Feverpitch/Deposit Photos

CHAPTER 12

Inference for Regression

Introduction

One of the most common uses of statistical methods in business and economics is to predict, or forecast, a response based on one or several explanatory (predictor) variables. In predictive analytics, these forecasts are then used by companies to make decisions. Here are some examples:

- Lime uses the day of the week, hour of the day, and current weather forecast to predict scooter- and bike-sharing demand around a city. This information is incorporated into the company's nightly redistribution strategy.
- Amazon wants to describe the relationship between dollars spent in its Digital Music department and dollars spent in its Online Grocery department by 18- to 25-year-olds this past year. This information will be used to determine a new advertising strategy.
- Panera Bread, when looking for a new store location, develops a model to predict profitability using the amount of traffic near the store, the proximity to competitive restaurants, and the average income level of the neighborhood.

Prediction is most straightforward when there is a straight-line relationship between a quantitative response variable y and a single quantitative explanatory variable x. This is **simple linear regression,** the topic of this chapter. In Chapter 13, we will consider the more common setting involving more than one explanatory (predictor) variable. Because both settings share many of the same ideas, we introduce inference for regression under the simple setting.

simple linear regression

CHAPTER OUTLINE

least-squares line, p. 83

In Chapter 2, we saw that the least-squares line can be used to predict y for a given value of x. Now we consider the use of significance tests and confidence intervals in this setting. To do this, we will think of the least-squares line, $b_0 + b_1x$, as an estimate of a regression line for the population—just as in Chapter 8, where we viewed the sample mean $\bar{x}$ as the estimate of the population mean μ, and in Chapter 10, where we viewed the sample proportion $\hat{p}$ as the estimate for the population proportion p.

parameters and statistics, p. 295

We write the population regression line as $\beta_0 + \beta_1x$. The numbers β_0 and β_1 are *parameters* that describe this population line. The numbers b_0 and b_1 are *statistics* calculated by fitting a line to a sample. The fitted intercept b_0 estimates the intercept of the population line β_0, and the fitted slope b_1 estimates the slope of the population line β_1.

ANOVA, p. 458

Our discussion begins with an overview of the simple linear regression model and inference about the slope β_1 and the intercept β_0. Because regression lines are most often used for prediction, we then consider inference about either the mean response or an individual future observation on y for a given value of the explanatory variable x. We conclude the chapter with more of the computational details, including the use of analysis of variance (ANOVA). If you plan to read Chapter 13 on regression involving more than one explanatory variable, these details will be very useful.

12.1 Inference about the Regression Model

When you complete this section, you will be able to:

- Describe the simple linear regression model in terms of a population regression line and the distribution of deviations of the response variable y from this line.
- Use linear regression output from statistical software to find the least-squares regression line and estimated regression standard deviation.
- Use plots of the residuals to visually check the assumptions of the simple linear regression model.
- Construct and interpret a confidence interval for the population intercept and for the population slope.
- Perform a significance test for the population intercept and for the population slope and summarize the results.

the one-way ANOVA model, p. 465

Simple linear regression studies the relationship between a quantitative response variable y and a quantitative explanatory variable x. We expect that different values of x will be associated with different mean responses for y. We encountered a situation similar to this in Chapter 9, when we considered the possibility that different treatment groups had different mean responses.

Figure 12.1 illustrates the statistical model from Chapter 9 for comparing the items per hour entered by three groups of financial clerks using new

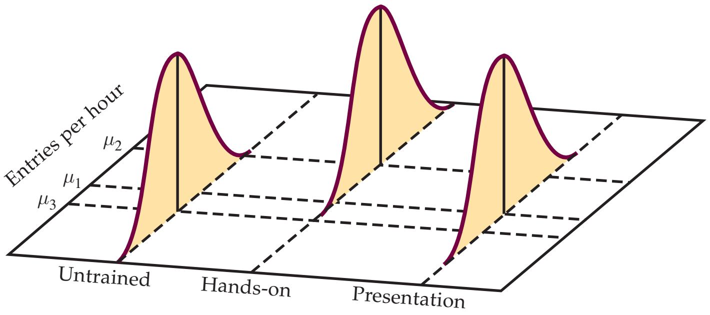


FIGURE 12.1 The statistical model for comparing the responses to three treatments. The responses vary within each treatment group according to a Normal distribution. The mean may be different in the three treatment groups.

data entry software. Group 1 received no training, Group 2 received one hour of hands-on training, and Group 3 attended an hour-long presentation describing the entry process. Entries per hour is the response variable y. Treatment (or type of training) is the explanatory variable. The model has two important parts:

- The mean entries per hour may be different in the three populations. These means are μ_1, μ_2 and μ_3 in Figure 12.1.

- Individual entries per hour vary within each population according to a Normal distribution. The three Normal curves in Figure 12.1 describe these responses. These Normal distributions have the same spread, indicating that the population standard deviations are assumed to be equal.

Statistical model for simple linear regression

In linear regression, the explanatory variable x is quantitative and can have many different values. Imagine, for example, giving different lengths x of hands-on training to different groups of clerks. We can think of these groups as belonging to **subpopulations,** one for each possible value of x. Each subpopulation consists of all individuals in the population having the same value of x. If we gave $x = 1$ hour of training to some subjects, $x = 2$ hours of training to some others, and $x = 4$ hours of training to some others, these three groups of subjects would be considered samples from the corresponding three subpopulations.

subpopulation

The statistical model for simple linear regression assumes that, for each value of x (or subpopulation), the response variable y is Normally distributed with a mean that depends on x. We use μ_y to represent these means. In general, the means μ_y can change as x changes according to any sort of pattern. In simple linear regression, we assume that the means all lie on a line when plotted against x.

To summarize, this model has two important parts:

- The mean entries per hour μ_y changes as the number of training hours x changes and these means all lie on a straight line; that is, $\mu_y = \beta_0 + \beta_1 x$.

- Individual entries per hour y for subjects with the same amount of training x vary according to a Normal distribution. This variation, measured by the standard deviation σ, is the same for all values of x.

Figure 12.2 illustrates this statistical model. The line describes how the mean response μ_y changes with x; it is called the **population regression line.** The three Normal curves show how the response y will vary for three different values of the explanatory variable x. Each curve is centered at its mean response μ_y. All three curves have the same spread, measured by their common standard deviation σ.

population regression line

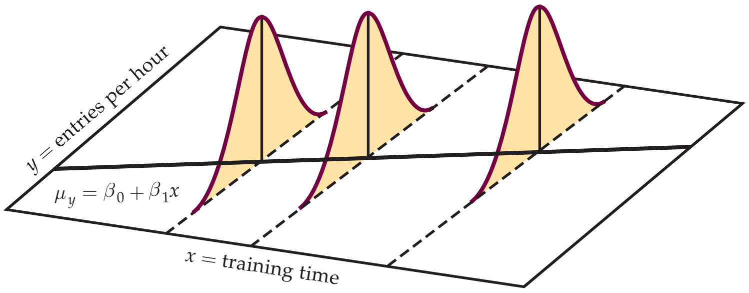


FIGURE 12.2 The statistical model for linear regression. The responses vary within each subpopulation according to a Normal distribution. The mean response is a straight-line function of the explanatory variable.

From data analysis to inference

The data for a simple linear regression problem are the n pairs of (x,y) observations. The model takes each x to be a fixed known quantity, like the hours of training that a clerk receives.[1] The response y for a given x is a Normal random variable. Our regression model describes the mean and standard deviation of this random variable.

We will use Case 12.1 to explain the fundamentals of simple linear regression. In practice, regression calculations are always done by software, so we rely on computer output for the arithmetic. Later in the chapter, we show formulas for doing the calculations. These formulas are useful in understanding analysis of variance (see Section 12.3) and multiple regression (see Chapter 13).

J. Countess/Getty Images

ENTRE

smoothed curve, p. 69

CASE 12.1 **The Relationship between Income and Education for Entrepreneurs** Numerous studies have shown that better-educated employees have higher incomes. Is this also true for entrepreneurs? Does more years of formal education translate into higher income? We know about the extremely successful entrepreneurs, such as Oprah Winfrey and her amazing rags-to-riches story. Cases like this, however, are anecdotal and most likely not representative of the population of entrepreneurs. One study explored this question using the National Longitudinal Survey of Youth (NLSY), which followed a large group of individuals aged 14 to 22 for roughly 10 years.[2] The researchers studied both employees and entrepreneurs, but we just focus on entrepreneurs here.

The researchers defined *entrepreneurs* as those individuals who were self-employed or who were the owner/director of an incorporated business. For each of these individuals, they recorded the education level and income. The education level (Educ) was defined as the years of completed schooling prior to starting the business. The income level (Inc) was the average annual total earnings since starting the business.

We consider a random sample of 100 entrepreneurs. Figure 12.3 is a scatterplot of the data with a fitted smoothed curve to help us visualize the relationship. The explanatory variable x is the entrepreneur's education level. The response variable y is the income level. ■

Let's briefly review some of the ideas from Chapter 2 regarding least-squares regression. We always start with a plot of the data, as in Figure 12.3,

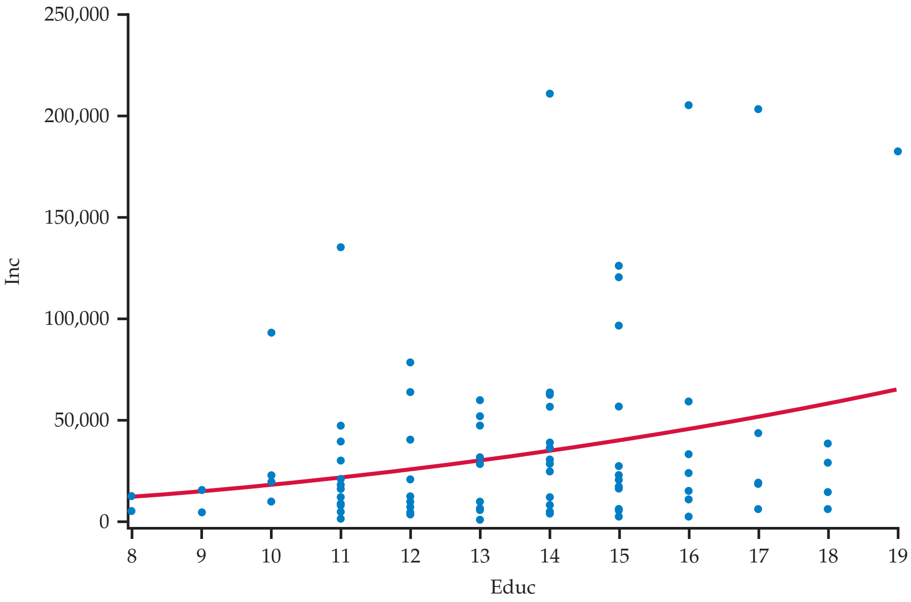


FIGURE 12.3 Scatterplot, with smoothed curve, of average annual income versus years of education for a sample of 100 entrepreneurs.

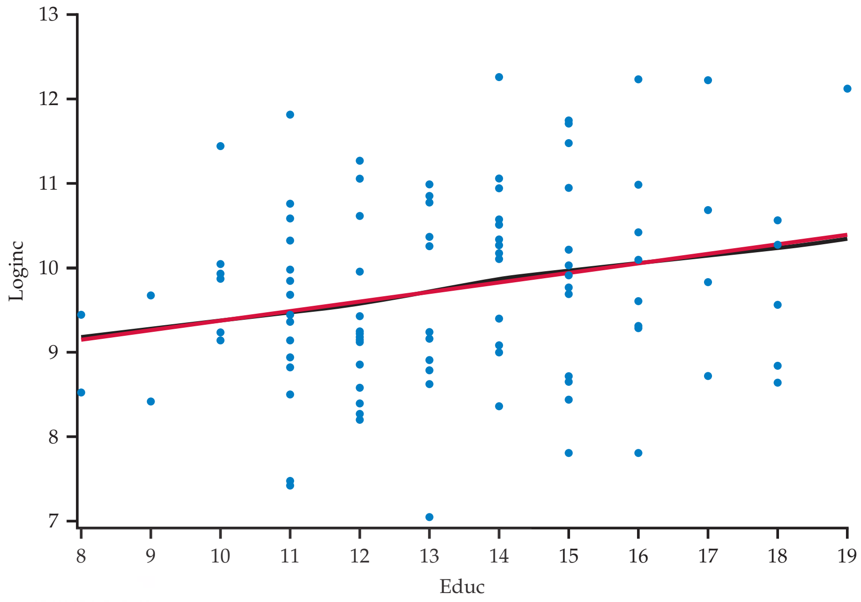


FIGURE 12.4 Scatterplot, with smoothed curve (black) and regression line (red), of log average annual income versus years of education for a sample of 100 entrepreneurs. The smoothed curve is almost the same as the least-squares regression line.

to verify that the relationship is approximately linear with no outliers. *There is no point in fitting a linear model if the relationship does not, at least approximately, appear linear.* For the data of Case 12.1, the smoothed curve looks roughly linear but the distributions of incomes about it are skewed to the right. At each education level, there are many small incomes and just a few very large incomes. It also looks like the smoothed curve is being pulled toward those very large incomes, suggesting those observations could be influential.

influential observations, p. 95

A common remedy for a skewed variable such as income is to consider transforming it prior to fitting a model. Here, the researchers considered the natural logarithm of income (Loginc). Figure 12.4 is a scatterplot of Loginc versus Educ with a fitted curve and the least-squares regression line. The smoothed curve nearly overlaps the fitted line, suggesting a very linear association. In addition, the observations in the y direction are more equally dispersed above and below this fitted line than with the curve in Figure 12.3. Lastly, those four very large incomes no longer appear to be influential. Given these results, we continue our discussion of least-squares regression using the transformed y data.

log transformation, p. 70

EXAMPLE 12.1

CASE 12.1

ENTRE

Prediction of Loginc from Educ The fitted line in Figure 12.4 is the least-squares regression line for predicting y (log income) from x (years of formal schooling). The equation of this line is

$$\hat{y} = 8.2546 + 0.1126x$$

or

$$\text{predicted Loginc} = 8.2546 + 0.1126 \times \text{Educ}$$

We can use the least-squares regression equation to find the predicted log income corresponding to a given education level. The difference between the observed value and the predicted value is the residual. For example, Entrepreneur 4 has 15 years of formal schooling and a log income of $y = 10.2274$. The predicted log income of this person is

residuals, p. 90

$$\hat{y} = 8.2546 + (0.1126)(15) = 9.9436$$

so the residual is

$$y - \hat{y} = 10.2274 - 9.9436 = 0.2838 \blacksquare$$

Recall that the least-squares line is the line that minimizes the sum of the squares of the residuals. The least-squares regression line also always passes through the point $(\bar{x}, \bar{y})$. These are helpful facts to remember when considering the fit of this line to a data set. You can also use the *Correlation and Regression* applet, introduced in Chapter 2, to visually explore residuals and the properties of the least-squares line.

interpretation of r^2, p. 88

In Section 2.2 (page 74), we discussed the correlation as a measure of linear association between two quantitative variables. In Section 2.3, we learned to interpret the square of the correlation as the fraction of the variation in y that is explained by x in a simple linear regression.

EXAMPLE 12.2

CASE 12.1 **Correlation between Loginc and Educ** For Case 12.1, the correlation between log income and education level is $r = 0.2394$. Because the squared correlation $r^2 = 0.0573$, indicating that the change in Loginc along the regression line as Educ increases explains only 5.7% of the variation. The remaining 94.3% is due to other differences among these entrepreneurs. The entrepreneurs in this sample live in different parts of the United States; some are single and others are married, and some may have had a difficult upbringing. All of these factors could be associated with income and, therefore, add to the variability if they are not included in the model. ■

APPLY YOUR KNOWLEDGE

CASE 12.1 **12.1 Predict Loginc.** In Case 12.1, Entrepreneur 12 has Educ = 13 years and a log income of $y = 10.7649$. Using the least-squares regression equation in Example 12.1, find the predicted Loginc and the residual for this individual.

