatory variables waist size x_1 and height x_2.

(a) Do you think body fat y and waist size x_1 are positively correlated? Explain.

(b) For a fixed waist size, height x_2 is negatively correlated with body fat y. Explain why.

(c) The slope of the simple linear regression line for predicting body fat from height for a sample of men is almost 0—say, 0.13. Knowing a man's height does not tell you much about his body fat. Do you think this parameter estimate would become negative if a multiple regression model with height x_2 and waist size x_1 was used to predict body fat? Explain.

29.23 Combining Relationships. Suppose that $x_1 = 2x_2 - 4$ so that x_1 and x_2 are positively correlated. Let $y = 3x_2 + 4$ so that y and x_2 are positively correlated.

(a) Use the relationship between x_1 and x_2 to find the linear relationship between y and x_1. Are y and x_1 positively correlated?

(b) Add the equations $x_1 = 2x_2 - 4$ and $y = 3x_2 + 4$ together and solve for y to obtain an equation relating y to both x_1 and x_2. Are the coefficients of both x's positive? Combining explanatory variables that are correlated can produce surprising results.

29.9 A Case Study for Multiple Regression

We will now look at a set of data with several explanatory variables to illustrate the process of arriving at a suitable multiple regression model. In the next section, we will use the model we have chosen for inference, including predicting the response variable.

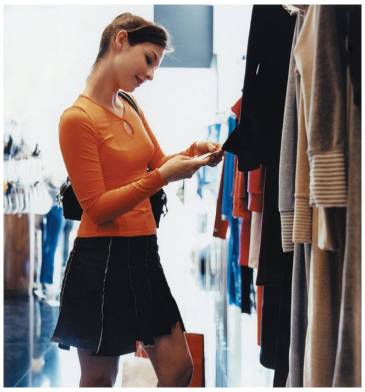
To build a multiple regression model, first examine the data for outliers and other deviations that might unduly influence your conclusions. Next, use descriptive statistics, especially correlations, to get an idea of which explanatory variables may be most helpful in explaining the response. Fit several models using combinations of these variables, paying attention to the individual t statistics to see if any variables contribute little in any particular model. Always think about the real-world setting of your data, and use common sense as part of the process.

EXAMPLE 29.18 Marketing Data for a Clothing Retailer

The data provided in Table 29.6 represent a random sample of 60 customers from a large clothing retailer.[8] The manager of the store is interested in predicting how much a customer will spend on his or her next purchase.

Our goal is to find a regression model for predicting the amount of a purchase from the available explanatory variables. Here is a short description of each variable.

Variable	Description
Amount	The net dollar amount spent by customers who made a purchase from this retailer
Recency	The number of months since the last purchase
Freq12	The number of purchases in the last 12 months
Dollar12	The dollar amount of purchases in the last 12 months
Freq24	The number of purchases in the last 24 months
Dollar24	The dollar amount of purchases in the last 24 months
Card	An indicator variable: *Card* = 1 for customers who have a private label credit card with the retailer, and *Card* = 0 for those who do not



Commercial Eye/Getty Images

TABLE 29.6 Data from clothing retailer

ID	Amount	Recency	*Freq*12	*Dollar*12	*Freq*24	*Dollar*24	Card
1	0	22	0	0	3	400	0
2	0	30	0	0	0	0	0
3	0	24	0	0	1	250	0
4	30	6	3	140	4	225	0
5	33	12	1	50	1	50	0
6	35	48	0	0	0	0	0
7	35	5	5	450	6	415	0
8	39	2	5	245	12	661	1
9	40	24	0	0	1	225	0
10	45	3	6	403	8	1138	0
11	48	6	3	155	4	262	0
12	50	12	1	42	7	290	0
13	50	5	2	100	8	700	1
14	50	8	3	144	4	202	0
15	50	1	10	562	13	595	1
16	50	2	3	166	4	308	0
17	50	4	4	228	4	228	0
18	50	5	5	322	7	717	1
19	55	13	0	0	6	1050	0
20	55	6	3	244	7	811	0
21	57	20	0	0	2	140	0
22	58	3	4	200	4	818	1
23	60	12	1	70	2	150	0
24	60	3	4	256	7	468	0
25	62	12	1	65	5	255	0
26	64	8	1	70	6	300	0
27	65	2	6	471	8	607	0
28	68	6	2	110	3	150	0
29	70	3	3	222	5	305	0
30	70	6	2	120	4	230	0
31	70	5	3	205	8	455	1
32	72	7	4	445	6	400	0
33	75	6	1	77	2	168	0
34	75	4	2	166	5	404	0
35	75	4	3	210	4	270	0
36	78	8	2	180	7	555	1
37	78	5	3	245	9	602	1
38	79	4	3	225	5	350	0
39	80	3	4	300	6	499	0
40	90	73	5	400	9	723	0
41	95	1	6	650	9	1006	1
42	98	6	2	215	3	333	0

(*Continued*)

TABLE 29.6 (*Continued*)

ID	Amount	Recency	*Freq*12	*Dollar*12	*Freq*24	*Dollar*24	Card
43	100	12	1	100	2	200	0
44	100	2	1	110	4	400	1
45	100	3	3	217	6	605	0
46	100	3	4	330	8	660	1
47	105	2	4	400	7	560	0
48	110	3	4	420	6	570	0
49	125	3	2	270	5	590	1
50	140	6	3	405	6	775	0
51	160	2	2	411	8	706	0
52	180	1	5	744	10	945	1
53	200	1	3	558	4	755	1
54	240	4	4	815	10	1150	1
55	250	3	3	782	10	1500	1
56	300	12	1	250	4	401	0
57	340	1	5	1084	7	1162	1
58	500	4	2	777	3	905	1
59	650	1	4	1493	7	2050	1
60	1,506,000	1	6	5000	11	8000	1

The response variable y is the amount of money spent by a customer. A careful examination of Table 29.6 reveals that the first three values for *Amount* are zero because some customers purchased items and then returned them. We are not interested in modeling returns, so these observations will be removed before proceeding. The last row of Table 29.6 indicates that one customer spent $1,506,000 in the store. A quick consultation with the manager reveals that this observation is a data entry error, so this customer will also be removed from our analysis. We can now proceed with the cleaned data on 56 customers.

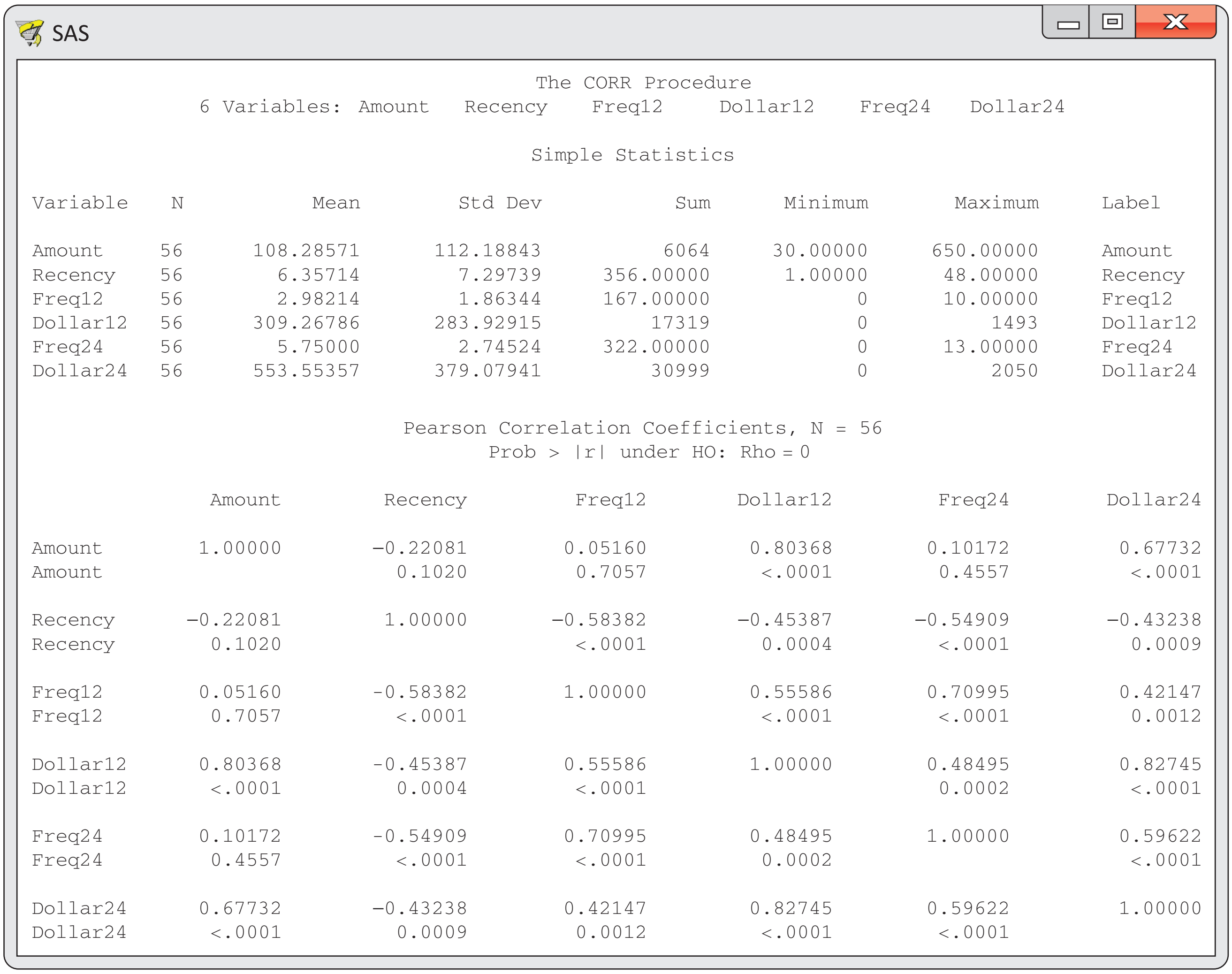
EXAMPLE 29.19 Relationships among the Variables

DATA CLOTHE

We won't go through all of the expected relationships among the variables, but we would certainly expect the amount of a purchase to be positively associated with the amount of money spent over the last 12 and the last 24 months. Speculating about how the frequency of purchases over the last 12 and 24 months is related to the purchase amount is not as easy. Some customers may buy small amounts of clothing on a regular basis, whereas others may purchase large amounts at less frequent intervals. Yet other people may purchase large amounts on a regular basis.

Descriptive statistics and a matrix of correlation coefficients for the six quantitative variables are shown in Figure 29.14. As expected, *Amount* is strongly correlated with past spending: $r = 0.80368$ with *Dollar12* and $r = 0.67732$ with *Dollar24*. However, the matrix also reveals that these explanatory variables are correlated with one another. Because the variables are dollar amounts in

SAS



```
                                  The CORR Procedure
            6 Variables:  Amount    Recency    Freq12     Dollar12    Freq24   Dollar24

                                  Simple Statistics

Variable    N          Mean       Std Dev           Sum      Minimum       Maximum      Label

Amount     56     108.28571     112.18843          6064     30.00000     650.00000      Amount
Recency    56       6.35714       7.29739     356.00000      1.00000      48.00000      Recency
Freq12     56       2.98214       1.86344     167.00000            0      10.00000      Freq12
Dollar12   56     309.26786     283.92915         17319            0          1493      Dollar12
Freq24     56       5.75000       2.74524     322.00000            0      13.00000      Freq24
Dollar24   56     553.55357     379.07941         30999            0          2050      Dollar24

                         Pearson Correlation Coefficients, N = 56
                                Prob > |r| under H0: Rho = 0

                 Amount        Recency        Freq12       Dollar12        Freq24       Dollar24

Amount          1.00000       -0.22081       0.05160       0.80368       0.10172        0.67732
Amount                          0.1020        0.7057        <.0001        0.4557         <.0001

Recency        -0.22081        1.00000      -0.58382      -0.45387      -0.54909       -0.43238
Recency          0.1020                       <.0001        0.0004        <.0001         0.0009

Freq12          0.05160       -0.58382       1.00000       0.55586       0.70995        0.42147
Freq12           0.7057         <.0001                      <.0001        <.0001         0.0012

Dollar12        0.80368       -0.45387       0.55586       1.00000       0.48495        0.82745
Dollar12         <.0001         0.0004        <.0001                      0.0002         <.0001

Freq24          0.10172       -0.54909       0.70995       0.48495       1.00000        0.59622
Freq24           0.4557         <.0001        <.0001        0.0002                       <.0001

Dollar24        0.67732       -0.43238       0.42147       0.82745       0.59622        1.00000
Dollar24         <.0001         0.0009        0.0012        <.0001        <.0001
```

FIGURE 29.14

Descriptive statistics and correlation coefficients, for Example 29.19.

overlapping time periods, there is a strong positive association, $r = 0.82745$, between *Dollar12* and *Dollar24*.

Recency (the number of months since the last purchase) is negatively associated with the purchase amount and with the four explanatory variables that indicate the number of purchases or the amount of those purchases. Perhaps recent customers (low Recency) tend to be regular customers and those who have not visited in some time (high *Recency*) include customers who often shop elsewhere. Customers with low *Recency* would then visit more frequently and spend more.

One common mistake in modeling is to include too many variables in the multiple regression model, especially variables that are related to one another. A hasty user of statistical software will include all explanatory variables along with some possible interaction terms and quadratic terms. Here's an example to show you what can happen.