12.2 Draw the fitted line. Suppose you fit 10 pairs of (x, y) data using least squares. Draw the fitted line if $\bar{x} = 5$, $\bar{y} = 4$, and the residual for the pair $(3, 4)$ is 1.

Having reviewed the basics of least-squares regression, we are now ready to discuss inference for regression. To do this:

- We regard the 100 entrepreneurs for whom we have data as a simple random sample from the population of all entrepreneurs in the United States.

- We use the regression line calculated from this sample as a basis for inference about the population. For example, for a given level of education, we want not just a prediction, but a prediction with a margin of error and a level of confidence for the log income of any entrepreneur in the United States.

Our statistical model assumes that the responses y are Normally distributed with a mean μ_y that depends upon x in a linear way. Specifically, the population regression line

$$\mu_y = \beta_0 + \beta_1 x$$

describes the relationship between the mean log income μ_y and the number of years of formal education x in the population. The slope β_1 is the average change in log income for each additional year of education. It turns out that a change in natural logs is a good approximation for the percent change [see Example 14.11 (page 698) for more details]. Thus, another way to view β_1 in this setting is as the average percent change in income for an additional year of education. The intercept β_0 is the mean log income when an entrepreneur has $x = 0$ years of formal education. This parameter, by itself, is not interesting in this example because zero years of education is very unusual. The value $x = 0$ is also well outside the data's range.

extrapolation, p. 100

Because the means μ_y lie on the line $\mu_y = \beta_0 + \beta_1 x$, they are all determined by β_0 and β_1. Thus, once we have estimates of β_0 and β_1, the linear relationship determines the estimates of μ_y for all values of x. Linear regression allows us to do inference not only for those subpopulations for which we have data, but also for those subpopulations corresponding to x's not present in the data. These x-values can be both within and outside the range of observed x's. *Use extreme caution when predicting outside the range of the observed x's, because there is no assurance that the same linear relationship between μ_y and x holds.*

We cannot observe the population regression line because the observed responses y vary about their means. In Figure 12.4, we see the least-squares regression line that describes the overall pattern of the data, along with the scatter of individual points about this line. The statistical model for linear regression makes the same distinction, as shown in Figure 12.2 with the line and three Normal curves. The population regression line describes the on-the-average relationship, whereas the Normal curves describe the variability in y for each value of x.

As we did in Chapter 9, we can think of this regression model as being of the form

$$\text{DATA} = \text{FIT} + \text{RESIDUAL}$$

DATA = FIT + RESIDUAL, p. 464

The FIT part of the model consists of the subpopulation means, given by the expression $\beta_0 + \beta_1 x$. The RESIDUAL part represents deviations of the data from the line of population means.

The model assumes that these deviations are Normally distributed with standard deviation σ. We use ε (the lowercase Greek letter epsilon) to stand for the RESIDUAL part of the statistical model. A response y is the sum of its mean and a chance deviation ε from the mean. The deviations ε represent "noise"—that is variations in y due to other causes that prevent the observed (x,y)-values from forming a perfectly straight line.

SIMPLE LINEAR REGRESSION MODEL

Given n observations of the explanatory variable x and the response variable y,

$$(x_1, y_1),\ (x_2, y_2),\ \ldots,\ (x_n, y_n)$$

The **statistical model for simple linear regression** states that the observed response y_i when the explanatory variable takes the value x_i is

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

Here, $\mu_y = \beta_0 + \beta_1 x_i$ is the mean response when $x = x_i$. The deviations ε_i are independent and Normally distributed with mean 0 and standard deviation σ.

The parameters of the model are β_0, β_1, and σ.

Use of a simple linear regression model can be justified in a wide variety of circumstances. Sometimes, we observe the values of two variables, and we formulate a model with one of these as the response variable and the other as the explanatory variable. This is the setting for Case 12.1, where the response variable is log income (Loginc) and the explanatory variable is the number of years of formal education (Educ). In other settings, the values of the explanatory variable are chosen by the persons designing the study. The scenario illustrated by Figure 12.2 is an example. Here, the explanatory variable is training time, which is set at a few carefully selected values. The response variable is the number of entries per hour.

APPLY YOUR KNOWLEDGE

12.3 Understanding a linear regression model. Consider a linear regression model for the number of financial entries per hour with $\mu_y = 56.82 + 2.4x$ and standard deviation $\sigma = 4.4$. The explanatory variable x is the number of hours of hands-on training.

(a) What is the slope of the population regression line?

(b) Explain clearly what this slope says about the change in the mean of y for an additional hour of training.

(c) What is the intercept of the population regression line?

(d) Explain clearly what this intercept says about the mean number of entries per hour.

12.4 Understanding a linear regression model, continued. Refer to the previous exercise.

(a) What is the subpopulation mean when $x = 3$ hours?

(b) What is the subpopulation distribution when $x = 3$ hours?

(c) Between what two values would approximately 95% of the observed responses y fall when $x = 3$ hours?

For the simple linear regression model to be valid, one essential assumption is that the relationship between the means of the response variable for the different values of the explanatory variable is approximately linear. This is the FIT part of the model. Another essential assumption concerns the RESIDUAL part of the model. The assumption states that the deviations are an SRS from a Normal distribution with mean zero and standard deviation σ. If the data are collected through some sort of random sampling, the SRS assumption is often easy to justify. This is the case in our two scenarios, in which both variables are observed in a random sample from a population or the response variable is measured at several predetermined values of the explanatory variable that were randomly assigned to clerks.

In many other settings, particularly in business applications, we analyze all of the data available and there is no random sampling. Here, we often justify the use of inference for simple linear regression by viewing the data as coming from some sort of process. Here is one example.

EXAMPLE 12.3

Profits and Foot Traffic Panera Bread wants to select the location for a new store. To help with this decision, company managers use information from all the current stores to determine the relationship between profits and foot traffic outside the establishment. The regression model they use says that

$$\text{Profits} = \beta_0 + \beta_1 \times \text{Foot Traffic} + \varepsilon$$

The slope β_1 is, as usual, a rate of change: it is the expected increase in annual profits associated with each additional person walking by the store. The intercept β_0 is needed to describe the line but has no interpretive importance because no stores have zero foot traffic. Nevertheless, foot traffic does not completely determine profit. The ε term in the model accounts for differences among individual stores with the same foot traffic. A store's proximity to other restaurants, for example, could be important but is not included in the FIT part of the model. In Chapter 13, we consider moving variables like this out of the RESIDUAL part of the model by allowing for more than one explanatory variable in the FIT part. ■

APPLY YOUR KNOWLEDGE

12.5 U.S. versus overseas stock returns. Returns on common stocks in the United States and overseas appear to be growing more closely correlated as various countries' economies become more interdependent. Suppose that the following population regression line connects the total annual returns (in percent) on two indexes of stock prices:

$$\text{Mean overseas return} = -0.3 + 0.12 \times \text{U.S. Return}$$

(a) What is β_0 in this line? What does this number say about overseas returns when the U.S. market is flat (0% return)?

(b) What is β_1 in this line? What does this number say about the relationship between U.S. and overseas returns?

(c) We know that overseas returns will vary in years that have the same return on U.S. common stocks. Write the regression model based on the population regression line given in the problem statement. What part of this model allows overseas returns to vary when U.S. returns remain the same?

12.6 Fixed and variable costs. In some mass-production settings, there is a linear relationship between the number x of units of a product in a production run and the total cost y of making these x units.

(a) Write a population regression model to describe this relationship.

(b) The fixed cost is the component of total cost that does not change as x increases. Which parameter in your model is the fixed cost?

(c) Which parameter in your model shows how total cost changes as more units are produced? Do you expect this number to be greater than 0 or less than 0? Explain your answer.

(d) Actual data from several production runs will not fall directly on a straight line. What term in your model allows variation among runs of the same size x?

Estimating the regression parameters

The method of least squares presented in Chapter 2 fits the least-squares line to summarize the relationship between the observed values of an explanatory variable and a response variable. Now we want to use this line as a basis for inference about a population from which our observations are a sample. In this setting, the slope b_1 and intercept b_0 of the least-squares line

$$\hat{y} = b_0 + b_1 x$$

estimate the slope β_1 and the intercept β_0 of the population regression line, respectively.

This inference should be done only when the statistical model for regression is reasonable. Model checks are needed and some judgment is required. Because many of these checks rely on the residuals, let's briefly review the methods introduced in Chapter 2 for fitting the linear regression model to data and then discuss the model checks.

Using the formulas from Chapter 2, the slope of the least-squares line is

$$b_1 = r\frac{s_y}{s_x}$$

and the intercept is

$$b_0 = \bar{y} - b_1\bar{x}$$

correlation, p. 75

Here, r is the correlation between the observed values of y and x, s_y is the standard deviation of the sample of y's, and s_x is the standard deviation of the sample of x's. Notice that if the estimated slope is 0, so is the correlation, and vice versa. We discuss this connection in more depth later in this section.

The remaining parameter to be estimated is σ, which measures the variation of y about the population regression line. More precisely, σ is the standard deviation of the Normal distribution of the deviations ε_i in the regression model. We don't observe these ε_i, so how can we estimate σ?

residuals, p. 90

Recall that the vertical deviations of the points in a scatterplot from the fitted regression line are the residuals. We use e_i for the residual of the ith observation:

$$\begin{aligned} e_i &= \text{Observed Response} - \text{Predicted Response} \\ &= y_i - \hat{y}_i \\ &= y_i - b_0 - b_1x_i \end{aligned}$$

The residuals e_i are the observable quantities that correspond to the unobservable model deviations ε_i. The e_i sum to 0, and the ε_i come from a population with mean 0. Because we do not observe the ε_i, we use the residuals to estimate σ and check the model assumptions of the ε_i.

To estimate σ, we work first with the variance and take the square root to obtain the standard deviation. For simple linear regression, the estimate of σ^2 is the average squared residual

$$\begin{aligned} s^2 &= \frac{1}{n-2}\sum e_i^2 \\ &= \frac{1}{n-2}\sum (y_i - \hat{y}_i)^2 \end{aligned}$$

We average by dividing the sum by $n - 2$ so as to make s^2 an unbiased estimator of σ^2. We subtract 2 from n because we're using the data to also estimate β_0 and β_1. In addition, it turns out that when any $n - 2$ residuals are known, we can find the other two residuals.

regression standard deviation σ

The quantity $n - 2$ is the degrees of freedom of s^2. The estimate of the **regression standard deviation $\boldsymbol{\sigma}$** is given by

$$s = \sqrt{s^2}$$

We call s the *regression standard error.*

ESTIMATING THE REGRESSION PARAMETERS

In the simple linear regression setting, we use the **slope $\boldsymbol{b_1}$** and **intercept $\boldsymbol{b_0}$** of the least-squares regression line to estimate the slope β_1 and intercept β_0 of the population regression line, respectively.

The standard deviation σ in the model is estimated by the **regression standard error**

$$s = \sqrt{\frac{1}{n-2}\sum (y_i - \hat{y}_i)^2}$$

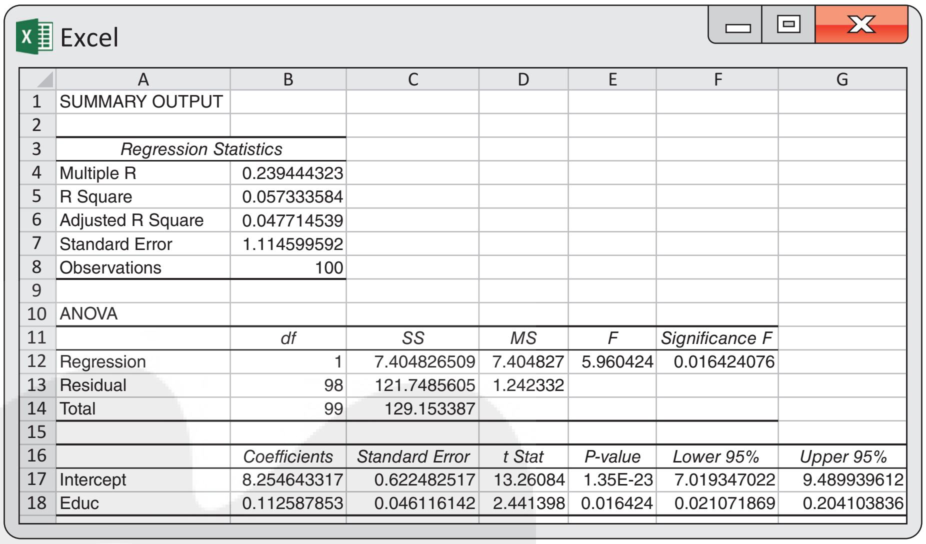
In practice, we use software to calculate b_1, b_0, and s from the (x,y) pairs of data. Here are the results for the income example of Case 12.1.

EXAMPLE 12.4

CASE 12.1

ENTRE

Reading Simple Regression Output Figure 12.5 displays Excel output for the regression of log income (Loginc) on years of education (Educ) for our sample of 100 entrepreneurs in the United States. In this output, we find the correlation $r = 0.2394$ and the squared correlation that we used in Example 12.2, along with the intercept and slope of the least-squares line. The regression standard error s is labeled simply "Standard Error."



Excel

	A	B	C	D	E	F	G
1	SUMMARY OUTPUT						
2							
3	*Regression Statistics*						
4	Multiple R	0.239444323					
5	R Square	0.057333584					
6	Adjusted R Square	0.047714539					
7	Standard Error	1.114599592					
8	Observations	100					
9							
10	ANOVA						
11		*df*	*SS*	*MS*	*F*	*Significance F*	
12	Regression	1	7.404826509	7.404827	5.960424	0.016424076	
13	Residual	98	121.7485605	1.242332			
14	Total	99	129.153387				
15							
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
17	Intercept	8.254643317	0.622482517	13.26084	1.35E-23	7.019347022	9.489939612
18	Educ	0.112587853	0.046116142	2.441398	0.016424	0.021071869	0.204103836

FIGURE 12.5 Excel output for the regression of log average income on years of education, for Example 12.4.

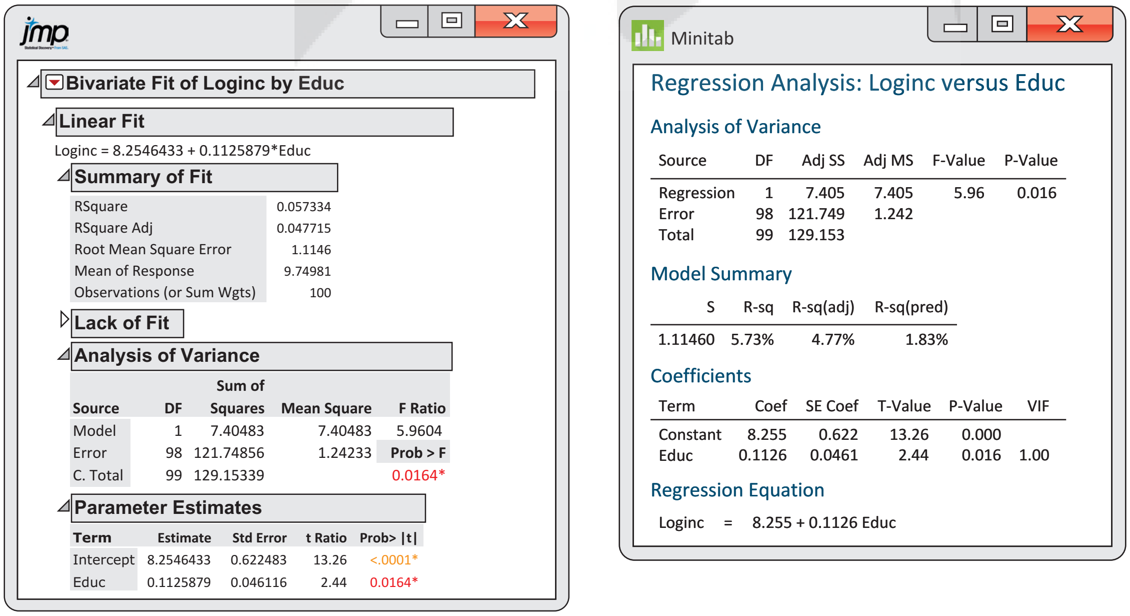
The three parameter estimates are

$$b_0 = 8.254643317 \qquad b_1 = 0.112587853 \qquad s = 1.114599592$$

After rounding, the fitted regression line is

$$\hat{y} = 8.2546 + 0.1126x$$

As usual, we ignore the parts of the output that we do not yet need. We will return to the output for additional information later.



jmp

Bivariate Fit of Loginc by Educ

Linear Fit

Loginc = 8.2546433 + 0.1125879*Educ

Summary of Fit

RSquare	0.057334
RSquare Adj	0.047715
Root Mean Square Error	1.1146
Mean of Response	9.74981
Observations (or Sum Wgts)	100

Lack of Fit

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	7.40483	7.40483	5.9604
Error	98	121.74856	1.24233	Prob > F
C. Total	99	129.15339		0.0164*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob> |t|
Intercept	8.2546433	0.622483	13.26	<.0001*
Educ	0.1125879	0.046116	2.44	0.0164*

Minitab

Regression Analysis: Loginc versus Educ

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	1	7.405	7.405	5.96	0.016
Error	98	121.749	1.242		
Total	99	129.153			

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
1.11460	5.73%	4.77%	1.83%

Coefficients

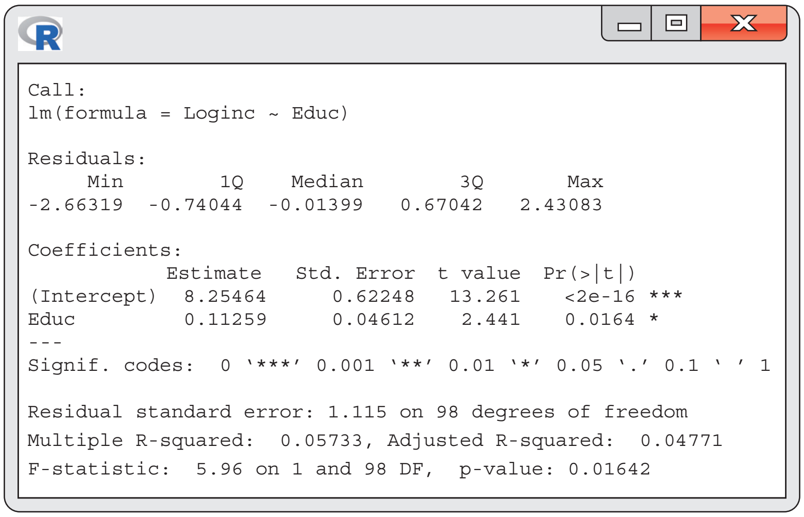
Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	8.255	0.622	13.26	0.000	
Educ	0.1126	0.0461	2.44	0.016	1.00

Regression Equation

Loginc = 8.255 + 0.1126 Educ

FIGURE 12.6 JMP, Minitab, and R outputs for the regression of log average income on years of education. The data are the same as in Figure 12.5.

FIGURE 12.6 Continued



```
Call:
lm(formula = Loginc ~ Educ)

Residuals:
     Min        1Q    Median        3Q       Max
-2.66319  -0.74044  -0.01399   0.67042   2.43083

Coefficients:
            Estimate   Std. Error  t value  Pr(>|t|)
(Intercept)  8.25464     0.62248   13.261    <2e-16 ***
Educ         0.11259     0.04612    2.441    0.0164 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.115 on 98 degrees of freedom
Multiple R-squared:  0.05733, Adjusted R-squared:  0.04771
F-statistic:  5.96 on 1 and 98 DF,  p-value: 0.01642
```

Figure 12.6 shows the regression output from three other software packages. Although the formats differ, you should be able to find the results you need. Once you know what to look for, you can understand statistical output from almost any software. ■

APPLY YOUR KNOWLEDGE

12.7 Research and development spending. The National Science Foundation collects data on research and development spending by universities and colleges in the United States.[3] Here are the data for the years 2012–2015: NSF

Year	2012	2013	2014	2015
Spending (billions of dollars)	65.9	67.1	67.3	68.7

(a) Create a scatterplot that shows the increase in research and development spending over time. Does the pattern suggest that the spending is increasing linearly over time? Explain your answer.

(b) Find the equation of the least-squares regression line for predicting spending from year. Add this line to your scatterplot.

(c) For each of the four years, find the residual. Use these residuals to calculate the regression standard error s. (Do these calculations with a calculator or spreadsheet.)

(d) Write the regression model for this setting. What are your estimates of the unknown parameters in this model?

(e) Use your least-squares equation to predict research and development spending for the year 2016. The actual spending for that year was $72.0 billion. Add this point to your plot and comment on how well the model predicted the actual outcome.

(*Comment:* These are *time series data*. Simple regression is often a good fit to time series data over a limited span of time. See Chapter 14 for methods designed specifically for use with time series.)

Conditions for regression inference

You can fit a least-squares line to any set of explanatory-response data when both variables are quantitative. The simple linear regression model, which is the basis for inference, imposes several conditions on this fit. *We should always verify these conditions before proceeding to inference.* There is no point in trying to do statistical inference if we cannot trust the results.

The conditions concern the population, but we can observe only our sample. Thus, in doing inference, we act as if **the sample is an SRS from the population.** For the study described in Case 12.1, the researchers used a national survey. Participants were chosen to be a representative sample of the United States, so we can treat this sample as an SRS. *The potential for bias should always be considered, especially when the sample includes volunteers.*

The next condition is that **there is a linear relationship in the population,** described by the population regression line. We can't observe the population line, so we check this condition by asking if the sample data show a roughly linear pattern in a scatterplot. We also check for any outliers or influential observations that could affect the least-squares fit.

outliers and influential observations, p. 95

The model also says that **the standard deviation of the responses about the population line is the same for all values of the explanatory variable.** In practice, this means the spread in the observations above and below the least-squares line should be roughly the same as x varies.

residual plots, p. 91

Plotting the residuals against the explanatory variable or against the predicted values is a helpful and frequently used visual aid to check both of these conditions. This technique is often better than creating a scatterplot because a residual plot magnifies any patterns that exist. The residual plot in Figure 12.7 for the data of Case 12.1 looks satisfactory. There is no obvious pattern in the residuals versus x, no data points seem out of the ordinary, and the residuals appear equally dispersed throughout the range of the explanatory variable.

The final condition is that **the response varies Normally about the population regression line.** If that is the case, we expect the residuals e_i to also be Normally distributed.[4] A Normal quantile plot or histogram of the residuals is commonly used to check this condition. For the data of Case 12.1, a Normal quantile plot of the residuals (Figure 12.8) shows no serious deviations

Normal quantile plot, p. 53

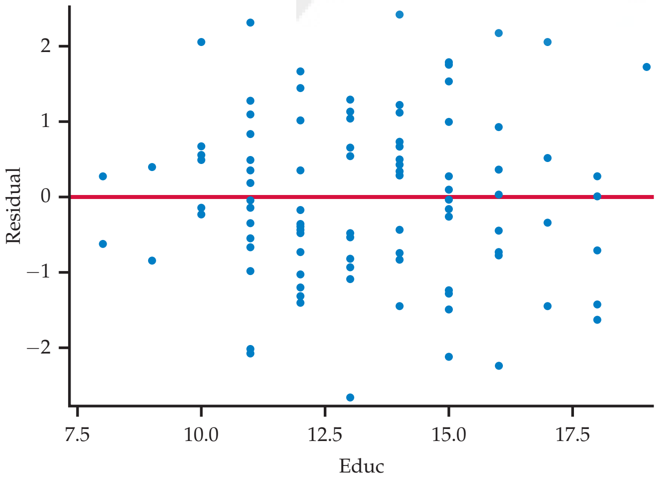


FIGURE 12.7 Plot of the regression residuals against the explanatory variable for the annual income data.

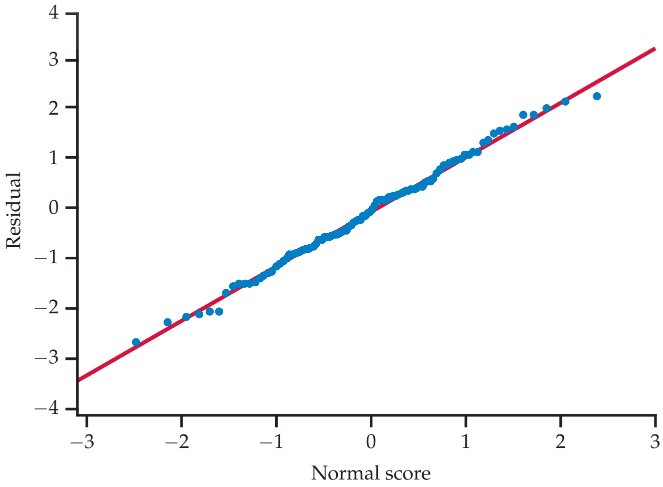


FIGURE 12.8 Normal quantile plot of the regression residuals for the average annual income data.

from a Normal distribution. The data give us no reason to doubt the simple linear regression model, so we proceed to inference.

Notice that Normality of the distributions of the response and explanatory variables is not required. The Normality condition applies to the distribution of the model deviations, which we assess using the residuals. For the entrepreneur problem, we transformed y to get a more linear relationship and residuals that are more Normal with constant variance. The fact that the distribution of the transformed y approaches Normality is purely a coincidence.

While not the case here, sometimes x is not a fixed known quantity but rather is measured with error. Even if all the conditions for linear regression are satisfied, *this regression model is not appropriate if the error in measuring x is large relative to the spread of the x's*. If this is a concern, seek expert advice, as more advanced inference methods are needed.

LINEAR REGRESSION MODEL CONDITIONS

To use the least-squares line as a basis for inference about a population, each of the following conditions should be approximately met:

- The sample is an SRS from the population.
- There a linear relationship between x and y.
- The standard deviation of the responses y about the population regression line is the same for all x.
- The model deviations are Normally distributed.

Confidence intervals and significance tests

Chapter 8 presented confidence intervals and significance tests for means and differences in means. In each case, inference rested on the standard errors of estimates and on t distributions. Inference for the slope and intercept in linear regression is similar in principle. For example, the t^*confidence intervals have the form

$$\text{estimate} \pm t^* \text{SE}_{\text{estimate}}$$

where t^* is a critical value of a t distribution. It is the formulas for the estimate and standard error that are different.

Confidence intervals and tests for the slope and intercept are based on the sampling distributions of the estimates b_1 and b_0. Here are some important facts about these sampling distributions when the simple linear regression model is true:

- Both b_1 and b_0 have Normal distributions.

unbiased estimator, p. 300

- The mean of b_1 is β_1 and the mean of b_0 is β_0. That is, the slope and intercept of the fitted line are unbiased estimators of the slope and intercept of the population regression line.
- The standard deviations of b_1 and b_0 are multiples of the regression standard deviation σ. (We give details later.)

central limit theorem, p. 313

Normality of b_1 and b_0 is a consequence of Normality of the individual deviations ε_i in the regression model. If the ε_i are not Normal, a general form of the central limit theorem tells us that the distributions of b_1 and b_0 will be approximately Normal when we have a large sample. On the one hand, this means **regression inference is robust against moderate lack of Normality.** *On the other hand, outliers and influential observations can invalidate the results of inference for regression.*

Because b_1 and b_0 have Normal sampling distributions, standardizing these estimates gives standard Normal z statistics. The standard deviations of these estimates are multiples of σ. Because we do not know σ, we estimate it by s, the regression standard error. When we do this, we get t distributions with degrees of freedom $n - 2$, the degrees of freedom of s. We give formulas for the standard errors SE_{b_1} and SE_{b_0} in Section 12.3. For now, we concentrate on the basic ideas and let software do the calculations.

INFERENCE FOR THE REGRESSION SLOPE

A **level *C* confidence interval** for the slope β_1 of the population regression line is

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

In this expression, t^* is the value for the $t(n - 2)$ density curve with area C between $-t^*$ and t^*. The **margin of error** is $m = t^* \text{SE}_{b_1}$.

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

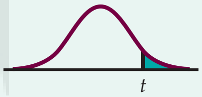
$$t = \frac{b_1 - \beta_1^*}{\text{SE}_{b_1}}$$

Most software provides the test of the hypothesis H_0: $\beta_1 = 0$. In that case, the t statistic reduces to

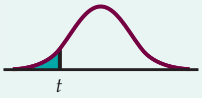
$$t = \frac{b_1}{\text{SE}_{b_1}}$$

The **degrees of freedom** are $n - 2$. In terms of a random variable T having the $t(n - 2)$ distribution, the P-value for a test of H_0 against

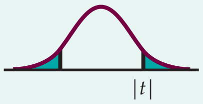
H_a: $\beta_1 > \beta_1^*$ is $P(T \geq t)$



H_a: $\beta_1 < \beta_1^*$ is $P(T \leq t)$



H_a: $\beta_1 \neq \beta_1^*$ is $2P(T \geq |t|)$



Formulas for confidence intervals and significance tests for the intercept β_0 are exactly the same, replacing b_1 and SE_{b_1} by b_0 and its standard error SE_{b_0}, respectively. *Although computer outputs may include a test of H_0: $\beta_0 = 0$, this information often has little practical value.* From the equation for the population regression line, $\mu_y = \beta_0 + \beta_1 x$, we see that β_0 is the mean response corresponding to $x = 0$. In many situations, this subpopulation does not exist or is not interesting. That is the case for Case 12.1, but Exercises 12.5 and 12.6 (page 577) are two settings where this information is meaningful.

The test of H_0: $\beta_1 = 0$ is always quite useful. When we substitute $\beta_1 = 0$ in the model, the x term drops out and we are left with

$$\mu_y = \beta_0$$

This model says that the mean of y does not vary with x. In other words, all the y's come from a single population with mean β_0, which we would estimate by $\bar{y}$ and then perform inference using the methods of Section 8.1. The hypothesis H_0: $\beta_1 = 0$, therefore, says that there is no straight-line relationship between y and x and that linear regression of y on x is of no value for predicting y.

EXAMPLE 12.5

Case 12.1

ENTRE

Does Loginc Increase with Educ? The Excel regression output in Figure 12.5 (page 579) for the entrepreneur problem contains the information needed for inference about the regression coefficients. You can see that the slope of the least-squares line is $b_1 = 0.1126$ and the standard error of this statistic is $\text{SE}_{b_1} = 0.0461$.

Given that the response y is on the log scale, this slope also approximates the percent change in the original variable for a unit change in x. In this case, one extra year of education is associated with an increase in income of approximately 11.3%.

A 95% confidence interval for the slope β_1 of the regression line in the population of all entrepreneurs in the United States is

$$\begin{aligned} b_1 \pm t^{*}\text{SE}_{b_1} &= 0.1126 \pm (1.984)(0.0461) \\ &= 0.1126 \pm 0.0915 \\ &= 0.0211 \text{ to } 0.2041 \end{aligned}$$

This interval contains only positive values, suggesting an increase in Loginc for an additional year of schooling. In terms of percent change, we are 95% confident that the average increase in income for one additional year of education is between 2.1% and 20.4%.