EXAMPLE 29.20 Including All Explanatory Variables

CLOTHE2

Create the following interaction terms and quadratic terms from the potential explanatory variables:

$$Int12 = Freq12 \times Dollar12$$

$$Int24 = Freq24 \times Dollar24$$

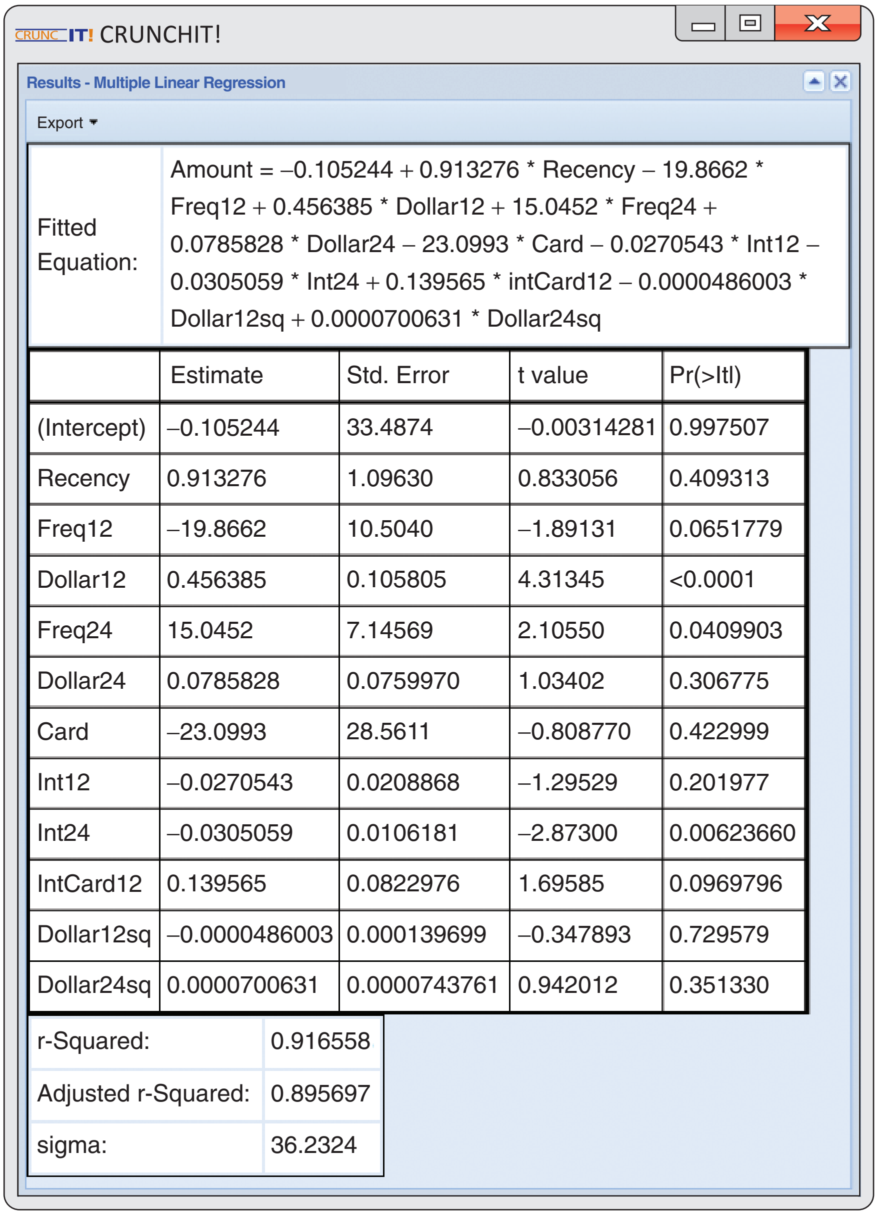
$$IntCard12 = Card \times Dollar12$$

$$Dollar12sq = Dollar12 \times Dollar12$$

$$Dollar24sq = Dollar24 \times Dollar24$$

Figure 29.15 shows the multiple regression output using all six explanatory variables provided by the manager and the five new variables. Most of the individual *t* statistics have *P*-values greater than 0.2, and only three have *P*-values less than 0.05.

CrunchIt!



Fitted Equation: Amount = –0.105244 + 0.913276 * Recency – 19.8662 * Freq12 + 0.456385 * Dollar12 + 15.0452 * Freq24 + 0.0785828 * Dollar24 – 23.0993 * Card – 0.0270543 * Int12 – 0.0305059 * Int24 + 0.139565 * intCard12 – 0.0000486003 * Dollar12sq + 0.0000700631 * Dollar24sq

	Estimate	Std. Error	t value	Pr(>Itl)
(Intercept)	–0.105244	33.4874	–0.00314281	0.997507
Recency	0.913276	1.09630	0.833056	0.409313
Freq12	–19.8662	10.5040	–1.89131	0.0651779
Dollar12	0.456385	0.105805	4.31345	<0.0001
Freq24	15.0452	7.14569	2.10550	0.0409903
Dollar24	0.0785828	0.0759970	1.03402	0.306775
Card	–23.0993	28.5611	–0.808770	0.422999
Int12	–0.0270543	0.0208868	–1.29529	0.201977
Int24	–0.0305059	0.0106181	–2.87300	0.00623660
IntCard12	0.139565	0.0822976	1.69585	0.0969796
Dollar12sq	–0.0000486003	0.000139699	–0.347893	0.729579
Dollar24sq	0.0000700631	0.0000743761	0.942012	0.351330

r-Squared:	0.916558
Adjusted r-Squared:	0.895697
sigma:	36.2324

FIGURE 29.15 CrunchIt! output for the multiple regression model, Example 29.20.

The model is successful at explaining 91.66% of the variation in the purchase amounts, but it is large and unwieldy. Management will have to measure all these variables to use the model in the future for prediction. This model does set a standard: removing explanatory variables can only reduce R_2, so no smaller model that uses some of these variables and no new variables can do better than $R_2 = 91.66\%$. But can a simpler model do almost as well?

Some statistical software provides automated algorithms to choose regression models. All possible regression algorithms are very useful. On the other hand, *algorithms that add or remove variables one at a time* often miss good models. We will not illustrate automated algorithms, but will build models by considering and evaluating various possible subsets of models.

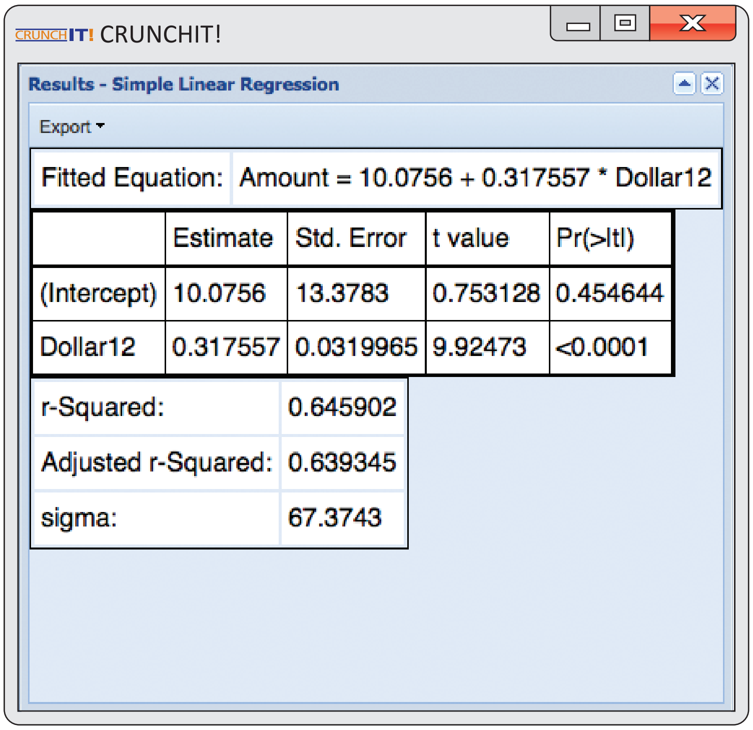
EXAMPLE 29.21 Highest Correlation

To start, let's look at a simple linear regression model with the single explanatory variable most highly correlated with *Amount*. The correlations in Figure 29.14 show that this explanatory variable is Dollar12. The least-squares regression line for predicting the purchase amount *y* is

$$\hat{y} = 10.0756 + 0.31756 Dollar12$$

Figure 29.16 shows the regression output for this simple linear regression model. This simple model has a low R^2 of 64.59%, so we need more explanatory variables.

CrunchIt!



Fitted Equation: Amount = 10.0756 + 0.317557 * Dollar12

	Estimate	Std. Error	t value	Pr(>ltl)
(Intercept)	10.0756	13.3783	0.753128	0.454644
Dollar12	0.317557	0.0319965	9.92473	<0.0001

r-Squared:	0.645902
Adjusted r-Squared:	0.639345
sigma:	67.3743

FIGURE 29.16
CrunchIt! output for the simple linear regression model in Example 29.21 using the dollar amount of purchases in the last 12 months (*Dollar*12) as the explanatory variable.

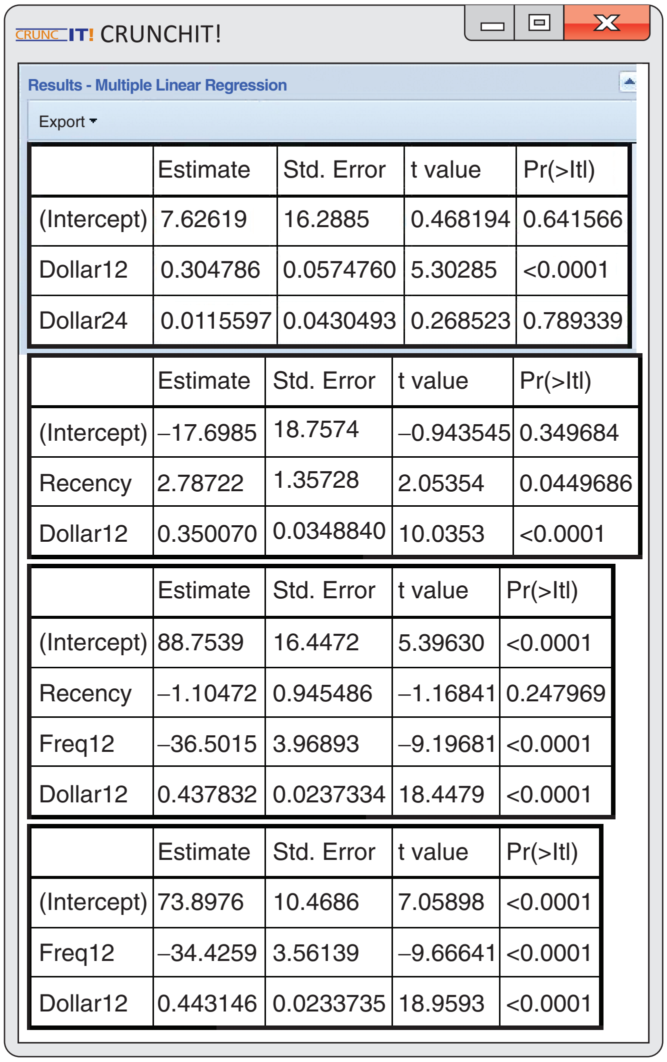
EXAMPLE 29.22 Including Other Explanatory Variables

Of the remaining explanatory variables, *Dollar*24 and *Recency* have the strongest associations with the purchase amounts. We will add these variables to try to improve our model. Rather than providing the complete computer output for each model, we will concentrate on the parameter estimates and individual *t* statistics provided in Figure 29.17. The fitted model using both *Dollar*12 and *Dollar*24 is

$$\hat{y} = 7.63 + 0.30 Dollar12 + 0.01 Dollar24$$

The *t* statistic for *Dollar*12 has dropped from 9.92 to 5.30, but it is still significant. However, if the amount of the purchases over the last 12 months (*Dollar*12) is already in the model, then adding the amount of purchases over the last 24 months (*Dollar*24) does not improve the model.

CrunchIt!



	Estimate	Std. Error	t value	Pr(>ItI)
(Intercept)	7.62619	16.2885	0.468194	0.641566
Dollar12	0.304786	0.0574760	5.30285	<0.0001
Dollar24	0.0115597	0.0430493	0.268523	0.789339

	Estimate	Std. Error	t value	Pr(>ItI)
(Intercept)	–17.6985	18.7574	–0.943545	0.349684
Recency	2.78722	1.35728	2.05354	0.0449686
Dollar12	0.350070	0.0348840	10.0353	<0.0001

	Estimate	Std. Error	t value	Pr(>ItI)
(Intercept)	88.7539	16.4472	5.39630	<0.0001
Recency	–1.10472	0.945486	–1.16841	0.247969
Freq12	–36.5015	3.96893	–9.19681	<0.0001
Dollar12	0.437832	0.0237334	18.4479	<0.0001

	Estimate	Std. Error	t value	Pr(>ItI)
(Intercept)	73.8976	10.4686	7.05898	<0.0001
Freq12	–34.4259	3.56139	–9.66641	<0.0001
Dollar12	0.443146	0.0233735	18.9593	<0.0001

FIGURE 29.17
CrunchIt! parameter estimates and individual *t* statistics for the models in Example 29.22.

Using *Recency* and *Dollar*12, we find the fitted model

$$\hat{y} = -17.7 + 0.35\textit{Dollar}12 + 2.79\textit{Recency}$$

Even though the *t* statistics associated with both explanatory variables are significant, the percent of variation in the purchase amounts explained by this model increases only to 67.2%.

The frequency of visits over the last 12 months (*Freq*12) was not strongly associated with the purchase amount, but may be helpful because dollar amount and frequency provide different information. The fitted model using all three explanatory variables is

$$\hat{y} = 88.75 + 0.44\textit{Dollar}12 - 1.1\textit{Recency} - 36.5\textit{Freq}12$$

The *t* statistic for *Dollar*12 jumps to 18.45, and the *t* statistic for *Recency* drops to −1.17, which is not significant. Eliminating *Recency* from the model, we obtain the fitted model

$$\hat{y} = 73.90 + 0.44\textit{Dollar}12 - 34.43\textit{Freq}12$$

This model explains 87.18% of the variation in the purchase amounts. That is almost as good as the big clumsy model in Example 29.20, but with only two explanatory variables. We might stop here, but we will take one more approach to the problem.

We have used the explanatory variables that were given to us by the manager to fit many different models. However, we have not thought carefully about the data and our objective. Thinking about the setting of the data leads to a new idea.

EXAMPLE 29.23 Creating a New Explanatory Variable

To predict the purchase amount for a customer, the average purchase over a recent time period might be helpful. We have the total amount and frequency of purchases over 12 months, so we can create a new variable

$$\textit{Purchase}12 = \frac{\textit{Dollar}12}{\textit{Freq}12}$$

If no purchases were made in the last 12 months, then *Purchase*12 is set to 0. Fitting a simple linear regression model with this new explanatory variable explains 87.64% of the variation in the purchase amounts. This is better than almost all of our previous models. Figure 29.18 shows the fitted model

$$\hat{y} = -22.99 + 1.34\textit{Purchase}12$$

on a scatterplot of *Amount* versus *Purchase*12 and the corresponding residual plot.

This new linear model provides a good fit. The residual plot in Figure 29.18 shows that low purchase amounts tend to be above the regression line, and moderate purchase amounts tend to be below the line. This suggests that a model with some curvature might improve the fit.

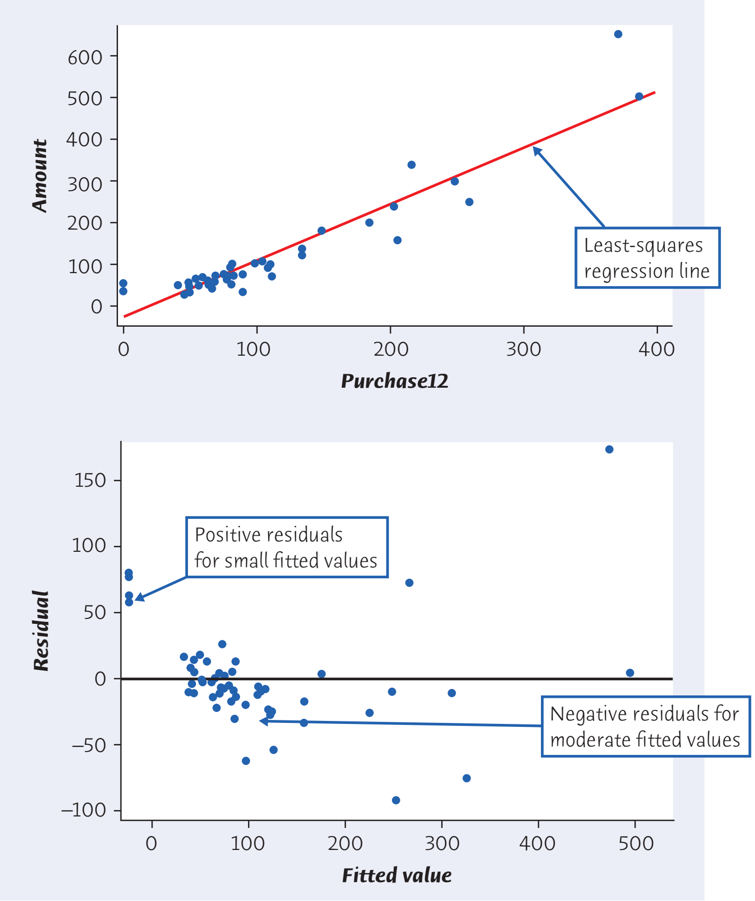


FIGURE 29.18 A scatterplot, including the simple linear regression line, and a residual plot, for Example 29.23.

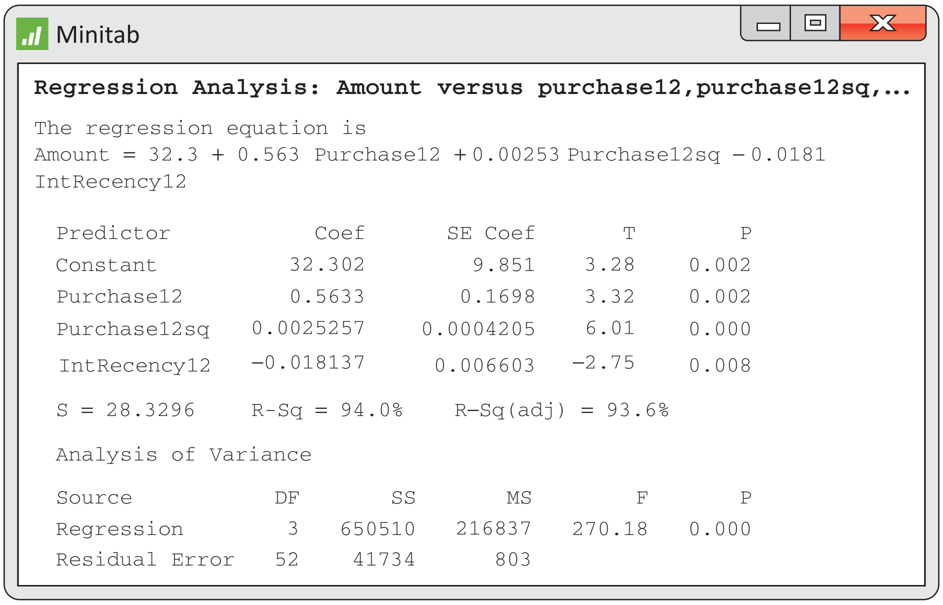
EXAMPLE 29.24 A Final Model

Create the variable *Purchase*12sq, the square of *Purchase*12, to allow some curvature in the model. Previous explorations also revealed that the dollar amount spent depends on how recently the customer visited the store, so an interaction term

$$IntRecency12 = Recency \times Dollar12$$

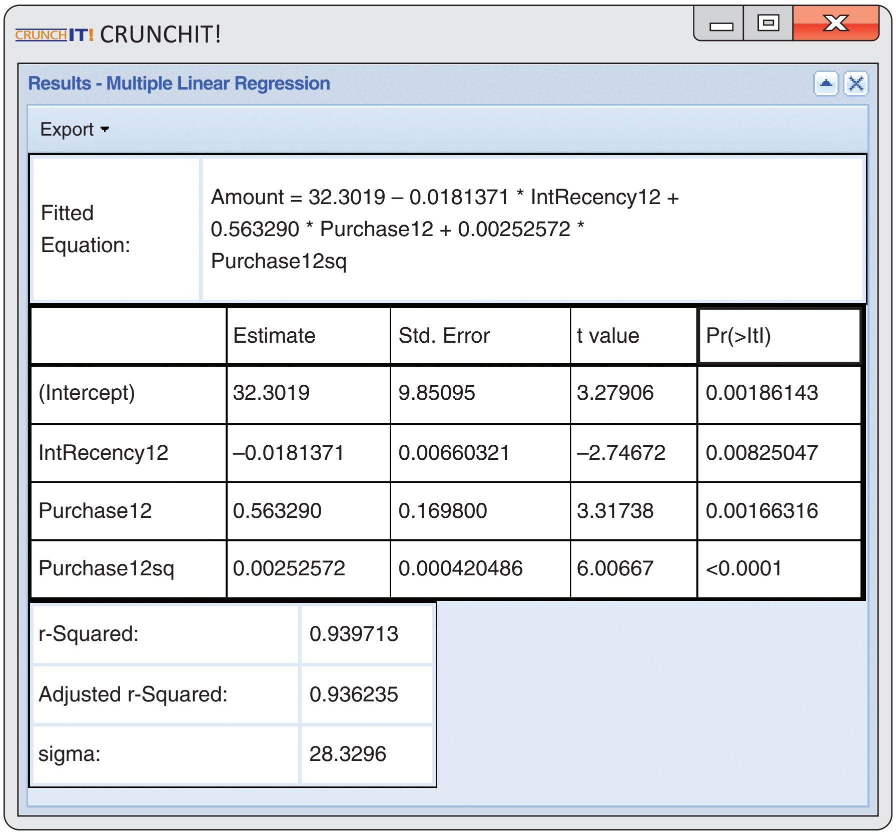
was created to incorporate this relationship into the model. The output for the multiple regression model using the three explanatory variables *Purchase*12, *Purchase*12sq, and *IntRecency*12 is shown in Figure 29.19. This model does a great job for the manager by explaining almost 94% of the variation in the purchase amounts.

Minitab



```
Regression Analysis: Amount versus purchase12,purchase12sq,...

The regression equation is
Amount = 32.3 + 0.563 Purchase12 + 0.00253 Purchase12sq − 0.0181
IntRecency12

  Predictor            Coef       SE Coef       T        P
  Constant           32.302         9.851    3.28    0.002
  Purchase12         0.5633        0.1698    3.32    0.002
  Purchase12sq    0.0025257     0.0004205    6.01    0.000
  IntRecency12    −0.018137      0.006603   −2.75    0.008

  S = 28.3296     R-Sq = 94.0%     R-Sq(adj) = 93.6%

  Analysis of Variance

  Source            DF        SS        MS         F        P
  Regression         3    650510    216837    270.18    0.000
  Residual Error    52     41734       803
```

CrunchIt!



Fitted Equation:	Amount = 32.3019 – 0.0181371 * IntRecency12 + 0.563290 * Purchase12 + 0.00252572 * Purchase12sq

	Estimate	Std. Error	t value	Pr(>ltl)
(Intercept)	32.3019	9.85095	3.27906	0.00186143
IntRecency12	–0.0181371	0.00660321	–2.74672	0.00825047
Purchase12	0.563290	0.169800	3.31738	0.00166316
Purchase12sq	0.00252572	0.000420486	6.00667	<0.0001

r-Squared:	0.939713
Adjusted r-Squared:	0.936235
sigma:	28.3296

FIGURE 29.19
Minitab and CrunchIt! output for the multiple regression model in Example 29.24.

APPLY YOUR KNOWLEDGE

29.24 Diamonds. Suppose that the couple shopping for a diamond in Example 29.15 (page 29-29) had used a quadratic regression model for the other quantitative variable, Depth. Use the data in Table 29.4 to answer the following questions. DIAMND

(a) What is the estimated quadratic regression model for mean total price based on the explanatory variable *Depth*?

(b) As you discovered in part (a), it is always possible to fit quadratic models, but we must decide if they are helpful. Is this model as informative to the couple as the model in Example 29.15? What percent of variation in the total price is explained by using the quadratic regression model with *Depth*?

29.25 Tuition and Fees. Information regarding tuition and fees at the University of Virginia from 1970 to 2014 is provided in Table 29.7.[9] Use statistical software to answer the following questions. TUITN

(a) Find the simple linear regression equation for predicting tuition and fees from year, and save the residuals and fitted values.

(b) The value of tuition and fees in 1971 is missing from the data set. Use the least-squares line to estimate this value.

(c) Does the estimate obtained in part (b) intuitively make sense to you? That is, are you happy with this estimate? Explain.

(d) Plot the residuals against year. What does the plot tell you about the adequacy of the linear fit?

(e) Will this linear model overestimate or underestimate the tuition and fees at this college in the 1990s?

(f) Because the residual plot shows a quadratic trend, it might be helpful to add a quadratic term to this model. Fit the quadratic regression model and provide the estimated model.

TABLE 29.7 Out-of-state tuition and fees (in dollars) at the University of Virginia

Year	Tuition and Fees	Year	Tuition and Fees	Year	Tuition and Fees
1970	1,069	1985	4,886	2000	17,409
1971	missing	1986	5,468	2001	18,268
1972	1,372	1987	5,796	2002	19,805
1973	1,447	1988	6,336	2003	21,984
1974	1,569	1989	7,088	2004	22,700
1975	1,619	1990	8,136	2005	24,100
1976	1,819	1991	9,564	2006	25,945
1977	1,939	1992	10,826	2007	27,750
1978	2,024	1993	12,254	2008	29,600
1979	2,159	1994	13,052	2009	31,672
1980	2,402	1995	14,006	2010	33,574
1981	2,646	1996	14,434	2011	36,570
1982	3,276	1997	15,030	2012	38,018
1983	3,766	1998	15,814	2013	39,844
1984	4,336	1999	16,603	2014	42,184

(g) Does the quadratic model provide a better fit than the linear model?

(h) Would you be willing to make inferences based on the quadratic model? Explain.

29.26 Fish Sizes. Table 29.8 contains data on the size of perch caught in a lake in Finland.[10] Use statistical software to help you analyze these data. PERCH

(a) Use the multiple regression model with two explanatory variables, length and width, to predict the weight of a perch. Provide the estimated multiple regression equation.

(b) How much of the variation in the weight of perch is explained by the model in part (a)?

(c) Does the ANOVA table indicate that at least one of the explanatory variables is helpful in predicting the weight of perch? Explain.

TABLE 29.8 Measurements on 56 perch

Observation Number	Weight (grams)	Length (cm)	Width (cm)	Observation Number	Weight (grams)	Length (cm)	Width (cm)
104	5.9	8.8	1.4	132	197.0	27.0	4.2
105	32.0	14.7	2.0	133	218.0	28.0	4.1
106	40.0	16.0	2.4	134	300.0	28.7	5.1
107	51.5	17.2	2.6	135	260.0	28.9	4.3
108	70.0	18.5	2.9	136	265.0	28.9	4.3
109	100.0	19.2	3.3	137	250.0	28.9	4.6
110	78.0	19.4	3.1	138	250.0	29.4	4.2
111	80.0	20.2	3.1	139	300.0	30.1	4.6
112	85.0	20.8	3.0	140	320.0	31.6	4.8
113	85.0	21.0	2.8	141	514.0	34.0	6.0
114	110.0	22.5	3.6	142	556.0	36.5	6.4
115	115.0	22.5	3.3	143	840.0	37.3	7.8
116	125.0	22.5	3.7	144	685.0	39.0	6.9
117	130.0	22.8	3.5	145	700.0	38.3	6.7
118	120.0	23.5	3.4	146	700.0	39.4	6.3
119	120.0	23.5	3.5	147	690.0	39.3	6.4
120	130.0	23.5	3.5	148	900.0	41.4	7.5
121	135.0	23.5	3.5	149	650.0	41.4	6.0
122	110.0	23.5	4.0	150	820.0	41.3	7.4
123	130.0	24.0	3.6	151	850.0	42.3	7.1
124	150.0	24.0	3.6	152	900.0	42.5	7.2
125	145.0	24.2	3.6	153	1015.0	42.4	7.5
126	150.0	24.5	3.6	154	820.0	42.5	6.6
127	170.0	25.0	3.7	155	1100.0	44.6	6.9
128	225.0	25.5	3.7	156	1000.0	45.2	7.3
129	145.0	25.5	3.8	157	1100.0	45.5	7.4
130	88.0	26.2	4.2	158	1000.0	46.0	8.1
131	180.0	26.5	3.7	159	1000.0	46.6	7.6

(d) Do the individual t tests indicate that both β_1 and β_2 are significantly different from zero? Explain.

(e) Create a new variable, called interaction, that is the product of length and width. Use the multiple regression model with three explanatory variables, length, width, and interaction, to predict the weight of a perch. Provide the estimated multiple regression equation.