The t statistic and P-value for the test of H_0: $\beta_1 = 0$ against the two-sided alternative H_a: $\beta_1 \neq 0$ appear in the columns labeled "*t Stat*" and "*P-value*." The t statistic for the significance of the regression is

$$t = \frac{b_1}{\text{SE}_{b_1}} = \frac{0.1126}{0.0461} = 2.44$$

and the P-value for the two-sided alternative is 0.0164. If we expected beforehand that income rises with education, our alternative hypothesis would be one-sided, H_a: $\beta_1 > 0$. The P-value for this H_a is one-half the two-sided value given by Excel; that is, $P = 0.0082$. In both cases, there is strong evidence that the mean log income level increases as education increases.

The t distribution for this problem has $n - 2 = 98$ degrees of freedom. Table D has no row for 98 degrees of freedom. In Excel, the critical value and P-value can be obtained by using the functions = T.INV(0.975, 98) and = T.DIST.2T(2.44, 98), respectively. If you do not have access to software, we suggest taking a conservative approach and using the next *lower* degrees of freedom in Table D (80 degrees of freedom). This makes our interval a bit wider than we actually need for 95% confidence and the P-value a bit larger. ■

conservative, p. 421

In this example, we can discuss percent change in income for a unit change in education because the response variable y is on the log scale and x is not. In business and economics, we often encounter models in which both variables are on the log scale. In these cases, the slope approximates the percent change in y for a 1% change in x. This relationship is known as **elasticity,** a very important concept in economic theory.

elasticity

APPLY YOUR KNOWLEDGE

Treasury bills and inflation. *When inflation is high, lenders require higher interest rates to make up for the loss of purchasing power of their money while it is loaned out. Table 12.1 displays the return for six-month Treasury bills (annualized) and the rate of inflation as measured by the change in the government's Consumer Price Index in the same year.*[5] *An inflation rate of 5% means that the same set of goods and services costs 5% more. The data cover 60 years, from 1958 to 2017. Figure 12.9 is a scatterplot of these data. Figure 12.10 shows Excel regression output for predicting T-bill return from inflation rate. Exercises 12.8 through 12.10 ask you to use this information.* INFLAT

12.8 Look at the data. Give a brief description of the form, direction, and strength of the relationship between the inflation rate and the return on Treasury bills. What is the equation of the least-squares regression line for predicting T-bill return?

12.9 Is there a relationship? What are the slope b_1 of the fitted line and its standard error? Use these numbers to test by hand the hypothesis that there is no straight-line relationship between inflation rate and T-bill return against the alternative that the return on T-bills increases as the rate of inflation increases. State the hypotheses, give both the t statistic and its degrees of freedom, and use Table D to approximate the P-value. Then compare your results with those given by Excel. (Excel's P-value rounded to 2.40E-10 is shorthand for 0.00000000024. We would report this as "< 0.0001.")

TABLE 12.1 Return on Treasury bills and rate of inflation

Year	T-bill percent	Inflation percent	Year	T-bill percent	Inflation percent	Year	T-bill percent	Inflation percent
1958	3.01	1.76	1978	7.58	9.02	1998	4.83	1.61
1959	3.81	1.73	1979	10.04	13.20	1999	4.75	2.68
1960	3.20	1.36	1980	11.32	12.50	2000	5.90	3.39
1961	2.59	0.67	1981	13.81	8.92	2001	3.34	1.55
1962	2.90	1.33	1982	11.06	3.83	2002	1.68	2.38
1963	3.26	1.64	1983	8.74	3.79	2003	1.05	1.88
1964	3.68	0.97	1984	9.78	3.95	2004	1.58	3.26
1965	4.05	1.92	1985	7.65	3.80	2005	3.39	3.42
1966	5.06	3.46	1986	6.02	1.10	2006	4.81	2.54
1967	4.61	3.04	1987	6.03	4.43	2007	4.44	4.08
1968	5.47	4.72	1988	6.91	4.42	2008	1.62	0.09
1969	6.86	6.20	1989	8.03	4.65	2009	0.28	2.73
1970	6.51	5.57	1990	7.46	6.11	2010	0.20	1.50
1971	4.52	3.27	1991	5.44	3.06	2011	0.10	2.96
1972	4.47	3.41	1992	3.54	2.90	2012	0.13	1.74
1973	7.20	8.71	1993	3.12	2.75	2013	0.09	1.50
1974	7.95	12.34	1994	4.64	2.67	2014	0.06	0.76
1975	6.10	6.94	1995	5.56	2.54	2015	0.16	0.73
1976	5.26	4.86	1996	5.08	3.32	2016	0.46	2.07
1977	5.52	6.70	1997	5.18	1.70	2017	1.05	2.11

FIGURE 12.9 Scatterplot of the percent return on Treasury bills against the rate of inflation the same year, for Exercises 12.8 to 12.10.

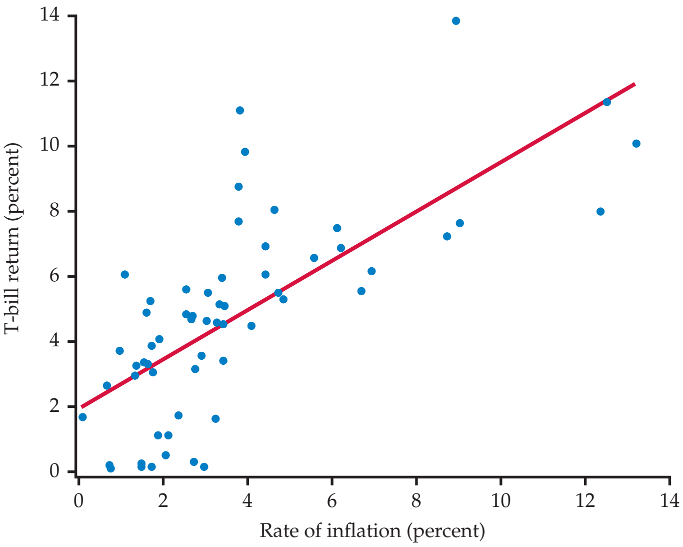
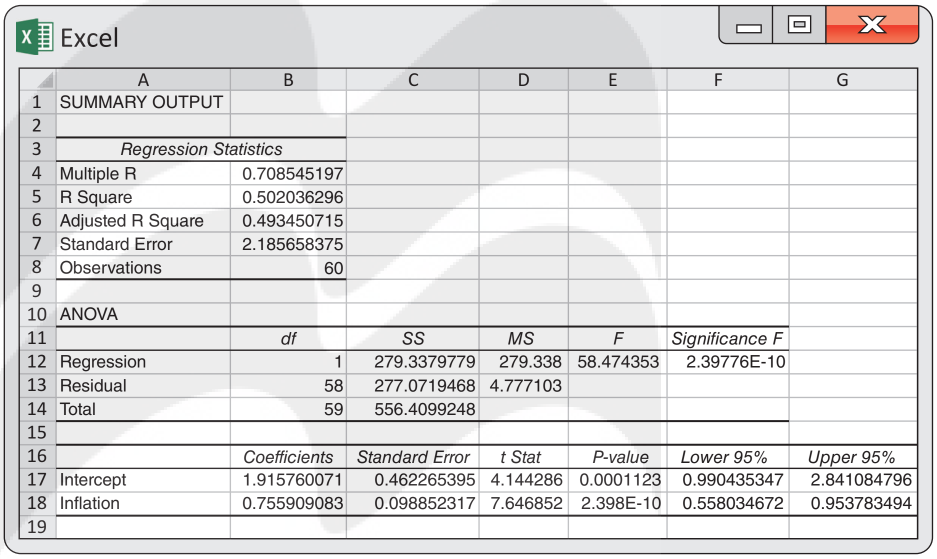


FIGURE 12.10 Excel output for the regression of the percent return on Treasury bills against the rate of inflation the same year, for Exercises 12.8 to 12.10.



	A	B	C	D	E	F	G
1	SUMMARY OUTPUT						
2							
3	*Regression Statistics*						
4	Multiple R	0.708545197					
5	R Square	0.502036296					
6	Adjusted R Square	0.493450715					
7	Standard Error	2.185658375					
8	Observations	60					
9							
10	ANOVA						
11		*df*	*SS*	*MS*	*F*	*Significance F*	
12	Regression	1	279.3379779	279.338	58.474353	2.39776E-10	
13	Residual	58	277.0719468	4.777103			
14	Total	59	556.4099248				
15							
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
17	Intercept	1.915760071	0.462265395	4.144286	0.0001123	0.990435347	2.841084796
18	Inflation	0.755909083	0.098852317	7.646852	2.398E-10	0.558034672	0.953783494
19							

12.10 Estimating the slope. Using Excel's values for b_1 and its standard error, find a 95% confidence interval for the slope β_1 of the population regression line. Compare your result with Excel's 95% confidence interval. What does the confidence interval tell you about the change in the T-bill return rate for a 1% increase in the inflation rate?

The word "regression"

To "regress" means to go backward. Why are statistical methods for predicting a response from an explanatory variable called "regression"? Sir Francis Galton (1822–1911) was the first to apply regression to biological and psychological data. He looked at examples such as the heights of children versus the heights of their parents. He found that the taller-than-average parents tended to have children who were also taller than average, but not as tall as their parents. Galton called this fact "regression toward mediocrity," and the name came to be applied to the statistical method. Galton also invented the correlation coefficient r and named it “correlation.”

Why are the children of tall parents shorter on the average than their parents? The parents are tall in part because of their genes. But they are also tall in part by chance. Looking at tall parents selects those in whom chance produced height. Their children inherit their genes, but not necessarily their good luck. As a group, the children are taller than average (genes), but their heights vary by chance about the average, some upward and some downward. The children, unlike the parents, were not selected because they were tall and thus, on average, are shorter. A similar argument can be used to describe why children of short parents tend to be taller than their parents.

Here’s another example. Students who score at the top on the first exam in a course are likely to do less well on the second exam. Does this show that they stopped studying? No—they scored high in part because they knew the material but also in part because they were lucky. On the second exam, they may still know the material but be less lucky. As a group, they will still do better than average but not as well as they did on the first exam. The students at the bottom on the first exam will tend to move up on the second exam, for the same reason.

regression fallacy

The **regression fallacy** is the assertion that *regression toward the mean* shows that there is some systematic effect at work: students with top scores now work less hard, or managers of last year’s best-performing mutual funds lose their touch this year, or heights get less variable with each passing generation as tall parents have shorter children and short parents have taller children. The Nobel economist Milton Friedman says, “I suspect that the regression fallacy is the most common fallacy in the statistical analysis of economic data.”[6] Beware.

APPLY YOUR KNOWLEDGE

12.11 Hot funds? Explain carefully to a naive investor why the mutual funds that had the highest returns this year will, as a group, probably do less well relative to other funds next year.

12.12 Mediocrity triumphant? In the early 1930s, a man named Horace Secrist wrote a book titled *The Triumph of Mediocrity in Business*. Secrist found that businesses that did unusually well or unusually poorly in one year tended to be nearer the average in profitability at a later year. Why is it a fallacy to say that this fact demonstrates an overall movement toward “mediocrity”?

Inference about correlation

The correlation between log income and level of education for the 100 entrepreneurs is $r = 0.2394$. This value appears in the Excel output in Figure 12.5 (page 579), where it is labeled “Multiple R.”[7] We might expect a positive correlation between these two measures in the population of all entrepreneurs in the United States. Is the sample result convincing evidence that this is true?

population correlation ρ

This question concerns a new population parameter, the **population correlation.** This is the correlation between the log income and level of education when we measure these variables for every member of the population. We call the population correlation ρ, the Greek letter rho. To assess the evidence that $\rho > 0$ in the population, we must test the hypotheses

$$H_0\colon \rho = 0$$

$$H_a\colon \rho > 0$$

It is natural to base the test on the sample correlation $r = 0.2394$. Indeed, most computer packages with routines to calculate sample correlations provide the result of this significance test. We can also use regression software by exploiting the close link between correlation and the regression slope. The population correlation ρ is zero, positive, or negative exactly when the slope β_1 of the population regression line is zero, positive, or negative, respectively. In fact, the t statistic for testing H_0: $\beta_1 = 0$ also tests H_0: $\rho = 0$. What is more, this t statistic can be written in terms of the sample correlation r.

TEST FOR ZERO POPULATION CORRELATION

To test the hypothesis H_0: $\rho = 0$, either use the t statistic for the regression slope or compute this statistic from the sample correlation r:

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

This t statistic has $n - 2$ degrees of freedom.

EXAMPLE 12.6

CASE 12.1

ENTRE

Correlation between Loginc and Educ The sample correlation between Loginc and Educ is $r = 0.2394$ for a sample of size $n = 100$. Figure 12.11 contains Minitab output for this correlation calculation. Minitab calls this a Pearson correlation to distinguish it from other kinds of correlations it can calculate. The P-value for a two-sided test of H_0: $\rho = 0$ is 0.016 and the P-value for our one-sided alternative is 0.008.

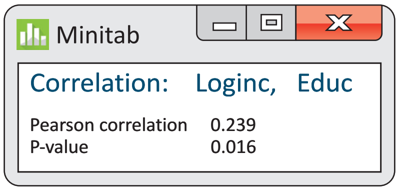
We can also get this result from the Excel output in Figure 12.5 (page 579). In the "Educ" line, notice that $t = 2.441$ with two-sided P-value 0.0164. Thus, $P = 0.00082$ for our one-sided alternative.

Finally, we can calculate t directly from r as follows:

$$\begin{aligned} t &= \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \\ &= \frac{0.2394\sqrt{100-2}}{\sqrt{1-(0.2394)^2}} \\ &= \frac{2.3699}{0.9709} = 2.441 \end{aligned}$$

If we are not using software, we can compare $t = 2.441$ with critical values from the t table (Table D) with 80 (largest row less than or equal to $n - 2 = 98$) degrees of freedom. ■

The alternative formula for the test statistic is convenient because it uses only the sample correlation r and the sample size n. Remember that correlation, unlike regression, does not require a distinction between the explanatory and response variables. For variables x and y, there are two regressions (y on x and x on y) but just one correlation. Both regressions produce the same t statistic.



Correlation: Loginc, Educ

```
Pearson correlation   0.239
P-value               0.016
```

FIGURE 12.11 Minitab output for the correlation between log average income and years of education, for Example 12.6.

The distinction between the regression setting and correlation is important only for understanding the conditions under which the test for zero population correlation makes sense. In the regression model, we take the values of the explanatory variable x as given. The values of the response y are Normal random variables, with means that are a straight-line function of x. In the model for testing correlation, we think of the setting where we obtain a random sample from a population and measure both x and y. Both are assumed to be Normal random variables. In fact, they are taken to be **jointly Normal.** This implies that the conditional distribution of y for each possible value of x is Normal, just as in the regression model.

jointly Normal

APPLY YOUR KNOWLEDGE

12.13 T-bills and inflation. We expect the interest rates on Treasury bills to rise when the rate of inflation rises and to fall when inflation falls. That is, we expect a positive correlation between the return on T-bills and the inflation rate.

(a) Find the sample correlation r for the 60 years in Table 12.1 in the Excel output in Figure 12.10 (page 586).

(b) From r, calculate the t statistic for testing correlation. What are its degrees of freedom? Use Table D to give an approximate P-value. Compare your result with the P-value from part (a).

(c) Verify that your t for correlation calculated in part (b) has the same value as the t for slope in the Excel output.

CASE 12.1 **12.14 Two regressions.** We have regressed Loginc on Educ, with the results appearing in Figures 12.5 and 12.6. Use software to regress Educ on Loginc for the same data. ENTRE

(a) What is the equation of the least-squares line for predicting years of education from log income? Is it a different line than the regression line in Figure 12.4? To answer this question, plot two points for each equation and draw a line connecting them.

(b) Verify that the two lines cross at the mean values of the two variables. That is, substitute the mean Educ into the line in Figure 12.5, and show that the predicted log income equals the mean of Loginc of the 100 subjects. Then substitute the mean Loginc into your new line, and show that the predicted years of education equals the mean Educ for the entrepreneurs.

(c) Verify that the two regressions give the same value of the t statistic for testing the hypothesis of zero population slope. You could use either regression to test the hypothesis of zero population correlation.