(f) How much of the variation in the weight of perch is explained by the model in part (e)?

(g) Does the ANOVA table indicate that at least one of the explanatory variables is helpful in predicting the weight of perch? Explain.

(h) Describe how the individual t statistics changed when the interaction term was added.

29.10 Inference for Regression Parameters

We discussed the general form of inference procedures for regression parameters earlier in the chapter, using software output. This section provides more details for the analysis of variance (ANOVA) table, the F test, and the individual t statistics for the multiple regression model with p explanatory variables, $\mu_y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$.

Software always provides the ANOVA table. The general form of the ANOVA table is shown below.

Source	Degrees of Freedom	Sum of Squares	Mean Square	F Statistic
Model	p	$\text{SSM} = \Sigma(\hat{y} - \bar{y})^2$	$\text{MSM} = \dfrac{SSM}{p}$	$F = \dfrac{MSM}{MSE}$
Error	$n - p - 1$	$\text{SSE} = \Sigma(y - \hat{y})^2$	$\text{MSE} = \dfrac{SSE}{n - p - 1}$	
Total	$n - 1$	$\Sigma(y - \bar{y})^2$		

EXAMPLE 29.25 A Quick Check

The final multiple regression model for the clothing retailer data in Example 29.24 is

$$\mu_y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$

where $x_1 = \textit{Purchase}12$, $x_2 = \textit{Purchase}12\textit{sq}$, and $x_3 = \textit{IntRecency}$. It is a good idea to check that the degrees of freedom from the ANOVA table on the output match the form above. This verifies that the software is using the number of observations and the number of explanatory variables that you intended. The model degrees of freedom is the number of explanatory variables, 3, and the total degrees of freedom (degrees of freedom for the model plus degrees of freedom for error) is the number of observations minus 1, $56 - 1 = 55$. We usually do not check the other calculations by hand, but knowing that the mean sum of squares is the sum of squares divided by the degrees of freedom and that the F statistic is the ratio of the mean sum of squares for each source helps us understand how the F statistic is formed.

Statistics In Your World

Do Good Looks Mean Good Money?

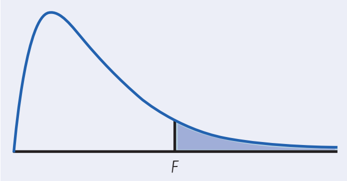
Experienced researchers who have spent decades studying physical attractiveness suggest that good looks translate into good money. In particular, studies suggest that "plain people earn 5% to 10% less than people of average looks, who in turn earn 3% to 8% less than those deemed good-looking." Other studies suggest that size is important also, with tall people earning considerably more over their careers than short people. Before you take a look in the mirror, it is important to understand that hiring managers say that the appearance of confidence is more attractive to them than physical beauty.

The first formal test in most multiple regression studies is the ANOVA F *test.* This test is used to check if the complete set of explanatory variables is helpful in predicting the response variable.

Analysis of Variance *F* Test

The analysis of variance F statistic tests the null hypothesis that all the regression coefficients (βs) except β_0 are equal to zero. The test statistic is

$$F = \frac{\text{Variation due to model}}{\text{Variation due to error}} = \frac{\text{MSM}}{\text{MSE}}$$



P-values come from the F distribution with p and $n - p - 1$ degrees of freedom.

To give formulas for the numerator and denominator of the F statistic, let $\hat{y}$ stand for predicted values and let $\bar{y}$ be the average of the response observations. The numerator of F is the mean square for the model:

$$\text{Variation due to model} = \frac{\Sigma(\hat{y} - \bar{y})^2}{p}$$

The denominator of F is the mean square for error:

$$\text{Variation due to error} = \frac{\Sigma(y - \hat{y})^2}{n - p - 1}$$

The P value for a test of H_0 against the alternative that at least one β parameter is not zero is the area to the right of F under an $F(p, n - p - 1)$ distribution.

EXAMPLE 29.26 Any Useful Predictors

The ANOVA table in Figure 29.19 (page 29-44) shows an F statistic of 270.18. The P-value provided on the output is the area to the right of 270.18 under an F distribution with 3 numerator and 52 denominator degrees of freedom. Because this area is so small (<0.001), we reject the hypothesis that the β coefficients associated with the three explanatory variables are all equal to zero. The three explanatory variables together do help predict the response.

As we have seen, individual t tests are helpful in identifying the explanatory variables that are useful predictors, but extreme caution is necessary when interpreting the results of these tests. Remember that an individual t assesses the contribution of its variable *after controlling for the effects of the other variables in this specific model.* That is, individual t's depend on the model in use, not just on the direct association between an explanatory variable and the response.

Confidence Intervals and Individual *t* Tests for Coefficients

A level C confidence interval for a regression coefficient β is $b \pm t^*\text{SE}_b$.

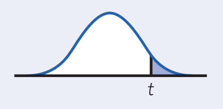
The critical value t^* is obtained from the t_{n-p-1} distribution.

The t statistic for testing the null hypothesis that a regression coefficient is equal to zero has the form

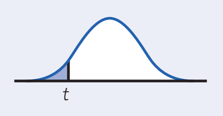
$$t = \frac{\text{Parameter estimate}}{\text{Standard error of estimate}} = \frac{b}{SE_b}$$

In terms of a random variable T having the t_{n-p-1} distribution, the P-value for a test of H_0 against

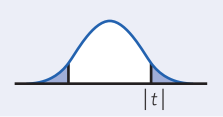
H_a: $\beta > 0$ is $P(T \geq t)$



H_a: $\beta < 0$ is $P(T \leq t)$



H_a: $\beta \neq 0$ is $2P(T \geq |t|)$



EXAMPLE 29.27 The Easiest Situation: All Predictors Are Helpful

The individual t statistics and corresponding P-values in Figure 29.19 (page 29-44) indicate that all three of the explanatory variables are useful predictors. All the P-values are below 0.01, which indicates very convincing evidence of statistical significance. The P-values are computed using a t distribution with 52 degrees of freedom. The degrees of freedom for error in the ANOVA table will always tell you which t distribution to use for the individual β coefficients.

The main purpose of most regression models is prediction. Construction of *confidence intervals for a mean response and prediction intervals for a future observation* with multiple regression models is similar to the methods we used for simple linear regression. The main difference is that we must now specify a list of values for all of the explanatory variables in the model. As we learned in Chapter 25, the additional uncertainty in predicting future observations will result in prediction intervals that are wider than confidence intervals.

Confidence and Prediction Intervals for Multiple Regression Response

A level C **confidence interval for the mean response** μ_y is $\hat{y} \pm t^*SE_{\hat{\mu}}$.

A level C **prediction interval for a single response** μ_y is $\hat{y} \pm t^*SE_{\hat{y}}$.

In both intervals, t^* is the critical value for the t_{n-p-1} density curve with area C between $-t$ and t.

EXAMPLE 29.28 Predicting Means and Future Clothing Purchases

Figure 29.20 provides the predicted values, 95% confidence limits for the mean purchase amount, and 95% prediction limits for a future purchase amount for each of the 56 observations in Table 29.6. The values of the explanatory variables don't appear, but they are needed to obtain the predicted values $\hat{y}$ and the endpoints of the intervals. As expected, the prediction intervals for future purchase amounts are always wider than the confidence intervals for the mean purchase amounts. You can also see that predicting future purchase amounts, even with a good model, is not an easy task. Several of the prediction intervals (for Observations 1 to 3, for example) include purchase amounts below zero. The manager will not give customers money for coming to the store, so the lower endpoint of the prediction intervals should be zero for practical purposes.

APPLY YOUR KNOWLEDGE

29.27 World Record Running Times. Exercise 29.15 (page 29-26) shows the progress of world record times (in seconds) for the 10,000-meter run for both men and women. DATA RECORD

(a) Provide the ANOVA table for the regression model with two regression lines, one for men and one for women.

(b) Are all the individual coefficients significantly different from zero? Set up the appropriate hypotheses, identify the test statistics and *P*-values, and make conclusions in the context of the problem.

29.28 Fish Sizes. Use explanatory variables length, width, and interaction from Exercise 29.26 (page 29-48) on the 56 perch to provide 95% confidence intervals for the mean and prediction intervals for future observations. Interpret both intervals for the 10th perch in the data set. What *t* distribution is used to provide both intervals? DATA PERCH

29.29 Clothing Retailer. Because the average purchase amount *Purchase*12 was such a good predictor, the manager would like you to consider another explanatory variable: the average purchase amount from the previous 12 months. Create the new variable DATA CLOTHE2

$$Purchase12b = \frac{Dollar24 - Dollar12}{Freq24 - Freq12}$$

and add it to the final model obtained in Example 29.24 (page 29-43).

(a) What is R^2 for this model? How does this value compare with R^2 in Example 29.24?

(b) What is the value of the individual *t* statistic for this new explanatory variable? How much did the individual *t* statistics change from their previous values?

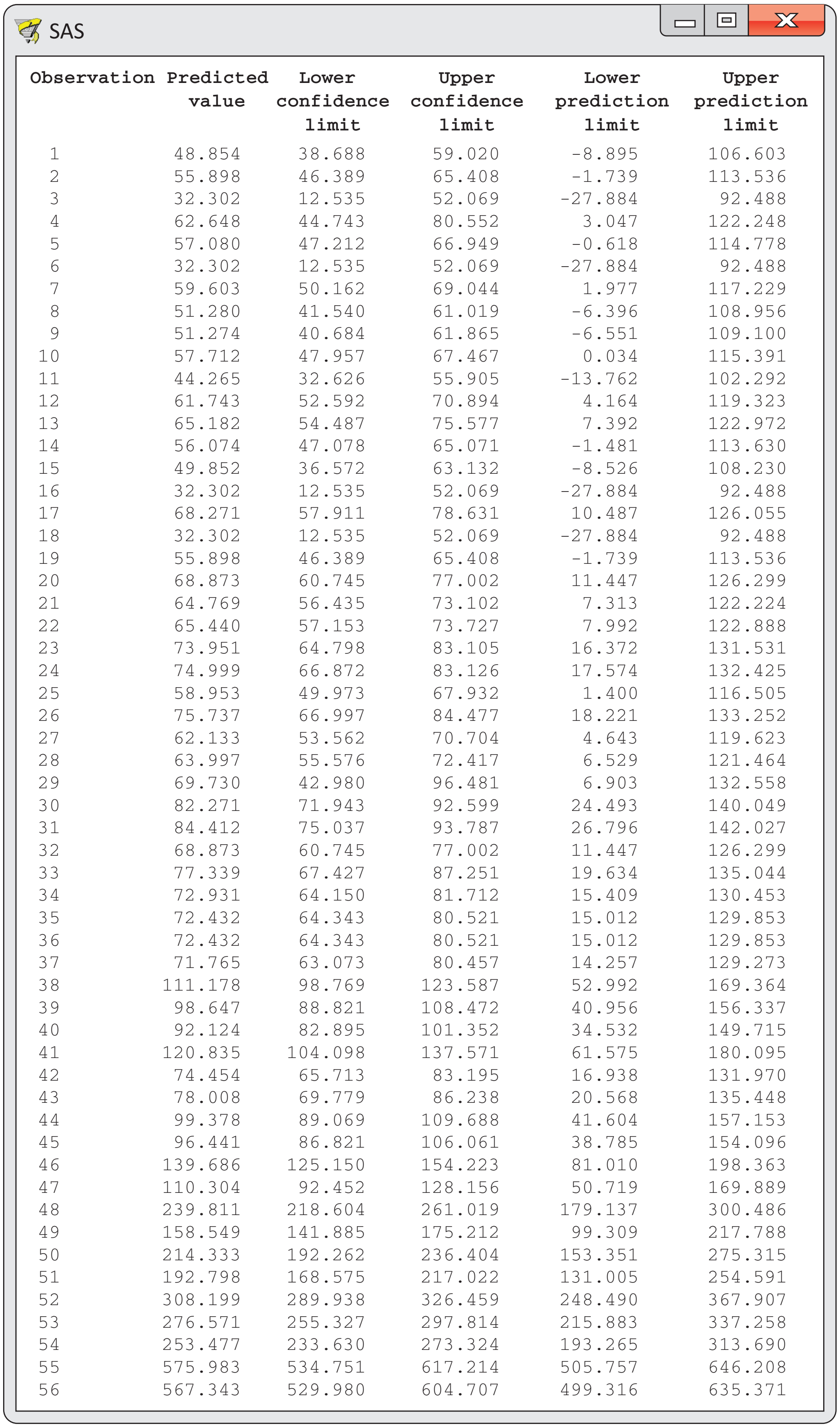
(c) Would you recommend this model over the model in Example 29.24? Explain.

29.11 Checking the Conditions for Inference

A full picture of the conditions for multiple regression requires much more than a few plots of the residuals. We will present only a few methods here because regression diagnostics is a subject that could fill an entire book.

Plot the response variable against each of the explanatory variables. These plots help you explore and understand potential relationships. Multiple regression models allow curvature and other interesting features that are not simple to check visually, especially when we get beyond two explanatory variables.

SAS



```
Observation Predicted   Lower      Upper       Lower      Upper
            value    confidence confidence  prediction prediction
                       limit      limit       limit      limit
   1         48.854    38.688     59.020     -8.895     106.603
   2         55.898    46.389     65.408     -1.739     113.536
   3         32.302    12.535     52.069    -27.884      92.488
   4         62.648    44.743     80.552      3.047     122.248
   5         57.080    47.212     66.949     -0.618     114.778
   6         32.302    12.535     52.069    -27.884      92.488
   7         59.603    50.162     69.044      1.977     117.229
   8         51.280    41.540     61.019     -6.396     108.956
   9         51.274    40.684     61.865     -6.551     109.100
  10         57.712    47.957     67.467      0.034     115.391
  11         44.265    32.626     55.905    -13.762     102.292
  12         61.743    52.592     70.894      4.164     119.323
  13         65.182    54.487     75.577      7.392     122.972
  14         56.074    47.078     65.071     -1.481     113.630
  15         49.852    36.572     63.132     -8.526     108.230
  16         32.302    12.535     52.069    -27.884      92.488
  17         68.271    57.911     78.631     10.487     126.055
  18         32.302    12.535     52.069    -27.884      92.488
  19         55.898    46.389     65.408     -1.739     113.536
  20         68.873    60.745     77.002     11.447     126.299
  21         64.769    56.435     73.102      7.313     122.224
  22         65.440    57.153     73.727      7.992     122.888
  23         73.951    64.798     83.105     16.372     131.531
  24         74.999    66.872     83.126     17.574     132.425
  25         58.953    49.973     67.932      1.400     116.505
  26         75.737    66.997     84.477     18.221     133.252
  27         62.133    53.562     70.704      4.643     119.623
  28         63.997    55.576     72.417      6.529     121.464
  29         69.730    42.980     96.481      6.903     132.558
  30         82.271    71.943     92.599     24.493     140.049
  31         84.412    75.037     93.787     26.796     142.027
  32         68.873    60.745     77.002     11.447     126.299
  33         77.339    67.427     87.251     19.634     135.044
  34         72.931    64.150     81.712     15.409     130.453
  35         72.432    64.343     80.521     15.012     129.853
  36         72.432    64.343     80.521     15.012     129.853
  37         71.765    63.073     80.457     14.257     129.273
  38        111.178    98.769    123.587     52.992     169.364
  39         98.647    88.821    108.472     40.956     156.337
  40         92.124    82.895    101.352     34.532     149.715
  41        120.835   104.098    137.571     61.575     180.095
  42         74.454    65.713     83.195     16.938     131.970
  43         78.008    69.779     86.238     20.568     135.448
  44         99.378    89.069    109.688     41.604     157.153
  45         96.441    86.821    106.061     38.785     154.096
  46        139.686   125.150    154.223     81.010     198.363
  47        110.304    92.452    128.156     50.719     169.889
  48        239.811   218.604    261.019    179.137     300.486
  49        158.549   141.885    175.212     99.309     217.788
  50        214.333   192.262    236.404    153.351     275.315
  51        192.798   168.575    217.022    131.005     254.591
  52        308.199   289.938    326.459    248.490     367.907
  53        276.571   255.327    297.814    215.883     337.258
  54        253.477   233.630    273.324    193.265     313.690
  55        575.983   534.751    617.214    505.757     646.208
  56        567.343   529.980    604.707    499.316     635.371
```

FIGURE 29.20
Predicted values, confidence limits for the mean purchase amount, and prediction limits for a future purchase amount, for Example 29.28.

Plot the residuals against the predicted values and against all of the explanatory variables in the model. These plots will allow you to check the condition that the standard deviation of the response about the multiple regression model is the same everywhere. They should show an unstructured horizontal band of points centered at 0. The mean of the residuals is always 0, just as in simple linear regression, so we continue to add a line at 0 to orient ourselves. Funnel or cone shapes indicate that this condition is not met and that the standard deviation of the residuals must be stabilized before making inferences. Other patterns in residual plots can sometimes be fixed by changing the model. For example, if you see a quadratic pattern, then you should consider adding a quadratic term for that explanatory variable.

Look for outliers and influential observations in all residual plots. To check the influence of a particular observation, you can fit your model with and without this observation. If the estimates and statistics do not change much, you can safely proceed. However, if there are substantial changes, you must begin a more careful investigation. Do not simply throw out observations to improve the fit and increase R^2. Ideally, we would like all of the explanatory variables to be independent and the observations on the response variable to be independent. As you have seen in this chapter, practical problems include explanatory variables that are not independent. Association between two or more explanatory variables can create serious problems in the model, so use correlations and scatterplots to check relationships.

To check the condition that the response should vary Normally about the multiple regression model, *make a histogram or stemplot of the residuals.* We can rely on the robustness of the regression methods when there is a slight departure from Normality, except for prediction intervals. As in the case of simple linear regression, we view prediction intervals from multiple regression models as rough approximations.

EXAMPLE 29.29 Checking Conditions

Figure 29.21 shows residual plots for the final model in Example 29.24 (page 29-43). The scatterplot shows that the variability for the larger predicted values is greater than the variability for the predicted values below 200. The constant-variance condition is not satisfied. Because most of the predicted values are below 200 and

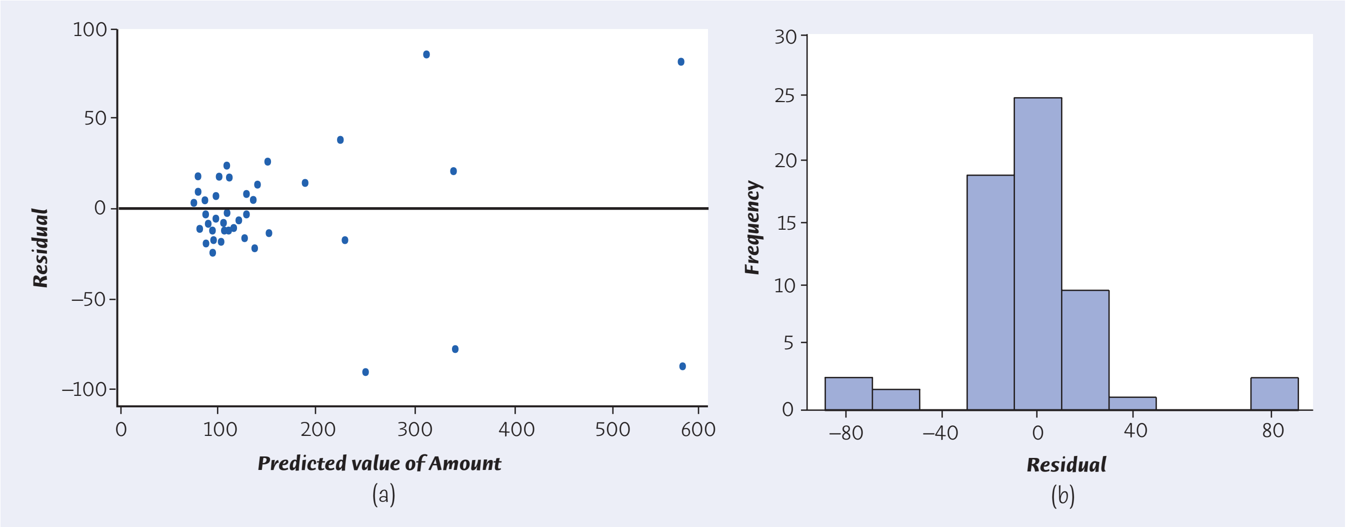


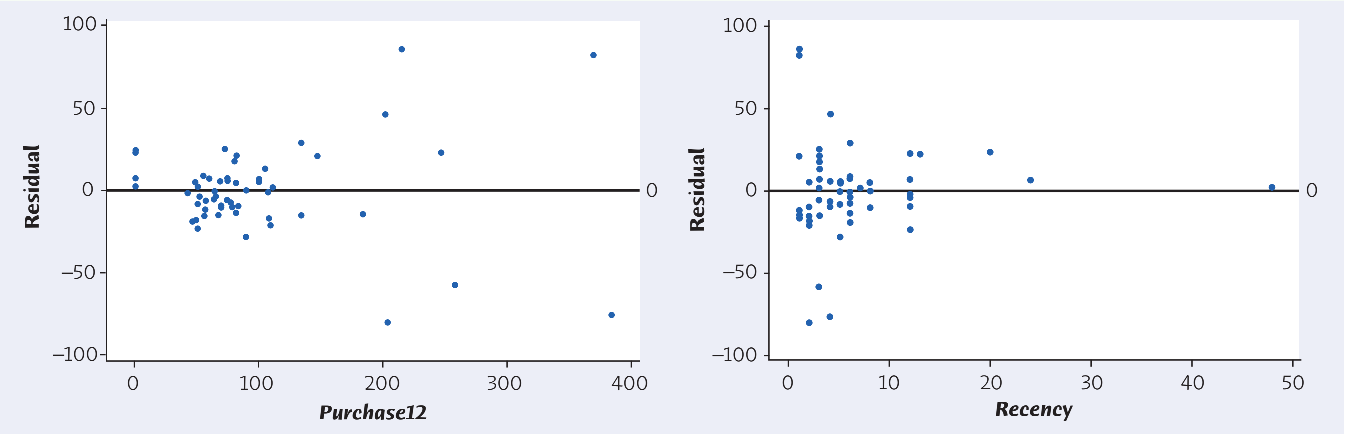
FIGURE 29.21
Residual plots for the multiple regression model in Example 29.24. (a) A scatterplot of the residuals against the predicted values. (b) A histogram of the residuals.

the variability is roughly constant in that range, we will not resort to more sophisticated methods to stabilize the variance.

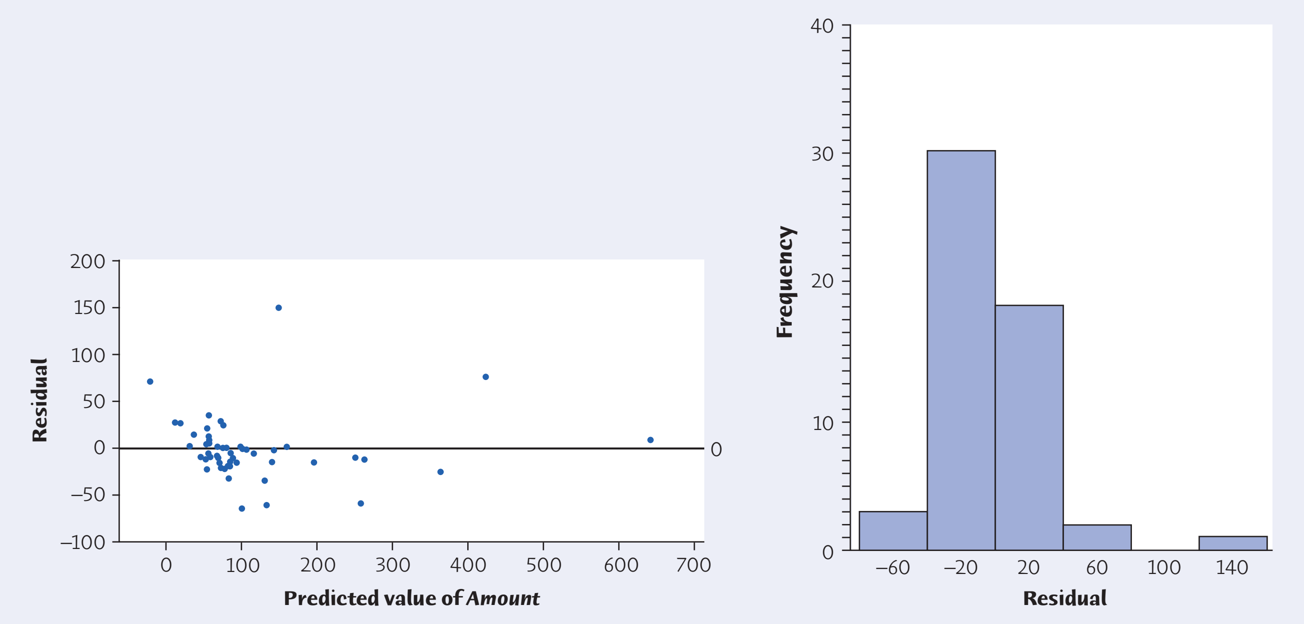
The histogram shows approximate perfect symmetry in the residuals. The residuals above 75 and below −75 are apparent on the scatterplot and the histogram. This is a situation in which we need to rely on the robustness of regression inference when there are slight departures from Normality.

APPLY YOUR KNOWLEDGE

29.30 Final Model for the Clothing Retailer Problem. The two residual plots below show the residuals for the final model in the clothing retailer problem plotted against *Purchase12* and *Recency*. Do the plots suggest any potential problems with the conditions for inference? Comment.



29.31 The Clothing Retailer Problem. The accompanying scatterplot and histogram show the residuals from the model in Example 29.20 with all explanatory variables, some interaction terms, and quadratic terms. Comment on both plots. Do you see any reason for concern in using this model for inference?



CHAPTER 29 SUMMARY

Chapter Specifics

- An indicator variable x_2 can be used to fit a regression model with two parallel lines. The mean response is

$$\mu_y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

  where x_1 is an explanatory variable.
- A multiple regression model with **two regression lines** includes an explanatory variable x_1, an indicator variable x_2, and an interaction term x_1x_2. The mean response is

$$\mu_y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$$

- The mean response μ_y for a general **multiple regression model** based on p explanatory variables $x_1, x_2, \ldots, x_p$ is

$$\mu_y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

- The **estimated regression model** is

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_p x_p$$

  where the b's are obtained by the method of least squares.
- The regression standard error s has $n - p - 1$ degrees of freedom and is used to estimate σ.
- The **sum of squares row in the analysis of variance (ANOVA) table** breaks the total variability in the responses into two pieces. One piece summarizes the variability due to the model, and the other piece summarizes the variability due to error:

  Total sum of squares = Model sum of squares + Error sum of squares

- The **squared multiple correlation coefficient** R^2 represents the proportion of variability in the response variable y that is explained by the explanatory variables $x_1, x_2, \ldots, x_p$ in a multiple regression model.
- To test the hypothesis that all the regression coefficients (β's), except β_0, are equal to zero, use the **ANOVA *F* statistic**. In other words, the **null model** says that the x's do not help predict y. The alternative is that the explanatory variables as a group are helpful in predicting y.
- **Individual *t* procedures** in regression inference have $n - p - 1$ degrees of freedom. These individual t procedures depend on the other explanatory variables specified in a multiple regression model. Individual t tests assess the contribution of one explanatory variable after controlling for the effects of the other variables in a model. The null hypothesis is written as H_0: $\beta = 0$ but interpreted as "the coefficient of x is 0 *in this model.*"
- **Confidence intervals** for the mean response μ_y have the form $\hat{y} \pm t^* SE_{\hat{\mu}}$. **Prediction intervals** for individual future responses y have the form $\hat{y} \pm t^* SE_{\hat{y}}$.

Statistics in Summary

Here are the most important skills you should have acquired from reading this chapter.

A. Preliminaries

1. Examine the data for outliers and other deviations that might influence your conclusions.
2. Use descriptive statistics, especially correlations, to get an idea of which explanatory variables may be most helpful in explaining the response.
3. Make scatterplots to examine the relationships between explanatory variables and a response variable.
4. Use software to compute a correlation matrix to explore the relationships between pairs of variables.