SECTION 12.1 SUMMARY

- **Least-squares regression** fits a straight line to data to predict a quantitative response variable y from a quantitative explanatory variable x. Inference about regression requires additional conditions.

- The **simple linear regression model** says that a **population regression line** $\mu_y = \beta_0 + \beta_1 x$ describes how the mean response in an entire population varies as x changes. The observed response y for any x has a Normal distribution with a mean given by the population regression line and with the same standard deviation σ for any value of x.

- The **parameters** of the simple linear regression model are the intercept β_0, the slope β_1, and the regression standard deviation σ. The **slope b_1** and

intercept b_0 of the least-squares line estimate the slope β_1 and intercept β_0 of the population regression line, respectively.

- The parameter σ is estimated by the **regression standard error**

$$s = \sqrt{\frac{1}{n-2}\sum(y_i - \hat{y}_i)^2}$$

where the differences between the observed and predicted responses are the **residuals**

$$e_i = y_i - \hat{y}_i$$

- Prior to inference, always examine the residuals for Normality, constant variance, and any other remaining patterns in the data. **Plots of the residuals** are commonly used as part of this examination.

- The regression standard error s has $n - 2$ **degrees of freedom.** Inference about β_0 and β_1 uses t distributions with $n - 2$ degrees of freedom.

- **Confidence intervals for the slope** of the population regression line have the form $b_1 \pm t^*\text{SE}_{b_1}$. In practice, you will use software to find the slope b_1 of the least-squares line and its standard error SE_{b_1}.

- To test the hypothesis that the population slope is zero, use the ***t* statistic** $t = b_1 / \text{SE}_{b_1}$, also given by software. This null hypothesis says that straight-line dependence on x has no value for predicting y.

- The t test for zero population slope also tests the null hypothesis that the **population correlation** is zero. This t statistic can be expressed in terms of the sample correlation, $t = r\sqrt{n-2} / \sqrt{1-r^2}$.

SECTION 12.1 EXERCISES

For Exercises 12.1 and 12.2, see page 574; for 12.3 and 12.4, see page 576; for 12.5 and 12.6, see page 577; for 12.7, see page 580; for 12.8 to 12.10, see pages 585–586; for 12.11 and 12.12, see page 587; and for 12.13 and 12.14, see page 589.

12.15 Assessment value versus sales price. Real estate is typically assessed annually for property tax purposes. This assessed value, however, is not necessarily the same as the fair market value of the property. Table 12.2 lists the sales price and assessed value for an SRS of 35 properties recently sold in a midwestern county.[8] Both variables are measured in thousands of dollars.

HSALES

(a) What proportion have a selling price greater than the assessed value? Do you think this proportion is a good estimate for the larger population of all homes recently sold? Explain your answer.

(b) Make a scatterplot with assessed value on the horizontal axis. Briefly describe the relationship between assessed value and selling price.

(c) Based on the scatterplot, there are two properties with very large assessed values. Do you think it is more appropriate to consider all 35 properties for linear regression analysis or to just consider the 33 properties? Explain your decision.

(d) Report the least-squares regression line for predicting selling price from assessed value using all 35 properties. What is the regression standard error?

(e) Now remove the two properties with the highest assessments and refit the model. Report the least-squares regression line and regression standard error.

(f) Compare the two sets of results. Describe how these large x values impact the results.

12.16 Assessment value versus sales price, continued. Refer to the previous exercise. Let's consider linear regression analysis using all 35 properties.

HSALES

(a) Obtain the residuals and plot them versus assessed value. Is there anything unusual to report? Describe the reasoning behind your answer.

(b) Do the residuals appear to be approximately Normal? Describe how you assessed this.

(c) Do you think all the conditions for inference are approximately met? Explain your answer.

(d) Construct a 95% confidence interval for the intercept and slope, and summarize the results.

TABLE 12.2 Sales price and assessed value (in thousands of $) of 35 homes in a midwestern county

Property	Sales price	Assessed value	Property	Sales price	Assessed value	Property	Sales price	Assessed value
1	116.9	94.9	13	200.0	205.6	25	200.0	200.6
2	161.0	160.0	14	146.6	152.9	26	162.5	92.3
3	202.0	233.3	15	215.0	167.4	27	256.8	251.0
4	300.0	255.1	16	125.0	139.3	28	286.0	184.3
5	137.5	123.9	17	139.9	128.2	29	90.0	102.0
6	178.0	157.4	18	238.0	198.2	30	284.3	272.4
7	350.0	395.5	19	120.9	93.4	31	229.9	217.0
8	150.9	126.8	20	142.5	92.3	32	235.0	199.7
9	122.5	109.7	21	282.2	257.6	33	419.0	335.8
10	270.5	241.9	22	279.0	243.5	34	149.0	209.8
11	267.5	254.4	23	110.0	109.2	35	255.4	258.1
12	174.9	135.0	24	130.0	125.1			

12.17 Are the assessment value and sales price different? Refer to the previous two exercises. HSALES

(a) Again create the scatterplot with assessed value on the horizontal axis. If, on average, sales price and the assessed value are the same, the population regression line should be $y = x$. Draw this line on your scatterplot and compare it to the least squares line.

(b) Explain why we cannot simply test H_0: $\beta_1 = 1$ versus the two-sided alternative to assess if the least-squares line is different from $y = x$.

(c) Use methods from Chapter 8 to test the hypothesis that, on average, the sales price equals the assessed value.

12.18 Are female CEOs older? A pair of researchers looked at the age and sex of large sample of CEOs.[9] To investigate the relationship between these two variables, they fit a regression model with age as the response variable and sex as the explanatory variable. The explanatory variable was coded $x = 0$ for males and $x = 1$ for females. The resulting least-squares regression line was

$$\hat{y} = 55.643 - 2.205x$$

(a) What is the expected age for a male CEO ($x = 0$)?

(b) What is the expected age for a female CEO ($x = 1$)?

(c) What is the difference in the expected age of female and male CEOs?

(d) Relate your answers to parts (a) and (c) to the least-squares estimates b_0 and b_1.

(e) The t statistic for testing H_0: $\beta_1 = 0$ was reported as -6.474. Based on this result, what can you conclude about the average ages of female and male CEOs?

(f) To compare the average age of male and female CEOs, the researchers could have instead performed a two-sample t test (Chapter 8). Will this regression approach provide the same result? Explain your answer.

TABLE 12.3 In-state tuition and fees (in dollars) for 33 public universities

School	2013	2017	School	2013	2017	School	2013	2017
Penn State	16,992	18,436	Ohio State	10,037	10,591	Texas	9790	10,136
Pittsburgh	17,100	19,080	Virginia	12,458	16,781	Nebraska	8075	8901
Michigan	13,142	14,826	California–Davis	13,902	14,382	Iowa	8061	8964
Rutgers	13,499	14,638	California–Berkeley	12,864	13,928	Colorado	10,529	12,086
Michigan State	12,908	14,460	California–Irvine	13,149	15,516	Iowa State	7726	8636
Maryland	9161	10,399	Purdue	9992	9992	North Carolina	8340	9005
Illinois	14,750	15,868	California–San Diego	13,302	14,028	Kansas	10,107	10,824
Minnesota	13,618	14,417	Oregon	9763	11,571	Arizona	10,391	11,877
Missouri	10,104	9787	Wisconsin	10,403	10,533	Florida	6263	6381
Buffalo	7022	7976	Washington	12,397	10,974	Georgia Tech	10,650	12,418
Indiana	10,209	10,533	UCLA	12,696	13,749	Texas A&M	8506	10,403

12.19 Public university tuition: 2013 versus 2017. Table 12.3 shows the in-state undergraduate tuition in 2013 and 2017 for 33 public universities.[10] TUIT

(a) Plot the data with the 2013 tuition on the x axis and describe the relationship. Are there any outliers or unusual values? Does a linear relationship between the tuition in 2013 and 2017 seem reasonable?

(b) Fit the simple linear regression model and give the least-squares regression line and regression standard error.

(c) Obtain the residuals and plot them versus the 2013 tuition amount. Describe anything unusual in the plot.

(d) Do the residuals appear to be approximately Normal? Explain.

(e) Remove any unusual observations and repeat parts (b)–(d).

(f) Compare the two sets of least-squares results. Describe any impact these unusual observations have on the results.

12.20 More on public university tuition. Refer to the previous exercise. Use all 33 observations for this exercise. TUIT

(a) Give the null and alternative hypotheses for examining if there is a linear relationship between 2013 and 2017 tuition amounts.

(b) Write down the test statistic and P-value for the hypotheses stated in part (a). State your conclusions.

(c) Construct a 95% confidence interval for the slope. What does this interval tell you about the annual percent increase in tuition between 2013 and 2017?

(d) The tuition at CashCow U was \$9200 in 2013. What is the predicted tuition in 2017?

(e) The tuition at Moneypit U was \$18,895 in 2013. What is the predicted tuition in 2017?

(f) Discuss the appropriateness of using the fitted equation to predict tuition for each of these universities.

12.21 The timing of initial public offerings. Initial public offerings (IPOs) have tended to group together in time and in sector of business. Some researchers hypothesize this clustering is due to managers either speeding up or delaying the IPO process in hopes of taking advantage of a "hot" market, which will provide the firm with high initial valuations of its stock.[11] The researchers collected information on 196 public offerings listed on the Warsaw Stock Exchange over a six-year period. For each IPO, they obtained the length of the IPO offering period (the time between the approval of the prospectus and the IPO date) and three market return rates. The first rate was for the period between the date the prospectus was approved and the "expected" IPO date. The second rate was for the period 90 days prior to the "expected" IPO date. The last rate was between the approval date and 90 days after the "expected" IPO date. The "expected" IPO date was the median length of the 196 IPO periods. They regressed the length of the offering period (in days) against each of the three rates of return. Here are the results:

Period	b_0	b_1	P-value	r
1	48.018	−129.391	0.0008	−0.238
2	49.478	−114.785	<0.0001	−0.414
3	47.613	−41.646	0.0463	−0.143

(a) What does this table tell you about the relationship between the IPO offering period and the three market return rates?

(b) The researchers argue that since the strongest correlation is for the second period and the weakest correlation is for the third period, there is evidence supporting their hypothesis. Do you agree with this conclusion? Explain your answer.

12.22 The relationship between log income and education level for employees. Recall Case 12.1 (page 572). The researchers also looked at the relationship between education and log income for employees. An employee was defined as a person whose main employment status is a salaried job. Based on a sample of 100 employees: EMPL

(a) Construct a scatterplot of log income versus education. Describe the relationship between the two variables. Is a linear relationship reasonable? Explain your answer.

(b) Report the least-squares regression line.

(c) Obtain the residuals and use them to assess the assumptions needed for inference.

(d) In Example 12.5 (page 584), we constructed a 95% confidence interval for the slope of the entrepreneur population; it was (0.0208 to 0.2044). Construct a 95% confidence interval for the slope of the employee population.

(e) Compare the two confidence intervals. Do you think there is a difference in the two slopes? Explain your answer.

12.23 Incentive pay and job performance. In the National Football League (NFL), incentive bonuses now account for roughly 25% of player compensation.[12] Does tying a player's salary to performance bonuses result in better individual or team success on the field? Focusing on linebackers, let's look at the relationship between a player's end-of-the-year production rating and the percent of his salary devoted to incentive payments in that same year. PERPLAY

(a) Use numerical and graphical methods to describe the two variables and summarize your results.

(b) Neither variable is Normally distributed. Does that necessarily pose a problem for performing linear regression? Explain.

(c) Construct a scatterplot of the data and describe the relationship. Are there any outliers or unusual values? Does a linear relationship between the percent of salary from incentive payments and player rating seem reasonable? Is it a very strong relationship? Explain.

(d) Run the simple linear regression and give the least-squares regression line.

(e) Obtain the residuals and assess whether the assumptions for the linear regression analysis are reasonable. Include all plots and numerical summaries that you used to make this assessment.

12.24 Incentive pay and job performance, continued. Refer to the previous exercise. PERPLAY

(a) Now run the simple linear regression for the variables square root of rating and percent of salary from incentive payments.

(b) Obtain the residuals and assess whether the assumptions for the linear regression analysis are reasonable. Include all plots and numerical summaries that you used to make this assessment.

(c) Construct a 95% confidence interval for the square root increase in rating given a 1% increase in the percent of salary from incentive payments.

(d) Consider the values 0%, 20%, 40%, 60%, and 80% salary from incentives. Compute the predicted rating for this model and for the one in Exercise 12.23. For the model in this exercise, you will need to square the predicted value to get back to the original units.

(e) Plot the predicted values versus the percents, and connect those values from the same model. For which regions of percent do the predicted values from the two models vary the most?

(f) Based on your comparison of the regression models (both predicted values and residuals), which model do you prefer? Explain.

12.25 Predicting public university tuition: 2008 versus 2017. Refer to Exercise 12.19. The data file also includes the in-state undergraduate tuition for the year 2008. TUIT

(a) Plot the data with the 2008 tuition on the *x* axis, then describe the relationship. Are there any outliers or unusual values? Does a linear relationship between the tuition in 2008 and 2017 seem reasonable?

(b) Fit the simple linear regression model and give the least-squares regression line and regression standard error.

(c) Obtain the residuals and plot them versus the 2008 tuition amount. Describe anything unusual in the plot.

(d) Do the residuals appear to be approximately Normal? Explain.

12.26 Compare the analyses. In Exercises 12.19 and 12.25, you used two different explanatory variables to predict university tuition in 2017. Summarize the two analyses and compare the results. If you had to choose between the two, which explanatory variable would you choose? Give reasons for your answers.

Age and income. *The data file for the following exercises contains the age and income of a random sample of 5712 men between the ages of 25 and 65 who have a bachelor's degree but no higher degree. Figure 12.12 is a scatterplot of these data. Figure 12.13 displays Excel output for regressing income on age. The line in the scatterplot is the least-squares regression line. Exercises 12.27 through 12.29 ask you to interpret this information.* INAGE

12.27 Looking at age and income. The scatterplot in Figure 12.12 has a distinctive form.

(a) Age is recorded as of the last birthday. How does this explain the vertical stacks of incomes in the scatterplot?

(b) Give some reasons that older men in this population might earn more than younger men. Give some reasons that younger men might earn more than older men. What do the data show about the relationship between age and income in the sample? Is the relationship very strong?

(c) What is the equation of the least-squares line for predicting income from age? What specifically does the slope of this line tell us?