B. Recognition

1. Recognize when a multiple regression model with parallel regression lines is appropriate.

2. Recognize when an interaction term needs to be added to fit a multiple regression model with two separate regression lines.
3. Recognize when a multiple regression model with several explanatory variables is appropriate.
4. Recognize the difference between the overall *F* test and the individual *t* tests.
5. Recognize that the parameter estimates, *t* statistics, and *P*-values for each explanatory variable depend on the specific model.
6. Inspect the data to recognize situations in which inference isn't safe: influential observations, strongly skewed residuals in a small sample, or nonconstant variation of the data points about the regression model.

C. Inference Using Software

1. Use software to find the estimated multiple regression model.
2. Explain the meaning of the regression parameters (β's) in any specific multiple regression model.
3. Understand the software output for regression. Find the regression standard error, the squared multiple correlation coefficient R^2, and the overall *F* test and *P*-value. Identify the parameter estimates, standard errors, individual *t* tests, and *P*-values.
4. Use that information to carry out tests and calculate confidence intervals for the β's.
5. Use R^2 and residual plots to assess the fit of a model.
6. Choose a model by comparing R^2-values, regression standard errors, and individual *t* statistics.
7. Explain the distinction between a confidence interval for the mean response and a prediction interval for an individual response.

Link It

Chapters 4, 5, and 25 discuss scatterplots, correlation, and regression. In these chapters, we studied how to use a single explanatory variable to predict a response, although in Chapter 4 we saw how to incorporate a categorical variable into a scatterplot. In this chapter, we extend the ideas of Chapters 4, 5, and 25 and learn how to use several explanatory variables to predict a response. The multiple regression model is similar to the simple linear regression model but with more explanatory variables. The conditions for inference, the methods for estimating and testing hypotheses about regression coefficients, prediction, and checking the conditions for inference are much like those discussed in Chapter 25. New concepts include the notion of an interaction, deciding which of several candidate regression models is best, and interpreting parameter estimates when several explanatory variables are included. We will encounter some of these new concepts again in Chapter 30.

Macmillan Learning Online Resources

If you are having difficulty with any of the sections of this chapter, this online resource should help prepare you to solve the exercises at the end of this chapter:

- LearningCurve provides you with a series of questions about the chapter that adjust to your level of understanding.

CHECK YOUR SKILLS

Many exercise bikes, elliptical trainers, and treadmills display basic information like distance, speed, calories burned per hour (or total calories), and duration of the workout. The data in Table 29.9 show the treadmill display's claimed calories per hour by speed for a 175-pound male using a Cybex treadmill at inclines of 0%, 2%, and 4%.

The relationship between speed and calories is different for walking and running, so we need an indicator for slow/fast. The variables created from Table 29.9 are

Calories = calories burned per hour

Mph = speed of the treadmill

Incline = the incline percent 0, 2, or 4

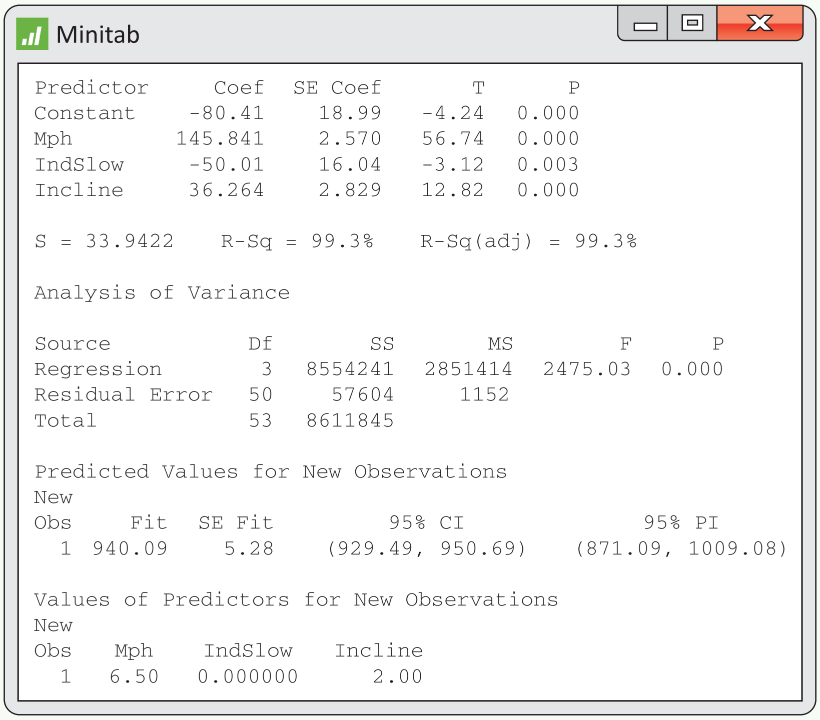
IndSlow = 1 for Mph $\leq$ 3 and IndSlow = 0 for Mph $>$ 3

Here is part of the Minitab output from fitting a multiple regression model to predict Calories from Mph, IndSlow, and Incline for the Cybex treadmill. Exercises 29.32 to 29.41 are based on this output.

TABLE 29.9 Cybex treadmill display's claimed calories per hour by speed and incline for a 175-pound man

	Incline		
Mph	0%	2%	4%
1.5	174	207	240
2.0	205	249	294
2.5	236	291	347
3.0	267	333	400
3.5	372	436	503
4.0	482	542	607
4.5	592	649	709
5.0	701	756	812
5.5	763	824	885
6.0	825	892	959
6.5	887	960	1032
7.0	949	1027	1105
7.5	1011	1094	1178
8.0	1073	1163	1252
8.5	1135	1230	1325
9.0	1197	1298	1398
9.5	1259	1365	1470
10.0	1321	1433	1544

Minitab



```
Predictor       Coef   SE Coef        T       P
Constant      -80.41     18.99    -4.24   0.000
Mph          145.841     2.570    56.74   0.000
IndSlow       -50.01     16.04    -3.12   0.003
Incline       36.264     2.829    12.82   0.000

S = 33.9422    R-Sq = 99.3%    R-Sq(adj) = 99.3%

Analysis of Variance

Source            Df        SS         MS         F       P
Regression         3   8554241    2851414   2475.03   0.000
Residual Error    50     57604       1152
Total             53   8611845

Predicted Values for New Observations
New
Obs     Fit   SE Fit          95% CI                  95% PI
  1  940.09     5.28    (929.49, 950.69)    (871.09, 1009.08)

Values of Predictors for New Observations
New
Obs    Mph    IndSlow     Incline
  1   6.50   0.000000        2.00
```

29.32 The number of parameters in this multiple regression model is

(a) 4. (b) 5. (c) 6.

29.33 The equation for predicting calories from these explanatory variables is

(a) $Calories = -80.41 + 145.84Mph - 50.01IndSlow + 36.26Incline$.

(b) $Calories = -4.24 + 56.74Mph - 3.12IndSlow + 12.82Incline$.

(c) $Calories = 18.99 + 2.57Mph + 16.04IndSlow + 2.83Incline$.

29.34 The regression standard error for these data is

(a) 0.993. (b) 33.94. (c) 1152.

29.35 To predict calories when walking ($Mph \leq 3$) with no incline, use the line

(a) $-80.41 + 145.84Mph$.

(b) $(-80.41 - 50.01) + 145.84Mph$.

(c) $[-80.41 + (2 \times 36.26)] + 145.84Mph$.

29.36 To predict calories when running ($Mph > 3$) with no incline, use the line

(a) $-80.41 + 145.84Mph$.

(b) $(-80.41 + 36.26) + 145.84Mph$.

(c) $[-80.41 + (2 \times 36.26)] + 145.84Mph$.

29.37 To predict calories when running on a 2% incline, use the line

(a) $-80.41 + 145.84Mph$.

(b) $(-80.41 - 50.01) + 145.84Mph$.

(c) $[-80.41 + (2 \times 36.26)] + 145.84Mph$.

29.38 Is there significant evidence that more calories are burned for higher speeds? To answer this question, test the hypotheses

(a) H_0: $\beta_0 = 0$ versus H_a: $\beta_0 > 0$.

(b) H_0: $\beta_1 = 0$ versus H_a: $\beta_1 > 0$.

(c) H_0: $\beta_1 = 0$ versus H_a: $\beta_1 \neq 0$.

29.39 Confidence intervals and tests for these data use the t distribution with degrees of freedom

(a) 3. (b) 50. (c) 53.

29.40 Orlando, a 175-pound man, plans to run 6.5 miles per hour for one hour on a 2% incline. He can be 95% confident that he will burn between

(a) 871 and 1009 calories.

(b) 929 and 951 calories.

(c) 906 and 974 calories.

29.41 Suppose that we also have data on a second treadmill, made by LifeFitness. An indicator variable for brand of treadmill, say *Treadmill* = 1 for Cybex and *Treadmill* = 0 for LifeFitness, is created for a new model. If the three explanatory variables above and the new indicator variable *Treadmill* were used to predict *Calories*, how many β parameters would need to be estimated in the new multiple regression model?

(a) 4 (b) 5 (c) 6

CHAPTER 29 EXERCISES

29.42 **A computer game.** A multimedia statistics learning system includes a test of skill in using the computer's mouse. The software displays a circle at a random location on the computer screen. The subject clicks in the circle with the mouse as quickly as possible. A new circle appears as soon as the subject clicks the old one. Table 5.3 (text page 161) gives data for one subject's trials, 20 with each hand. Distance is the distance from the cursor location to the center of the new circle, in units whose actual size depends on the size of the screen. Time is the time required to click in the new circle, in milliseconds.[11] COMGAME

(a) Specify the population multiple regression model for predicting time from distance separately for each hand. Make sure you include the interaction term that is necessary to allow for the possibility of having different slopes. Explain in words what each β in your model means.

(b) Use statistical software to find the estimated multiple regression equation for predicting time from distance separately for each hand. What percent of variation in the distances is explained by this multiple regression model?

(c) Explain how to use the estimated multiple regression equation in part (b) to obtain the least-squares line for each hand. Draw these lines on a scatterplot of time versus distance.

29.43 **Bank wages and length of service.** We assume that our wages will increase as we gain experience and become more valuable to our employers. Wages also increase because of inflation. By examining a sample of employees at a given point in time, we can look at part of the picture. How does length of service (LOS) relate to wages? Table 29.10 gives data on the LOS in months and wages for 60 women who work in Indiana banks.

TABLE 29.10 **Bank wages, length of service, and bank size**

Wages	Length of Service	Size	Wages	Length of Service	Size	Wages	Length of Service	Size
48.3355	94	Large	64.1026	24	Large	41.2088	97	Small
49.0279	48	Small	54.9451	222	Small	67.9096	228	Small
40.8817	102	Small	43.8095	58	Large	43.0942	27	Large
36.5854	20	Small	43.3455	41	Small	40.7000	48	Small
46.7596	60	Large	61.9893	153	Large	40.5748	7	Large
59.5238	78	Small	40.0183	16	Small	39.6825	74	Small
39.1304	45	Large	50.7143	43	Small	50.1742	204	Large
39.2465	39	Large	48.8400	96	Large	54.9451	24	Large
40.2037	20	Large	34.3407	98	Large	32.3822	13	Small
38.1563	65	Small	80.5861	150	Large	51.7130	30	Large
50.0905	76	Large	33.7163	124	Small	55.8379	95	Large
46.9043	48	Small	60.3792	60	Large	54.9451	104	Large
43.1894	61	Small	48.8400	7	Large	70.2786	34	Large
60.5637	30	Large	38.5579	22	Small	57.2344	184	Small
97.6801	70	Large	39.2760	57	Large	54.1126	156	Small
48.5795	108	Large	47.6564	78	Large	39.8687	25	Large
67.1551	61	Large	44.6864	36	Large	27.4725	43	Small
38.7847	10	Small	45.7875	83	Small	67.9584	36	Large
51.8926	68	Large	65.6288	66	Large	44.9317	60	Small
51.8326	54	Large	33.5775	47	Small	51.5612	102	Large

Wages are yearly total income divided by the number of weeks worked. We have multiplied wages by a constant for reasons of confidentiality.[12] BWAGES

(a) Plot wages versus LOS using different symbols for size of the bank. There is one woman with relatively high wages for her length of service. Circle this point and do not use it in the rest of this exercise.

(b) Would you be willing to use a multiple regression model with parallel slopes to predict wages from LOS for the two different bank sizes? Explain.

(c) Fit a model that will allow you to test the hypothesis that the slope of the regression line for small banks is equal to the slope of the regression line for large banks. Conduct the test for equal slopes.

(d) Are the conditions for inference met for your model in part (c)? Construct appropriate residual plots and comment.

29.44 Mean annual temperatures for two California cities. Table 29.11 contains data on the mean annual temperatures (degrees Fahrenheit) for the years 1951–2014 at two locations in California: Pasadena and Redding.[13] TEMPS

(a) Plot the temperatures versus year using different symbols for the two cities.

(b) Would you be willing to use a multiple regression model with parallel slopes to predict temperatures from year for the two different cities? Explain.

(c) Fit a model that will allow you to test the hypothesis that the slope of the regression line for Pasadena is

TABLE 29.11 Mean annual temperatures (°F) in two California cities

	Mean Temperature			Mean Temperature			Mean Temperature	
Year	Pasadena	Redding	Year	Pasadena	Redding	Year	Pasadena	
1951	62.27	62.02	1977	64.47	63.89	2003	66.31	63.13
1952	61.59	62.27	1978	64.21	64.05	2004	65.71	63.57
1953	62.64	62.06	1979	63.76	60.38	2005	66.40	62.62
1954	62.88	61.65	1980	65.02	60.04	2006	66.80	62.60
1955	61.75	62.48	1981	65.80	61.95	2007	66.50	62.70
1956	62.93	63.17	1982	63.50	59.14	2008	67.70	62.90
1957	63.72	62.42	1983	64.19	60.66	2009	66.90	62.70
1958	65.02	64.42	1984	66.06	61.72	2010	65.60	61.40
1959	65.69	65.04	1985	64.44	60.50	2011	65.50	61.10
1960	64.48	63.07	1986	65.31	61.76	2012	67.50	62.40
1961	64.12	63.50	1987	64.58	62.94	2013	67.30	63.40
1962	62.82	63.97	1988	65.22	63.70	2014	69.80	65.00
1963	63.71	62.42	1989	64.53	61.50			
1964	62.76	63.29	1990	64.96	62.22			
1965	63.03	63.32	1991	65.60	62.73			
1966	64.25	64.51	1992	66.07	63.59			
1967	64.36	64.21	1993	65.16	61.55			
1968	64.15	63.40	1994	64.63	61.63			
1969	63.51	63.77	1995	65.43	62.62			
1970	64.08	64.30	1996	65.76	62.93			
1971	63.59	62.23	1997	66.72	62.48			
1972	64.53	63.06	1998	64.12	60.23			
1973	63.46	63.75	1999	64.85	61.88			
1974	63.93	63.80	2000	66.25	61.58			
1975	62.36	62.66	2001	64.96	63.03			
1976	64.23	63.51	2002	65.10	63.28			

equal to the slope of the regression line for Redding. Conduct the test for equal slopes.