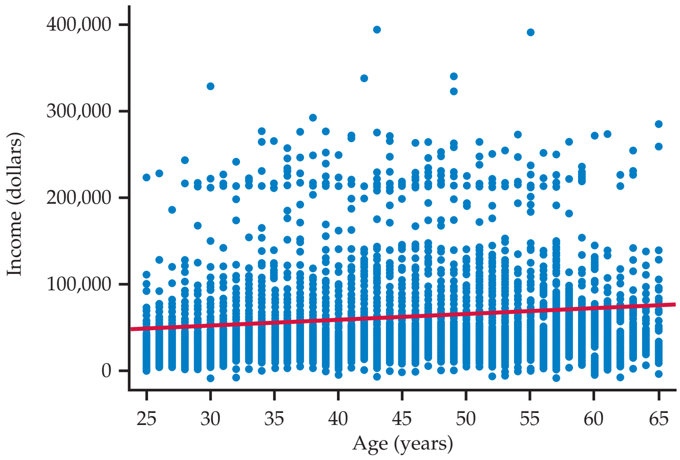
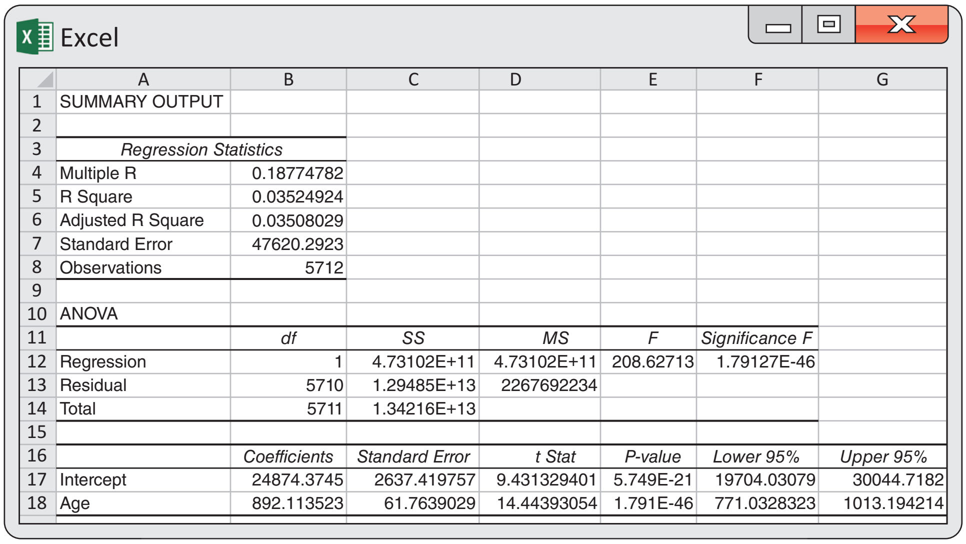


FIGURE 12.12 Scatterplot of income against age for a random sample of 5712 men aged 25 to 65, for Exercises 12.27 to 12.29.



	A	B	C	D	E	F	G
1	SUMMARY OUTPUT						
2							
3	*Regression Statistics*						
4	Multiple R	0.18774782					
5	R Square	0.03524924					
6	Adjusted R Square	0.03508029					
7	Standard Error	47620.2923					
8	Observations	5712					
9							
10	ANOVA						
11		*df*	*SS*	*MS*	*F*	*Significance F*	
12	Regression	1	4.73102E+11	4.73102E+11	208.62713	1.79127E-46	
13	Residual	5710	1.29485E+13	2267692234			
14	Total	5711	1.34216E+13				
15							
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
17	Intercept	24874.3745	2637.419757	9.431329401	5.749E-21	19704.03079	30044.7182
18	Age	892.113523	61.7639029	14.44393054	1.791E-46	771.0328323	1013.194214

FIGURE 12.13 Excel output for the regression of income on age, for Exercises 12.27 to 12.29.

12.28 Income increases with age. We see that older men do, on average, earn more than younger men, but the increase is not very rapid. (Note that the regression line describes many men of different ages—data on the same men over time might show a different pattern.)

(a) We know even without looking at the Excel output that there is highly significant evidence that the slope of the population regression line is greater than 0. Why do we know this?

(b) Excel gives a 95% confidence interval for the slope of the population regression line. What is this interval?

(c) Give a 99% confidence interval for the slope of the population regression line.

12.29 Was inference justified? You see from Figure 12.12 that the incomes of men at each age are (as expected) not Normal but right-skewed.

(a) How is this apparent on the plot?

(b) Nonetheless, your confidence interval in the previous exercise will be quite accurate even though it is based on Normal distributions. Why?

12.30 Regression to the mean? Suppose a large population of test takers take the GMAT. You fear some cheating may have occurred so you ask those people who scored in the top 10% to take the exam again.

(a) If their scores, on average, decrease, is this evidence that there was cheating? Explain your answer.

(b) If these same people were asked to take the test a third time, would you expect their scores to decline even further? Explain your answer.

12.31 T-bills and inflation. Exercises 12.8 through 12.10 interpret the part of the Excel output in Figure 12.10 (page 586) that concerns the slope—that is, the rate at which T-bill returns increase as the rate of inflation increases. Use this output to answer questions about the intercept.

(a) The intercept β_0 in the regression model is meaningful in this example. Explain what β_0 represents. Why should we expect β_0 to be greater than 0?

(b) What values does Excel give for the estimated intercept b_0 and its standard error SE_{b_0}?

(c) Is there good evidence that β_0 is greater than 0?

(d) Write the formula for a 95% confidence interval for β_0. Verify that the hand calculation (using the Excel values for b_0 and SE_{b_0}) agrees approximately with the output in Figure 12.10.

12.32 Is the correlation significant? Two studies looked at the relationship between customer-relationship management (CRM) implementation and organizational structure. One study reported a correlation of $r = 0.33$ based on a sample of size $n = 25$. The second study reported a correlation of $r = 0.22$ based on a sample of size $n = 62$. For each, test the null hypothesis that the population correlation $\rho = 0$ against the one-sided alternative $\rho > 0$. Are the results significant at the 5% level? What conclusions would you draw based on both studies?

12.33 Correlation between the prevalences of adult binge drinking and underage drinking. A group of researchers compiled data on the prevalence of adult binge drinking and the prevalence of underage drinking in 42 states.[13] A correlation of 0.32 was reported.

(a) Test the null hypothesis that the population correlation $\rho = 0$ against the alternative $\rho > 0$. Are the results significant at the 5% level?

(b) Explain this correlation in terms of the direction of the association and the percent of variability in the prevalence of underage drinking that is explained by the prevalence of adult binge drinking.

TABLE 12.4 Net new money (millions of $) flowing into stock and bond mutual funds

Year	Stocks	Bonds	Year	Stocks	Bonds	Year	Stocks	Bonds
1984	4336	13,058	1996	216,937	3141	2008	−215,757	30,039
1985	6643	63,127	1997	227,106	29,166	2009	2013	371,123
1986	20,386	102,618	1998	156,875	74,656	2010	−24,385	232,351
1987	19,231	6797	1999	187,565	−4767	2011	−129,363	117,734
1988	−14,948	−4488	2000	315,705	−50,115	2012	−152,678	306,256
1989	6774	−1226	2001	33,483	88,463	2013	159,481	−70,771
1990	12,915	6833	2002	−29,310	141,865	2014	25,458	43,600
1991	39,888	59,258	2003	144,077	32,750	2015	−75,620	−25,270
1992	78,983	70,989	2004	171,945	−15,102	2016	−258,030	106,897
1993	127,260	72,169	2005	123,938	25,294	2017	−159,640	260,162
1994	114,525	−61,362	2006	147,804	59,448			
1995	124,392	−5922	2007	73,307	110,609			

(c) The researchers collected information from 42 of 50 states so almost all the data available was used in the analysis. Provide an argument for the use of statistical inference in this setting.

12.34 Stocks and bonds. How is the flow of investors' money into stock mutual funds related to the flow of money into bond mutual funds? Table 12.4 shows the net new money flowing into stock and bond mutual funds in the years 1984 to 2017, in millions of dollars.[14] "Net" means that funds flowing out are subtracted from those flowing in. If more money leaves than arrives, the net flow will be negative. FLOW

(a) Make a scatterplot with cash flow into stock funds as the explanatory variable. Find the least-squares line for predicting net bond investments from net stock investments. What do the data suggest?

(b) Is there statistically significant evidence of some straight-line relationship between the flows of cash into bond funds and stock funds? (State the hypotheses, give a test statistic and its *P*-value, and state your conclusion.)

(c) Generate a plot of the residuals versus year. Describe any unusual patterns you see in this plot.

(d) Given the 2008 financial crisis and its lingering effects, remove the data for the years after 2007 and refit the remaining years. Is there statistically significant evidence of a straight-line relationship?

(e) Compare the least-squares regression lines and regression standard errors using all the years and using only the years before 2008.

(f) How would you report these results in a paper? In other words, how would you handle the difference in relationship before and after 2008?

12.35 Size and selling price of houses. Table 12.5 describes a random sample of 30 houses sold in a

TABLE 12.5 Selling price and size of homes

Price ($1000)	Size (sq ft)	Price ($1000)	Size (sq ft)	Price ($1000)	Size (sq ft)
268	1897	142	1329	83	1378
131	1157	107	1040	125	1668
112	1024	110	951	60	1248
112	935	187	1628	85	1229
122	1236	94	816	117	1308
128	1248	99	1060	57	892
158	1620	78	800	110	1981
135	1124	56	492	127	1098
146	1248	70	792	119	1858
126	1139	54	980	172	2010

Midwest city during a recent year.[15] In this exercise, we examine the relationship between size and price. HSIZE

(a) Plot the selling price versus the number of square feet. Describe the pattern. Does r^2 suggest that size is quite helpful for predicting selling price?

(b) Do a linear regression analysis. Give the least-squares line and the results of the significance test for the slope. What does your test tell you about the relationship between house size and selling price?

12.36 Are inflows into stocks and bonds correlated? Is the correlation between net flow of money into stock mutual funds and into bond mutual funds significantly different from 0? Use the regression analysis you did in Exercise 12.34, part (b), to answer this question with no additional calculations. FLOW

12.37 Do larger houses have higher prices? We expect that there is a positive correlation between the sizes of houses in the same market and their selling prices. HSIZE

(a) Use the data in Table 12.5 to test this hypothesis. (State the hypotheses, find the sample correlation r and the t statistic based on it, and give an approximate P-value and your conclusion.)

(b) How do your results in part (a) compare to the test of the slope in Exercise 12.35, part (b)?

(c) To what extent do you think that these results would apply to other regions in the United States?

12.38 Highway MPG and CO_2 emissions. Let's investigate the relationship between highway miles per gallon (MPGHwy) and carbon dioxide emissions (CO2 Emissions) for cars that use premium gasoline as reported by Natural Resources Canada.[16] PREM

(a) Make a scatterplot of the data and describe the pattern.

(b) Plot MPGHwy versus the logarithm of CO_2 emissions. Are these points closer to a straight line?

(c) Regress MPGHwy by the logarithm of CO_2 emissions. Give a 95% confidence interval for the slope of the population regression line. Describe what this interval tells you in terms of percent change in CO_2 emissions for every one mile increase in highway miles per gallon.

12.39 Influence? Your scatterplot in Exercise 12.35 shows one house whose selling price is quite high for its size. Rerun the analysis without this outlier. Does this one house influence r^2, the location of the least-squares line, or the t statistic for the slope in a way that would change your conclusions? HSIZE

12.40 Correlation between the observed and predicted y's. Using your choice of software, fit the data of Case 12.1 and obtain the log income predicted values $\hat{y}$. Then compute the correlation between these predicted values $\hat{y}$ and log income values y and compare it to the correlation between x and y that is reported in Example 12.2 (page 574). Describe what you find. ENTRE

12.2 Using the Regression Line

When you complete this section, you will be able to:

- Construct and interpret a confidence interval for a mean response when $x = x^*$.
- Construct and interpret a prediction interval for a future observation when $x = x^*$.
- Identify when a prediction interval should be constructed instead of a confidence interval.

Predictive analytics involves the use of various techniques from statistics to make predictions that help support decision making. One of the most common reasons to fit a line to data is to predict the response to a particular value of the explanatory, or predictor, variable. The method is simple: just substitute the value of x into the equation of the line. For example, the least-squares line for predicting log income of entrepreneurs from their years of education (Case 12.1) is

$$\hat{y} = 8.2546 + 0.1126x$$

For an Educ of 16, our least-squares regression equation gives the prediction

$$\hat{y} = 8.2546 + (0.1126)(16) = 10.0562$$

Confidence and prediction intervals

In terms of inference, there are two different uses of this prediction. First, we can estimate the *mean* log income in the subpopulation of entrepreneurs with 16 years of education. Second, we can predict the log income of *one individual entrepreneur* with 16 years of education.

For each use, the actual prediction is the same, $\hat{y} = 10.0562$. *It is the margin of error that is different.* Individual entrepreneurs with 16 years of education don't all have the same log income. Thus, we need a larger margin of error when predicting an individual's log income than when estimating the mean log income of all entrepreneurs who have 16 years of education.

To emphasize the distinction between predicting a single outcome and estimating the mean of all outcomes in the subpopulation, we use different terms for the two resulting intervals.

- To estimate the *mean* response, we use a *confidence interval*. This is an ordinary confidence interval for the parameter

$$\mu_y = \beta_0 + \beta_1 x^*$$

The regression model says that μ_y is the mean of responses y when $x = x^*$. It is a fixed number whose value we don't know because we don't know β_0 and β_1.

prediction interval, p. 383

- To estimate an *individual* response y, we use a *prediction interval.* A prediction interval estimates a single random response y rather than a parameter like μ_y. Even if we know β_0 and β_1, the response y is not a fixed number. The model says that y varies Normally with a mean that depends on x.

Fortunately, the meaning of a prediction interval is very much like the meaning of a confidence interval. A 95% prediction interval, like a 95% confidence interval, is right 95% of the time in repeated use. Consider doing the following many times:

1. Draw a sample of n observations (x, y) and one additional observation (x^*, y).
2. Calculate the 95% prediction interval for y when $x = x^*$ using the n observations.

Being right means that the y value in the additional observation will be in the calculated interval 95% of the time.

Each interval has the usual form

$$\hat{y} \pm t^*\text{SE}$$

where t^*SE is the margin of error. The main distinction is that because it is more difficult to predict a single observation (random variable) than the mean of a subpopulation (fixed value), the margin of error for the prediction interval is wider than the margin of error for the confidence interval. Formulas for computing these quantities are given in Section 12.3. For now, we rely on software to do the arithmetic.

CONFIDENCE AND PREDICTION INTERVALS FOR REGRESSION RESPONSE

A level C **confidence interval for the mean response** μ_y when x takes the value x^* is

$$\hat{y} \pm t^*\text{SE}_{\hat{\mu}}$$

Here, $\text{SE}_{\hat{\mu}}$ is the standard error for estimating a mean response. The **margin of error** is $m = t^*\text{SE}_{\hat{\mu}}$.

A level C **prediction interval for a single observation** on y when x takes the value x^* is

$$\hat{y} \pm t^* \mathrm{SE}_{\hat{y}}$$

Here, $\mathrm{SE}_{\hat{y}}$ is the standard error for predicting an individual response and the **margin of error** is $m = t^* \mathrm{SE}_{\hat{y}}$.

The standard error $\mathrm{SE}_{\hat{y}}$ is larger than the standard error $\mathrm{SE}_{\hat{\mu}}$.

In both cases, t^* is the value for the $t(n-2)$ density curve with area C between $-t^*$ and t^*.

Before moving on to the examples, it is important to note that predicting an individual response is an exception to the general fact that regression inference is robust against lack of Normality. *The prediction interval relies on Normality of individual observations, not just on the approximate Normality of statistics like the slope b_1 and intercept b_0 of the least-squares line.* In practice, this means that we should regard prediction intervals as rough approximations.

EXAMPLE 12.7

CASE 12.1

ENTRE

Predicting Loginc from Educ Jacob Brown is an entrepreneur with Educ = 16 years of education. We don't know his log income, but we can use the data on other entrepreneurs to predict it.

Statistical software usually allows prediction of the response for each x-value in the data and also for new values of x. Here is the output from the prediction option in the Minitab regression command for $x^* = 16$ when we ask for 95% intervals:

```
    Fit        SE Fit           95% CI                 95% PI
  10.0560     0.167802    (9.72305, 10.3890)    (7.81924, 12.2929)
```

The "Fit" entry gives the predicted log income, 10.0560. This agrees with our hand calculation within the rounding error. Minitab gives both 95% intervals; you must then choose which one you want. We are predicting a single response, so the prediction interval "95% PI" is the right choice. We are 95% confident that Jacob's log income lies between 7.81924 and 12.2929. This is a wide range because the data are widely scattered about the least-squares line. The 95% confidence interval for the mean log income of all entrepreneurs with EDUC = 16, given as "95% CI," is much narrower. ■

Note that Minitab reports only one of the two standard errors—the standard error for estimating the mean response, $\mathrm{SE}_{\hat{\mu}} = 0.1678$. A graph will help us to understand the difference between the two types of intervals.

EXAMPLE 12.8

CASE 12.1

ENTRE

Comparing the Two Intervals Figure 12.14 displays the data, the least-squares line, and both intervals. The confidence interval for the mean is the solid red vertical line. The prediction interval for Jacob's individual log income level is the dashed black vertical line. You can see that the prediction interval is much wider and that it matches the vertical spread of entrepreneurs' log incomes about the regression line. ■

Some software packages will graph the intervals for all values of the explanatory variable within the range of the data. With this type of display, we can see the difference between the two types of intervals across the range of x.

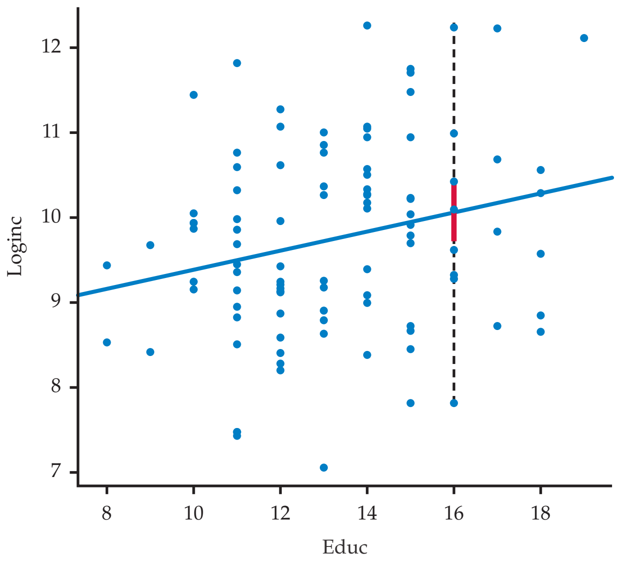


FIGURE 12.14 Confidence interval for mean log income (solid red) and prediction interval for individual log income (dashed black) for an entrepreneur with 16 years of education. Both intervals are centered on the predicted value from the least-squares line, which is $\hat{y} = 10.056$ for $x^* = 16$.

EXAMPLE 12.9

CASE 12.1 **Graphing the Confidence Intervals** The confidence intervals for the log income data are graphed in Figure 12.15. For each value of Educ, we see the estimated value on the solid line and the confidence limits on the dashed curves. ■

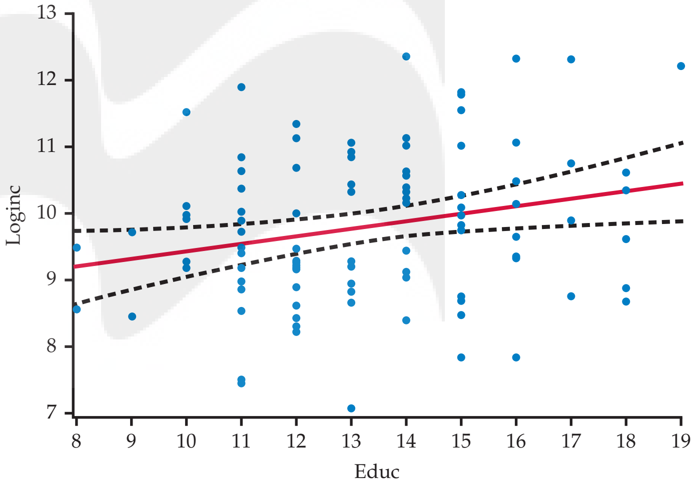


FIGURE 12.15 95% confidence intervals for mean response for the annual income data, for Example 12.9.

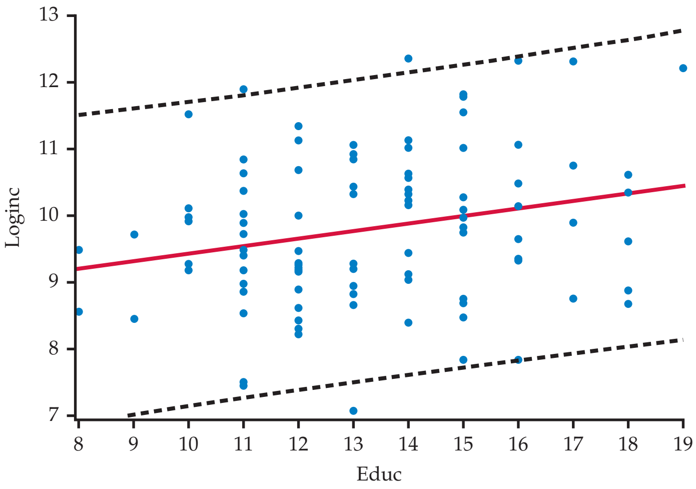
Notice that the intervals get wider as the values of Educ move away from the mean of this variable. This phenomenon reflects the fact that we have less information for estimating means that correspond to extreme values of the explanatory variable.

EXAMPLE 12.10

CASE 12.1 **Graphing the Prediction Intervals** The prediction intervals for the log income data are graphed in Figure 12.16. As with the confidence intervals, we see the predicted values on the solid line and the prediction limits on the dashed curves. ■

It is much easier to see the curvature of the confidence limits in Figure 12.15 than the curvature of the prediction limits in Figure 12.16. One reason

FIGURE 12.16 95% prediction intervals for individual response for the annual income data, for Example 12.10.



for this is that the prediction intervals in Figure 12.16 are dominated by the entrepreneur-to-entrepreneur variation. On the one hand, because the prediction intervals are concerned with individual predictions, they contain a very large proportion of the observations. On the other hand, the confidence intervals are designed to contain mean values and are not concerned with individual observations.

back-transform

Because we transformed income to better fit the model conditions for simple linear regression, our predictions are on the log scale (in log dollars). To get a prediction on the original scale (dollars), we can **back-transform** the predictions from our linear model. The inverse of the logarithm function is the exponential function. Thus, we could naively estimate the average income by exponentiating the log dollars estimate and exponentiating the endpoints of the confidence interval to get an approximate 95% confidence interval for the average income. For example, the average income for entrepreneurs with $x = 16$ years of education is $e^{10.0560} = \$23,295$ and the 95% confidence interval is ($16,698, $32,500).

logNormal distribution

These calculations, however, provide an estimate and confidence interval for the median income rather than the mean. In fact, the regression model implies that income has a **logNormal distribution.** An estimate and confidence interval for the mean income can still be constructed from the regression model estimates but the calculations are more complicated. We suggest seeking expert advice when a situation like this arises.

APPLY YOUR KNOWLEDGE

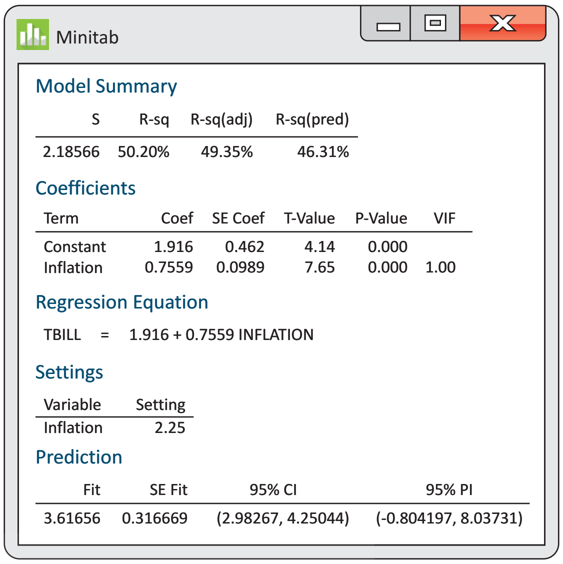
12.41 Predicting the mean Loginc. In Example 12.7, software predicts the mean log income of entrepreneurs with 16 years of education to be $\hat{y} = 10.0560$. We also see that the standard error of this estimated mean is $\text{SE}_{\hat{\mu}} = 0.167802$. These results come from data on 100 entrepreneurs.

(a) Use these facts and $t^* = 1.984$ to verify by hand Minitab's 95% confidence interval for the mean log income when Educ = 16.

(b) Use the same information to construct a 90% confidence interval for the mean log income when Educ = 16. Make sure to specify what degrees of freedom are used to obtain t^*.

12.42 Predicting the return on Treasury bills. Table 12.1 (page 585) gives data on the rate of inflation and the percent return on Treasury bills for 60 years. Figures 12.9 and 12.10 analyze these data. You think that next year's inflation rate will be 2.25%. Figure 12.17 displays part of the

FIGURE 12.17 Minitab output for the regression of the percent return on Treasury bills against the rate of inflation in the same year, for Exercise 12.42. The output includes predictions of the T-bill return when the inflation rate is 2.25%.



```
Model Summary

      S       R-sq   R-sq(adj)   R-sq(pred)
2.18566     50.20%      49.35%       46.31%

Coefficients

Term          Coef   SE Coef   T-Value   P-Value    VIF
Constant     1.916     0.462      4.14     0.000
Inflation   0.7559    0.0989      7.65     0.000   1.00

Regression Equation

TBILL   =   1.916 + 0.7559 INFLATION

Settings

Variable    Setting
Inflation      2.25

Prediction

    Fit     SE Fit              95% CI                     95% PI
3.61656   0.316669   (2.98267, 4.25044)   (-0.804197, 8.03731)
```

Minitab regression output, including predicted values for $x^* = 2.25$. The basic output agrees with the Excel results in Figure 12.10.

(a) Verify the predicted value $\hat{y} = 3.617$ from the equation of the least-squares line.

(b) What is your 95% interval for predicting next year's return on Treasury bills?

BEYOND THE BASICS

Nonlinear regression

nonlinear models

The simple linear regression model assumes that the relationship between the response variable and the explanatory variable can be summarized with a straight line. When the relationship is not linear, we can sometimes transform one or both of the variables so that the relationship becomes linear. Case 12.1 is an example in which the relationship of $\log y$ with x is linear. In other circumstances, we use **nonlinear models** that directly express a curved relationship using parameters that are not just intercepts and slopes.

Here is a typical example of a model that involves parameters β_0 and β_1 in a nonlinear way:

$$y_i = \beta_0 x_i^{\beta_1} + \varepsilon_i$$

This nonlinear model still has the form

$$\text{DATA} = \text{FIT} + \text{RESIDUAL}$$

The FIT term describes how the mean response μ_y depends on x. Figure 12.18 shows the form of the mean response for several values of β_1 when $\beta_0 = 1$. Choosing $\beta_1 = 1$ produces a straight line, but other values of β_1 result in a variety of curved relationships.

We cannot write simple formulas for the estimates of the parameters β_0 and β_1, but software can calculate both estimates and approximate standard errors for the estimates. If the deviations ε_i follow a Normal distribution, we can do inference both on the model parameters and for prediction. The details become more complex, but the ideas remain the same as those we have studied.

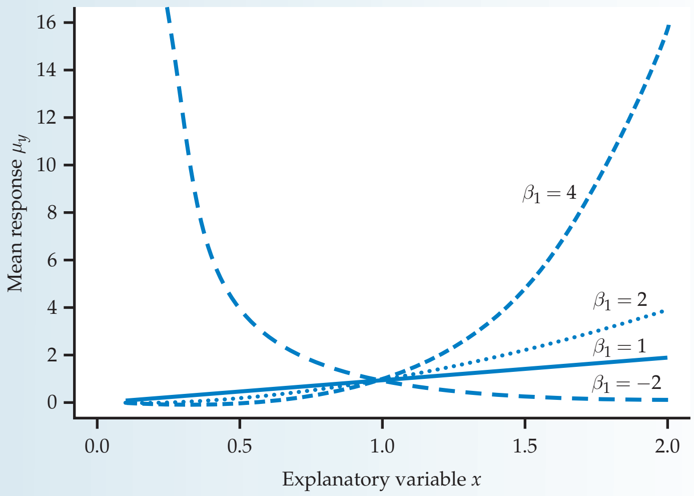


FIGURE 12.18 The nonlinear model $\mu_y = \beta_0 x^{\beta_1}$ includes these and other relationships between the explanatory variable x and the mean response.

SECTION 12.2 SUMMARY

- The **estimated mean response** for the subpopulation corresponding to the value x^* of the explanatory variable is found by substituting $x = x^*$ in the equation of the least-squares regression line:

$$\text{estimated mean response} = \hat{y} = b_0 + b_1 x^*$$

- The **predicted value of the response** y for a single observation from the subpopulation corresponding to the value x^* of the explanatory variable is found in exactly the same way:

$$\text{predicted individual response} = \hat{y} = b_0 + b_1 x^*$$

- **Confidence intervals for the mean response** μ_y when x has the value x^* have the form

$$\hat{y} \pm t^* \text{SE}_{\hat{\mu}}$$

- **Prediction intervals** for an individual response y have a similar form with a larger standard error:

$$\hat{y} \pm t^* \text{SE}_{\hat{y}}$$

In both cases, t^* is the value for the $t(n - 2)$ density curve with area C between $-t^*$ and t^*. Software often gives these intervals. The standard error $\text{SE}_{\hat{y}}$ for an individual response is larger than the standard error $\text{SE}_{\hat{\mu}}$ for a mean response because it must account for the variation of individual responses around their mean.

SECTION 12.2 EXERCISES

For Exercises 12.41 and 12.42, see pages 600–601.

Many of the following exercises require use of software that will calculate the intervals required for predicting mean response and individual response.

12.43 More on public university tuition. Refer to Exercises 12.19 and 12.20 (page 592). TUIT

(a) The tuition at CashCow U was \$9200 in 2013. Find the 95% prediction interval for its tuition in 2017.

(b) The tuition at Moneypit U was \$18,895 in 2013. Find the 95% prediction interval for its tuition in 2017.

(c) Compare the widths of these two intervals. Which is wider and why?

12.44 More on assessment value versus sales price. Refer to Exercise 12.17 (page 591). Suppose we're interested in determining whether the population regression line differs from $y = x$. We'll look at this three ways. HSALES

(a) Construct a 95% confidence interval for each property in the data set. If the model $y = x$ is reasonable, then the assessed value used to predict the sales price should be in the interval. Is this true for all (x, y) pairs?

(b) The model $y = x$ means $\beta_0 = 0$ and $\beta_1 = 1$. Test each of these hypotheses. Is there enough evidence to reject either of them?

(c) Recall that not rejecting H_0 does not imply H_0 is true. A test of "equivalence" would be a more appropriate method to assess similarity. Suppose that, for the slope, a difference within ±0.05% is considered not different. Construct a 90% confidence interval for the slope and see if it falls entirely within the interval (0.95, 1.05). If it does, we would conclude that the slope is not different from 1. What is your conclusion using this method?

12.45 Predicting 2017 tuition from 2013 tuition. Refer to Exercise 12.19 (page 592). TUIT

(a) Find a 95% confidence interval for the mean tuition amount corresponding to a 2013 tuition of $10,403.

(b) Find a 95% prediction interval for a future response corresponding to a 2013 tuition of $10,403.

(c) Write a short paragraph interpreting the meaning of the intervals in terms of public universities.

(d) Do you think that these results can be applied to private universities? Explain why or why not.

12.46 Predicting 2017 tuition from 2008 tuition. Refer to Exercise 12.25 (page 593). TUIT

(a) Find a 95% confidence interval for the mean tuition amount corresponding to a 2008 tuition of $7568.

(b) Find a 95% prediction interval for a future response corresponding to a 2008 tuition of $7568.

(c) Write a short paragraph interpreting the meaning of the intervals in terms of public universities.

(d) Do you think that these results can be applied to private universities? Explain why or why not.

12.47 Compare the estimates. Case 20 in Table 12.3 (Wisconsin) has a 2008 tuition of $7568 and a 2013 tuition of $10,403. A predicted 2017 tuition amount based on 2013 tuition was computed in Exercise 12.45, while one based on the 2008 tuition was computed in Exercise 12.46. Compare these two estimates and explain why they differ. Use the idea of a prediction interval to interpret these results.