(d) Are the conditions for inference met for your model in part (c)? Construct appropriate residual plots and comment.

29.45 **Growth of pine trees.** The Department of Biology at Kenyon College conducted an experiment to study the growth of pine trees. In April 1990, volunteers planted 1000 white pine (Pinus strobus) seedlings at the Brown Family Environmental Center. The seedlings were planted in two grids, distinguished by 10- and 15-foot spacings between the seedlings. Table 29.12 (page 29-60) shows the first 10 rows of a subset of the data collected by students at Kenyon College.[14] SEEDS

Variable	Description
Row	Row number in pine plantation
Col	Column number in pine plantation
*Hgt*90	Tree height at time of planting (cm)
*Hgt*96	Tree height in September 1996 (cm)
*Diam*96	Tree trunk diameter in September 1996 (cm)
*Hgt*97	Tree height in September 1997 (cm)
*Diam*97	Tree trunk diameter in September 1997 (cm)
*Spread*97	Widest lateral spread in September 1997 (cm)
*Needles*97	Needle length in September 1997 (mm)
*Deer*95	Type of deer damage in September 1995: 1 = none, 2 = browsed
*Deer*97	Type of deer damage in September 1997: 1 = none, 2 = browsed
*Cover*95	Amount of thorny cover in September 1995: 0 = none, 1 = < 1/3, 2 = between 1/3 and 2/3, 3 = > 2/3
Fert	Indicator for fertilizer: 0 = no, 1 = yes
Spacing	Distance (in feet) between trees (10 or 15)

(a) Use tree height at the time of planting (*Hgt*90) and the indicator variable for fertilizer (*Fert*) to fit a multiple regression model for predicting *Hgt*97. Specify the estimated regression model and the regression standard error. Are you happy with the fit of this model? Comment on the value of R^2 and the plot of the residuals against the predicted values.

(b) Construct a correlation matrix with *Hgt*90, *Hgt*96, *Diam*96, *Grow*96, *Hgt*97, *Diam*97, *Spread*97, and *Needles*97. Which variable is most strongly correlated with the response variable of interest (*Hgt*97)? Does this make sense to you?

(c) Add tree height in September 1996 (*Hgt*96) to the model in part (a). Does this model do a better job of predicting tree height in 1997? Explain.

(d) What happened to the individual t statistic for *Hgt*90 when *Hgt*96 was added to the model? Explain why this change occurred.

(e) Fit a multiple regression model for predicting *Hgt*97 based on the explanatory variables *Diam*97, *Hgt*96, and *Fert*. Summarize the results of the individual t tests. Does this model provide a better fit than the previous models? Explain by comparing the values of R^2 and s for each model.

(f) Does the parameter estimate for the variable indicating whether a tree was fertilized or not have the sign you expected? Explain. (Experiments can produce surprising results!)

(g) Do you think that the model in part (e) should be used for predicting growth in other pine seedlings? Think carefully about the conditions for inference.

29.46 **Heating a home.** The Sanchez household is about to install solar panels to reduce the cost of heating their house. In order to know how much the solar panels help, they record their consumption of natural gas before the solar panels are installed. Gas consumption is higher in cold weather, so the relationship between outside temperature and gas consumption is important. Here are the data for 16 consecutive months.[15] HEATING

Month	Nov.	Dec.	Jan.	Feb.	Mar.	Apr.	May	June
Degree-days	24	51	43	33	26	13	4	0
Gas used	6.3	10.9	8.9	7.5	5.3	4.0	1.7	1.2
Month	**July**	**Aug.**	**Sept.**	**Oct.**	**Nov.**	**Dec.**	**Jan.**	**Feb.**
Degree-days	0	1	6	12	30	32	52	30
Gas used	1.2	1.2	2.1	3.1	6.4	7.2	11.0	6.9

Outside temperature is recorded in degree-days, a common measure of demand for heating. A day's degree-days are the number of degrees its average temperature falls below 65°F. Gas used is recorded in hundreds of cubic feet.

(a) Create an indicator variable, say *INDwinter*, which is 1 for the months of November, December, January, and February. Make a plot of all the data using a different symbol for winter months.

(b) Fit the model with two regression lines, one for winter months and one for other months, and identify the estimated regression lines.

(c) Do you think that two regression lines were needed to explain the relationship between gas used and degree-days? Explain.

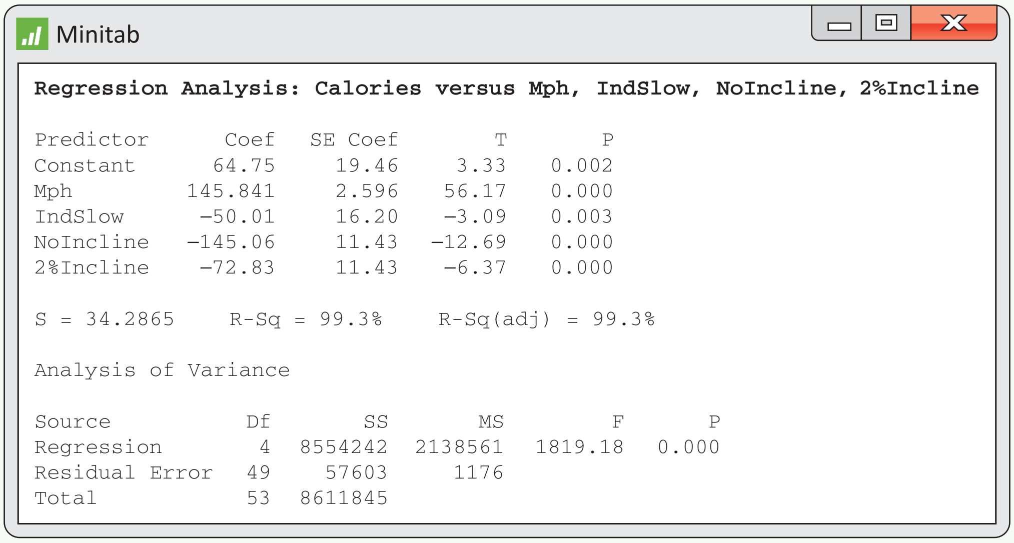
29.47 **Burning calories with exercise.** Many exercise bikes, elliptical trainers, and treadmills display basic information like distance, speed, calories burned per hour (or total calories), and duration of the workout. Let's take another look at the data in Table 29.9 (page 29-56) that were

TABLE 29.12 Measurements on pine seedlings at Brown Family Environmental Center

Row	Col	Hgt90	Hgt96	Diam96	Grow96	Hgt97	Diam97	Spread97	Needles97	Deer95	Deer97	Cover95	Fert	Spacing
1	1	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	0	0	15
1	2	14.0	284.0	4.2	96.0	362	6.60	162	66.0	0	1	2	0	15
1	3	17.0	387.0	7.4	110.0	442	9.30	250	77.0	0	0	1	0	15
1	4	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	0	0	15
1	5	24.0	294.0	3.9	70.0	369	7.00	176	72.0	0	0	2	0	15
1	6	22.0	310.0	5.6	84.0	365	6.90	215	76.0	0	0	1	0	15
1	7	18.0	318.0	5.4	96.0	356	7.60	238	74.5	0	0	0	0	15
1	8	32.0	328.0	5.4	88.0	365	7.70	219	60.5	0	0	1	0	15
1	9	n.a.	157.0	1.3	64.0	208	2.00	127	56.0	1	1	2	0	15
1	10	22.0	282.0	4.5	83.0	329	6.10	209	79.5	0	1	2	1	15

Note: n.a. indicates that data are not available.

Minitab



```
Regression Analysis: Calories versus Mph, IndSlow, NoIncline, 2%Incline

Predictor       Coef   SE Coef        T       P
Constant       64.75     19.46     3.33   0.002
Mph          145.841     2.596    56.17   0.000
IndSlow       -50.01     16.20    -3.09   0.003
NoIncline    -145.06     11.43   -12.69   0.000
2%Incline     -72.83     11.43    -6.37   0.000

S = 34.2865    R-Sq = 99.3%    R-Sq(adj) = 99.3%

Analysis of Variance

Source          Df        SS       MS        F      P
Regression       4   8554242  2138561  1819.18  0.000
Residual Error  49     57603     1176
Total           53   8611845
```

used for the Check Your Skills exercises. Scatterplots show different linear relationships for each incline, one for slow speeds and another for faster speeds, so the following indicator variables were created: TRDMILL

IndSlow = 1 for *Mph* $\le$ 3 and *IndSlow* = 0 for *Mph* > 3.

NoIncline = 1 for 0% incline and *NoIncline* = 0 for other inclines

2%*Incline* = 1 for a 2% incline and 2%*Incline* = 0 for other inclines

Above is part of the Minitab output from fitting a multiple regression model to predict *Calories* from *Mph*, *IndSlow*, *NoIncline*, and 2%*Incline* for the Cybex.

(a) Use the Minitab output to estimate each parameter in this multiple regression model for predicting calories burned with the Cybex machine. Don't forget to estimate σ.

(b) How many separate lines are fitted with this model? Do the lines all have the same slope? Identify each fitted line.

(c) Do you think that this model provides a good fit for these data? Explain.

(d) Is there significant evidence that more calories are burned for higher speeds? State the hypotheses, identify the test statistic and *P*-value, and provide a conclusion in the context of this question.

29.48 Burning calories with exercise. Table 29.13 provides data on speed and calories burned per hour for a 175-pound male using two different treadmills (a Cybex and a LifeFitness) at inclines of 0%, 2%, and 4%. TRDMILL2

(a) Create a scatterplot of calories against miles per hour using six different plotting symbols, one for each combination of incline level and machine.

(b) Create an indicator variable for brand of treadmill, say *Treadmill* = 1 for Cybex and *Treadmill* = 0 for LifeFitness. Fit a multiple regression model to predict *Calories* from *Mph*, *IndSlow*, *NoIncline*, 2%*Incline*, and *Treadmill*.

(c) Does the model provide a good fit for these data? Explain.

(d) Is there a significant difference in the relationship between calories and speed for the two different treadmills?

29.49 Metabolic rate and body mass. Metabolic rate, the rate at which the body consumes energy, is important in studies of weight gain, dieting, and exercise. The accompanying table gives data on the lean body mass and resting metabolic rate for 12 women and 7 men who are subjects in a study of dieting. Lean body mass, given in kilograms, is a person's weight leaving out all fat. Metabolic rate is measured in calories burned per 24 hours, the same calories used to describe the energy content of foods. The researchers believe that lean body mass is an important influence on metabolic rate. METAB2

Subject	Sex	Mass	Rate	Subject	Sex	Mass	Rate
1	M	62.0	1792	11	F	40.3	1189
2	M	62.9	1666	12	F	33.1	913
3	F	36.1	995	13	M	51.9	1460
4	F	54.6	1425	14	F	42.4	1124
5	F	48.5	1396	15	F	34.5	1052
6	F	42.0	1418	16	F	51.1	1347
7	M	47.4	1362	17	F	41.2	1204
8	F	50.6	1502	18	M	51.9	1867
9	F	42.0	1256	19	M	46.9	1439
10	M	48.7	1614				

TABLE 29.13 Treadmill display's claimed calories per hour by speed for a 175-pound man

	Incline			Incline		
Mph	Cybex-0%	Cybex-2%	Cybex-4%	Life-0%	Life-2%	Life-4%
1.5	174	207	240	178	212	246
2.0	205	249	294	10	256	301
2.5	236	291	347	243	300	356
3.0	267	333	400	276	343	411
3.5	372	436	503	308	387	466
4.0	482	542	607	341	431	522
4.5	592	649	709	667	718	769
5.0	701	756	812	732	789	845
5.5	763	824	885	797	860	922
6.0	825	892	959	863	930	998
6.5	887	960	1032	928	1015	1075
7.0	949	1027	1105	993	1072	1151
7.5	1011	1094	1178	1058	1143	1228
8.0	1073	1163	1252	1123	1214	1304
8.5	1135	1230	1325	1189	1285	1381
9.0	1197	1298	1398	1254	1356	1457
9.5	1259	1365	1470	1319	1426	1534
10.0	1321	1433	1544	1384	1497	1610

(a) Make a scatterplot of the data, using different symbols or colors for men and women. Summarize what you see in the plot.

(b) Use the model with two regression lines to predict metabolic rate from lean body mass for the different genders. Summarize the results.

(c) The parameter associated with the interaction term is often used to decide if a model with parallel regression lines can be used. Test the hypothesis that this parameter is equal to zero, and comment on whether or not you would be willing to use the more restrictive model with parallel regression lines for these data.

29.50 Student achievement and self-concept. In order to determine if student achievement is related to self-concept, as measured by the Piers-Harris Children's Self-Concept Scale, data were collected on 78 seventh-grade students from a rural midwestern school. Table 29.14 shows the records for the first 10 students on the following variables:[16] ACHIEVE

Variable	Description
OBS	Observation number ($n = 78$, some gaps in numbers)
GPA	GPA from school records
IQ	IQ test score from school records
AGE	Age in years, self-reported
SEX	$1 = \text{F}$, $2 = \text{M}$, self-reported
RAW	Raw score on Piers-Harris Children's Self-Concept Scale
C1	Cluster 1 within self-concept: behavior
C2	Cluster 2: school status
C3	Cluster 3: physical
C4	Cluster 4: anxiety
C5	Cluster 5: popularity
C6	Cluster 6: happiness

We will investigate the relationship between GPA and only three of the explanatory variables:

- IQ, the student's score on a standard IQ test
- C2, the student's self-assessment of his or her school status
- C5, the student's self-assessment of his or her popularity

Use statistical software to analyze the relationship between students' GPA and their IQ, self-assessed school status (C2), and self-assessed popularity (C5).

(a) One observation is an extreme outlier when all three explanatory variables are used. Which observation

TABLE 29.14 Student achievement and self-concept scores data for 78 seventh-grade students

OBS	GPA	IQ	AGE	SEX	RAW	C1	C2	C3	C4	C5	C6
001	7.940	111	13	2	67	15	17	13	13	11	9
002	8.292	107	12	2	43	12	12	7	7	6	6
003	4.643	100	13	2	52	11	10	5	8	9	7
004	7.470	107	12	2	66	14	15	11	11	9	9
005	8.882	114	12	1	58	14	15	10	12	11	6
006	7.585	115	12	2	51	14	11	7	8	6	9
007	7.650	111	13	2	71	15	17	12	14	11	10
008	2.412	97	13	2	51	10	12	5	11	5	6
009	6.000	100	13	1	49	12	9	6	9	6	7
010	8.833	112	13	2	51	15	16	4	9	5	8

is this? Give the observation number, and explain how you found it using regression output. Find this observation in the data list. What is unusual about it?

(b) Software packages often identify unusual or influential observations. Have any observations been identified as unusual or influential? If so, identify these points on a scatterplot of GPA versus IQ.

(c) C2 (school status) is the aspect of self-concept most highly correlated to GPA. If we carried out the simple linear regression of GPA on C2, what percent of the variation in students' GPAs would be explained by the straight-line relationship between GPA and C2?

(d) You know that IQ is associated with GPA, and you are not studying that relationship. Because C2 and IQ are positively correlated ($r = 0.547$), a significant relationship between C2 and GPA might occur just because C2 can "stand in" for IQ. Does C2 still contribute significantly to explaining GPA after we have allowed for the relationship between GPA and IQ? (Give a test statistic, its P-value, and your conclusion.)

(e) A new student in this class has IQ 115 and C2 score 14. What do you predict this student's GPA to be? (Just give a point prediction, not an interval.)

29.51 Children's perception of reading difficulty. Table 29.15 contains measured and self-estimated reading ability data for 60 fifth-grade students randomly sampled from one elementary school.[17] The variables are listed in the accompanying table. READING

(a) Is the relationship between measured (READ) and self-estimated (EST) reading ability the same for both boys and girls? Create an indicator variable for gender and fit an appropriate multiple regression model to answer the question.

Variable	Description
OBS	Observation number for each individual
SEX	Sex of the individual
LSS	Median grade level of student's selection of "best for me to read" (8 repetitions, each with four choices at grades 3, 5, 7, and 9 level)
IQ	IQ score
READ	Score on reading subtest of the Metropolitan Achievement Test
EST	Student's own estimate of his or her reading ability, scale 1 to 5 (1 = low)

(b) Fit a multiple regression model for predicting IQ from the explanatory variables LSS, READ, and EST. Are you happy with the fit of this model? Explain.

(c) Use residual plots to check the appropriate conditions for your model.

(d) Only two of the three explanatory variables in your model in part (b) have parameters that are significantly different from zero according to the individual t tests. Drop the explanatory variable that is not significant, and add the interaction term for the two remaining explanatory variables. Are you surprised by the results from fitting this new model? Explain what happened to the individual t tests for the two explanatory variables.

29.52 Florida real estate. The table on text page 628 gives the appraised market values and actual selling prices (in thousands of dollars) of condominium units sold in a beachfront building over a 164-month period. CONDOS

TABLE 29.15 Measured and self-estimated reading ability data for 60 fifth-grade students randomly sampled from one elementary school

OBS	SEX	LSS	IQ	READ	EST	OBS	SEX	LSS	IQ	READ	EST
1	F	5.00	145	98	4	31	M	7.00	106	55	4
2	F	8.00	139	98	5	32	M	6.00	124	70	4
3	M	6.00	126	90	5	33	M	8.00	115	82	5
4	F	5.33	122	98	5	34	M	8.40	133	94	5
5	F	5.60	125	55	4	35	F	5.00	116	75	4
6	M	9.00	130	95	3	36	F	6.66	102	80	3
7	M	5.00	96	50	4	37	F	5.00	127	85	4
8	M	4.66	110	50	4	38	M	6.50	117	88	5
9	F	4.66	118	75	4	39	F	5.00	109	70	3
10	F	8.20	118	75	5	40	M	5.50	137	80	4
11	M	4.66	101	65	4	41	M	6.66	117	55	4
12	M	7.50	142	68	5	42	M	6.00	90	65	2
13	F	5.00	134	80	4	43	F	4.00	103	30	1
14	M	7.00	124	10	4	44	F	5.50	114	74	5
15	M	6.00	112	67	4	45	M	5.00	139	80	5
16	M	6.00	109	83	3	46	M	6.66	101	70	2
17	F	5.33	134	90	4	47	F	8.33	122	60	4
18	M	6.00	113	90	5	48	F	6.50	105	45	2
19	M	6.00	81	55	3	49	F	4.00	97	45	1
20	F	6.00	113	83	4	50	M	5.50	89	55	4
21	M	6.00	123	65	4	51	M	5.00	102	30	2
22	F	4.66	94	25	3	52	F	4.00	108	10	4
23	M	4.50	100	45	3	53	M	4.66	110	40	1
24	F	6.00	136	97	4	54	M	5.33	128	65	1
25	M	5.33	109	75	4	55	M	5.20	114	15	2
26	F	3.60	131	70	4	56	M	4.00	112	62	2
27	M	4.00	117	23	3	57	F	3.60	114	98	4
28	M	6.40	110	45	3	58	M	6.00	102	52	2
29	F	6.00	127	70	2	59	F	4.60	82	23	1
30	F	6.00	124	85	5	60	M	5.33	101	35	2

(a) Find the multiple regression model for predicting selling price from appraised market value and month.

(b) Find and interpret the squared multiple correlation coefficient for your model.

(c) What is the regression standard error for this model?

(d) Hamada owns a unit in this building appraised at $802,600. Use your model to predict the selling price for Hamada's unit at month 164.

(e) Plot the residuals for your model against both explanatory variables and comment on the appearance of these plots.

29.53 Diamonds. Consider the diamond data of which Table 29.4 (page 29-30) is an excerpt. We are interested in predicting the total price of a diamond. Fit a simple linear regression model using *Carat* as the explanatory variable. DIAMND

TABLE 29.16 Catalog-spending data for 9 individuals from a very large database

Spending Ratio	Age	Length of Residence	Income	Total Assets	Security Assets	Short-Term Liquidity	Long-Term Liquidity	Wealth Index	Spending Volume	Spending Velocity
11.83	0	2	3	122	27	225	422	286	503	285
16.83	35	3	5	195	36	220	420	430	690	570
11.38	46	9	5	123	24	200	420	290	600	280
31.33	41	2	2	117	25	222	419	279	543	308
1.90	46	7	9	493	105	310	500	520	680	100
84.13	46	15	5	138	27	340	450	440	440	50
2.15	46	16	4	162	25	230	430	360	690	180
38.00	56	31	6	117	27	300	440	400	500	10
136.28	48	8	5	119	23	250	430	360	610	0

Collectible Gifts	Brick/ Mortar	Martha's Home	Sunday Ads	Theme Collections	Custom Decorating	Retail Kids	Teen Wear	Car Lovers	Country Collections
1	0	0	1	0	1	1	1	0	1
0	1	1	0	0	1	1	0	0	0
1	0	0	1	1	1	1	0	0	1
1	0	0	1	1	0	0	0	0	1
0	1	1	0	0	1	0	0	0	0
0	1	1	0	0	1	0	0	1	0
1	0	0	1	0	0	0	0	0	1
1	1	1	0	1	1	1	1	1	0
1	0	1	0	1	1	0	0	0	1

(a) Identify the least-squares line for predicting *Total Price* from *Carat*.

(b) Does the model provide a good fit? Comment on the residual plots. How much variation in price can be explained with this regression line?

(c) Create a new variable *Caratsq* = *Carat* × *Carat*. Fit a quadratic model using *Carat* and *Caratsq* and verify that your estimates for each parameter match those provided in Example 29.15 (page 29-29).

(d) Does the quadratic term *Caratsq* improve the fit of the model? Comment on the residual plots and the value of R^2.

(e) The individual t statistics look at the contribution of each variable when the other variables are in the model. State and test the hypotheses of interest for the quadratic term in your model.

29.54 Diamonds. Use the data in Table 29.4 (page 29-30) to fit the multiple regression model with two explanatory variables, *Carat* and *Depth*, to predict the *TotalPrice* of diamonds. Don't forget to include the interaction term in your model. DIAMND

(a) Identify the estimated multiple regression equation.

(b) Conduct the overall F test for the model.

(c) Identify the estimated regression parameters, standard errors, and t statistics with P-values.

(d) Prepare residuals plots and comment on whether the conditions for inference are satisfied.

(e) What percent of variation in *Total Price* is explained by this model?

(f) Find an estimate for σ and interpret this value.

29.55 Catalog spending. This realistic modeling project requires much more time than a typical exercise. Table 29.16 shows catalog-spending data for the first 9 of 200 randomly selected individuals from a very large (more than 20,000 households) database.[18] We are interested in developing a model to predict spending ratio. There are no missing values in the data set, but there are some incorrect entries that must be identified and removed before completing the analysis. Income is coded as an ordinal value, ranging from 1 to 12. Age can be regarded as quantitative, and any value less than 18 is invalid. Length of residence (LOR) is a value ranging from zero to someone's age. LOR should not be higher than age. All of the catalog variables are represented by indicator variables; either the consumer bought and the variable is coded as 1 or the consumer didn't buy and the variable is coded as 0. The other variables can be viewed as indexes for measuring assets, liquidity, and spending. Find a multiple regression model for predicting the amount of money that consumers will spend on catalog shopping, as measured by spending ratio. Your goal is to identify the best model you can. Remember to check the conditions for inference as you evaluate your models. CATALOG

EXPLORING THE WEB

29.56 Are Gas Prices Driving Elections? The *Chance* website discusses the use of regression to predict the margin of victory in presidential elections since 1948 from the price of gas (in 2008 dollars). Read the article at `www.causeweb.org/wiki/chance/index.php/Chance_News_72`. Use the data in the article to do the following.

(a) Fit a simple linear regression model using gas price to predict margin of victory. Do your results agree with those reported in the article?

(b) Use the incumbent party as an indicator variable (code Democrats as 1 and Republicans as 0), and add this to your simple linear regression model. What is the value of R^2?

(c) Now add gross domestic product (GDP) to your regression model in part (b). What is the value of R^2?

29.57 Historical Tuition and Fees. You can find data on past tuition and fees at several colleges by doing a Google search on "historical tuition and fees." Select one of the colleges you find and determine whether the data show the same pattern as you observed in Exercise 29.25 (page 29-45). You should try to find data going back at least 20 years (at the time we searched, we were able to find data for Clemson University, University of Pennsylvania, University of California, Univeristy of Colorado, Oregon State University, and William & Mary). The data may not be in spreadsheet format, and you may have to enter or cut and paste it into a spreadsheet to carry out your analysis.

Notes and Data Sources

1. Data on gas mileage are from the U.S. Department of Energy website at `http://www.fueleconomy.gov/feg/download.shtml`. The data given are a random sample of size 48 from all 1209 2016 model cars and trucks listed at the website.
2. Data were estimated from a scatterplot in Philipp Heeb, Mathias Kolliker, and Heinz Richner, "Bird-ectoparasite interactions, nest humidity, and ectoparasite community structure," *Ecology*, 81 (2000), pp. 958–968.
3. For more details, see H. Hoppeler and E. Weibel, "Scaling functions to body size: Theories and facts," *Journal of Experimental Biology*, 208 (2005), pp. 1573–1574.
4. We thank Professor Haruhiko Itagaki and his students, Andrew Vreede and Marissa Stearns, for providing data on tobacco hornworm caterpillars (*Manduca sexta*).
5. For more details, see Michael H. Kutner, Christopher J. Nachtsheim, and John Neter, *Applied Linear Regression Models*, 4th ed., McGraw-Hill, 2004.
6. Diamond database downloaded from AwesomeGems.com on July 28, 2005.
7. We thank Terry Klopcik for providing data from a physics lab on radioactive decay.
8. We thank David Cameron for providing data from a clothing retailer.
9. Found online at `https://avillage.web.virginia.edu/iaas/instreports/studat/dd/fees.htm`.
10. The data in Table 29.8 are part of a larger data set in the *Journal of Statistics Education* archive, accessible via the Internet. The original source is Pekka Brofeldt, "Bidrag till kaennedom on fiskbestondet i vaara sjoear. Laengelmaevesi," in T. H. Jaervi, Finlands fiskeriet, vol. 4, *Meddelanden utgivna av fiskerifoereningen i Finland*, Helsinki, 1917. The data were put in the archive (with information in English) by Juha Puranen of the University of Helsinki.
11. P. Velleman, *ActivStats* 2.0, Addison Wesley Interactive, 1997.
12. These data were provided by Professor Shelly MacDermid, Department of Child Development and Family Studies, Purdue University, from a study reported in S. M. MacDermid et al., "Is small beautiful? Work-family tension, work conditions, and organizational size," *Family Relations*, 44 (1994), pp. 159–167.
13. Data from the U.S. Historical Climatology Network, `http://cdiac.ornl.gov/epubs/ndp/ushcn/ushcn_map_interface.html`.
14. I thank Ray and Pat Heithaus for providing data on the pine seedlings at the Brown Family Environmental Center.
15. Data provided by Robert Dale, Purdue University.
16. Darlene Gordon, "The relationships among academic self-concept, academic achievement, and persistence with academic self-attribution, study habits, and perceived school environment," PhD thesis, Purdue University, 1997.
17. James T. Fleming, "The measurement of children's perception of difficulty in reading materials," *Research in the Teaching of English*, 1 (1967), pp. 136–156.
18. I thank David Cameron for providing the random sample of 200 observations from a large catalog-spending database.