12.48 Is the price right? Refer to Exercise 12.35 (page 595), where the relationship between the size of a home and its selling price is examined. HSIZE

(a) Suppose that you have a client who is thinking about purchasing a home in this area that is 1750 square feet in size. The asking price is $180,000. What advice would you give this client?

(b) Answer the same question for a client who is looking at a 1300-square-foot home that is selling for $110,000.

12.49 Predicting income from age. Figures 12.12 and 12.13 (pages 593 and 594) analyze data on the age and income of 5712 men between the ages of 25 and 65. Here is Minitab output predicting the income for ages 30, 40, 50, and 60 years:

Prediction

Fit	SE Fit	95% CI	95% PI
51638	948	(49780, 53496)	(−41735, 145010)
60559	637	(59311, 61807)	(−32803, 153921)
69480	822	(67870, 71091)	(−23888, 162848)
78401	1307	(75840, 80963)	(−14988, 171790)

(a) Use the regression line from Figure 12.12 to verify the "Fit" for age 30 years.

(b) Report the 95% confidence interval for the income of all 30-year-old men.

(c) Joseph is 30 years old. You don't know his income, so give a 95% prediction interval based on his age alone. How useful do you think this interval is?

12.50 Predict what? The two 95% intervals for the income of 30-year-olds given in Exercise 12.49 are very different. Explain briefly to someone who knows no statistics why the second interval is so much wider than the first. Start by looking at 30-year-olds in Figure 12.12.

12.51 Predicting income from age, continued. Use the computer outputs in Figure 12.13 and Exercise 12.49 to give a 90% confidence interval for the mean income of all 40-year-old men.

12.52 T-bills and inflation. Figure 12.17 (page 601) gives part of a regression analysis of the data in Table 12.1 relating the return on Treasury bills to the rate of inflation. The output includes prediction of the T-bill return when the inflation rate is 2.25%.

(a) Use the output to give a 90% confidence interval for the mean return on T-bills in all years having 2.25% inflation.

(b) You think that next year's inflation rate will be 2.25%. It isn't possible, without complicated arithmetic, to give a 90% prediction interval for next year's T-bill return based on the output displayed. Why not?

12.53 Two confidence intervals. The data used for Exercise 12.49 include 195 men who are 30 years old. The mean income of these men is $\bar{y} = \$49,880$ and the standard deviation of these 195 incomes is $s_y = \$38,250$.

(a) Use the one-sample t procedure to give a 95% confidence interval for the mean income μ_y of 30-year-old men.

(b) Why is this interval different from the 95% confidence interval for μ_y in the regression output? (*Hint:* Which data are used by each method?)

12.54 Size and selling price of houses. Table 12.5 (page 595) gives data on the size in square feet of a random sample of houses sold in a Midwest county along with their selling prices. HSIZE

(a) Find the mean size $\bar{x}$ of these houses and also their mean selling price $\bar{y}$. Give the equation of the least-squares regression line for predicting price from size, and use it to predict the selling price of a house of mean size. (You knew the answer, right?)

(b) Zoey and Aiden are selling a house in this Midwest county whose size is equal to the mean of this sample. Give an interval that predicts the price they will receive with 95% confidence.

(c) Compare the prediction interval you used in part (b) to the prediction interval for a new observation from a $N(\mu, \sigma)$ population described in Chapter 7 (page 385).

12.3 Some Details of Regression Inference

When you complete this section, you will be able to

- Use ANOVA table output to perform the ANOVA F test and draw appropriate conclusions regarding H_0: $\beta_1 = 0$.
- Use ANOVA table output to compute the square of the sample correlation and provide an interpretation of it in terms of explained variation.
- Perform, using a calculator or spreadsheet, inference in simple linear regression when software is not available.
- Distinguish the formulas for the standard error that we use for a confidence interval for the mean response and the standard error that we use for a prediction interval when $x = x^*$.

We have assumed that you will use software to handle regression in practice. If you do, it is much more important to understand what the standard error of the slope SE_{b_1} means than it is to know the formula your software uses to find its numerical value. For that reason, we have not yet given formulas for the standard errors. We have also not explained the block of output from software that is labeled "ANOVA" or "Analysis of Variance." This section addresses both of these omissions.

Standard errors

In this section, we give the formulas for all the standard errors we have encountered, for two reasons. First, you may want to see how these formulas can be obtained from facts you already know. The second reason is more practical: some software (in particular, spreadsheet programs) does not automate inference for prediction. Fortunately, almost all software does the hard work of calculating the regression standard error s. With s in hand, the rest is straightforward—but only if you know the details.

Tests and confidence intervals for the slope of a population regression line start with the slope b_1 of the least-squares line and with its standard error SE_{b_1}. If you are willing to skip some messy algebra, it is easy to see where SE_{b_1} and the similar standard error SE_{b_0} of the intercept come from.

1. The regression model takes the explanatory values x_i to be fixed numbers and the response values y_i to be independent random variables all having the same standard deviation σ.

2. The least-squares slope is $b_1 = rs_y/s_x$. Here is the first bit of messy algebra that we skip: it is possible to write the slope b_1 as a linear function of the responses, $b_1 = \sum a_i y_i$. The coefficients a_i depend on the x_i, so they are fixed numbers, too.

rules for variances, pp. 240–241

3. We can find the variance of b_1 by applying the rule for the variance of a sum of independent random variables; it is just $\sigma^2 \sum a_i^2$. A second piece of messy algebra shows that this simplifies to

$$\sigma_{b_1}^2 = \frac{\sigma^2}{\sum (x_i - \bar{x})^2}$$

The standard deviation σ about the population regression line is, of course, not known. If we estimate it by the regression standard error s based on the residuals from the least-squares line, we get the standard error of b_1. Here are the results for both the slope and the intercept.

STANDARD ERRORS FOR SLOPE AND INTERCEPT

The standard error of the slope b_1 of the least-squares regression line is

$$\text{SE}_{b_1} = \frac{s}{\sqrt{\sum(x_i - \bar{x})^2}}$$

The standard error of the intercept b_0 is

$$\text{SE}_{b_0} = s\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum(x_i - \bar{x})^2}}$$

The critical fact is that both standard errors are multiples of the regression standard error s. In a similar manner, accepting the results of yet more messy algebra, we get the standard errors for the two uses of the regression line that we have studied.

STANDARD ERRORS FOR TWO USES OF THE REGRESSION LINE

The standard error for estimating the mean response when the explanatory variable x takes the value x^* is

$$\text{SE}_{\hat{\mu}} = s\sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum(x_i - \bar{x})^2}}$$

The standard error for predicting an individual response when $x = x^*$ is

$$\text{SE}_{\hat{y}} = s\sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum(x_i - \bar{x})^2}}$$

$$= \sqrt{\text{SE}_{\hat{\mu}}^2 + s^2}$$

Once again, both standard errors are multiples of s. The only difference between the two prediction standard errors is the extra 1 under the square root sign in the standard error for predicting an individual response. This added term reflects the additional variation in individual responses, just as it did in Section 7.4 (page 385) when we considered predicting a new observation from a $N(\mu, \sigma)$ population. It also implies that $\text{SE}_{\hat{y}}$ is always greater than $\text{SE}_{\hat{\mu}}$.

EXAMPLE 12.11

Prediction Intervals from a Spreadsheet In Example 12.7, we used statistical software to predict the log income of Jacob, who has Educ = 16 years of education. Suppose that we have only the Excel spreadsheet. The prediction interval then requires some additional work.

Step 1. From the Excel output in Figure 12.5 (page 579), we know that $s = 1.1146$. Excel can also find the mean and variance of the Educ x for the 100 entrepreneurs. They are $\bar{x} = 13.28$ and $s_x^2 = 5.901$.

Step 2. We need the value of $\sum(x_i - \bar{x})^2$. Recalling the definition of the variance, we see that this is just

$$\begin{aligned}\sum(x_i - \bar{x})^2 &= (n-1)s_x^2\\ &= (99)(5.901) = 584.2\end{aligned}$$

Step 3. The standard error for predicting Jacob's log income from his years of education, $x^* = 16$, is

$$\begin{aligned}\text{SE}_{\hat{y}} &= s\sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum(x_i - \bar{x})^2}}\\ &= 1.1146\sqrt{1 + \frac{1}{100} + \frac{(16 - 13.28)^2}{584.2}}\\ &= 1.1146\sqrt{1 + \frac{1}{100} + \frac{7.3984}{584.2}}\\ &= (1.1146)(1.01127) = 1.12716\end{aligned}$$

Step 4. We predict Jacob's log income from the least-squares line (Figure 12.5 again):

$$\hat{y} = 8.2546 + (0.1126)(16) = 10.0562$$

This agrees with the "Fit" from software in Example 12.7. The 95% prediction interval requires the 95% critical value for $t(98)$. Using Excel, the function $=$ T.INV(0.975, 98) gives $t^* = 1.984$. The interval is

$$\begin{aligned}\hat{y} \pm t^*\text{SE}_{\hat{y}} &= 10.0562 \pm (1.984)(1.12716)\\ &= 10.0562 \pm 2.2363\\ &= 7.8199 \text{ to } 12.2925\end{aligned}$$

This agrees with the software result in Example 12.7, with a small difference due to roundoff. ■

The formulas for the standard errors of mean estimation and predicting an individual response show us one more thing about prediction. They both contain the term $(x^* - \bar{x})^2$, the squared distance of the value x^* for which we want to do prediction from the mean $\bar{x}$ of the x-values in our data. We see that prediction is most accurate (smallest margin of error) at the mean and grows less accurate as we move away from the mean of the explanatory variable. *If you know the values of x for which you want to do prediction, try to collect data centered near these values.*

APPLY YOUR KNOWLEDGE

12.55 T-bills and inflation. Figure 12.10 (page 586) gives the Excel output for regressing the annual return on Treasury bills on the annual rate of inflation. The data appear in Table 12.1 (page 585). Starting with the regression standard error $s = 2.1857$ from the output and the variance of the inflation rates in Table 12.1 (use your calculator), find the standard error of the regression slope SE_{b_1}. Check your result against the Excel output. INFLAT

12.56 Predicting T-bill return. Figure 12.17 (page 601) uses statistical software to predict the return on Treasury bills in a year when the inflation rate is 2.25%. Let's do this calculation without specialized software. Figure 12.10 contains Excel regression output. Use a calculator or software to find the variance s_x^2 of the annual inflation rates in Table 12.1 (page 585). From this information, find the 95% prediction interval for the T-bill return. Check your result against the software output in Figure 12.17. INFLAT

Analysis of variance for regression

sums of squares, p. 471

Software output for regression problems, such as those in Figures 12.5, 12.6, and 12.10, reports values under the heading of "ANOVA" or "Analysis of Variance." Analysis of variance (ANOVA) is the term for statistical analyses that break down the variation in data into separate pieces that correspond to different sources of variation. We used it in Chapter 9 to compare several population means by breaking down the total variation, expressed by sums of squares, into the variation among groups and the variation within groups. In the regression setting, the observed variation in the responses y_i also comes from two sources:

- As the explanatory variable x moves, it pulls the response with it along the regression line. In Figure 12.4, for example, entrepreneurs with 15 years of education generally have higher log incomes than those entrepreneurs with 9 years of education. The least-squares line drawn on the scatterplot describes this tie between x and y.
- When x is held fixed, y still varies because not all individuals who share a common x have the same response y. There are several entrepreneurs with 11 years of education, and their log income values are scattered above and below the least-squares line.

squared correlation r^2, p. 88

analysis of variance equation

We discussed these sources of variation in Chapter 2, where the main point was that the squared correlation r^2 is the proportion of the total variation in the responses that comes from the first source, the straight-line tie between x and y. Analysis of variance for regression expresses these two sources of variation in algebraic form so that we can calculate the breakdown of overall variation into two parts. Skipping quite a bit of messy algebra, the **analysis of variance equation** that always holds is

$$\begin{array}{ccccc} \text{total variation in } y & = & \text{variation along the line} & + & \text{variation about the line} \\ \text{SST} & = & \text{SSR} & + & \text{SSE} \\ \sum(y_i - \bar{y})^2 & = & \sum(\hat{y}_i - \bar{y})^2 & + & \sum(y_i - \hat{y}_i)^2 \end{array}$$

ANOVA table, p. 475

This breakdown is commonly summarized in the form of an ANOVA table:

Source	Degrees of freedom	Sum of squares	Mean square	*F*
Regression	DFR = 1	SSR = $\sum(\hat{y}_i - \bar{y})^2$	MSR = SSR/DFR	MSR/MSE
Residual	DFE = $n - 2$	SSE = $\sum(y_i - \hat{y}_i)^2$	MSE = SSE/DFE	
Total	DFT = $n - 1$	SST = $\sum(y_i - \bar{y})^2$		

The "total variation in y," which we label SST, is expressed by the sum of the squares of the deviations $y_i - \bar{y}$. Similar to Chapter 9, it is just $n - 1$ times the variance of the responses. The "variation along the line," labeled SSR, has the same form but is the variation among the *predicted* responses $\hat{y}_i$. The predicted responses lie on the least-squares regression line—they show how y moves in response to x. The more y moves in response to x, the larger this term will be. The "variation about the line," labeled SSE, is the sum of squares of the *residuals* $y_i - \hat{y}_i$. It measures the size of the scatter of the observed responses above and below the line. If all the responses fell exactly on a straight line, the residuals would all be 0. In such a case, there would be no variation about the line (SSE = 0) and the total variation would equal the variation along the line (SST = SSR). The other extreme is when $b_1 = 0$. In that case, $\hat{y}_i = \bar{y}$ so SSR = 0 and SST = SSE.

EXAMPLE 12.12

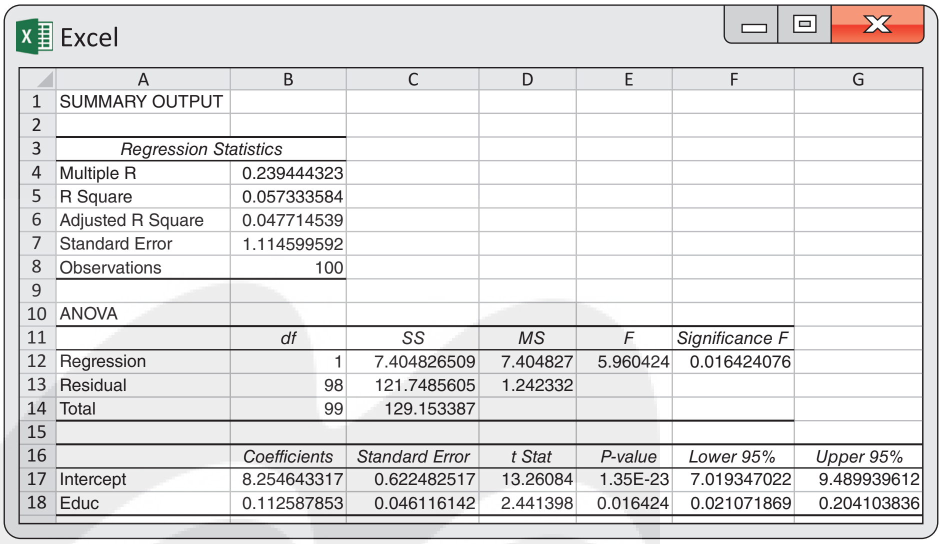
Case 12.1

ENTRE

ANOVA for Entrepreneur Income Study Figure 12.19 repeats Figure 12.5; it shows the Excel output for the regression of log income on years of education (Case 12.1). The three terms in the analysis of variance equation appear under the "SS" heading, reflecting the fact that each of the three terms is a sum of squared quantities. You can read the output as follows:

$$\begin{array}{ccccc} \text{SST} & = & \text{SSR} & + & \text{SSE} \\ 129.1534 & = & 7.4048 & + & 121.7486 \end{array}$$

FIGURE 12.19 Excel output for the regression of log annual income on years of education, for Examples 12.12 and 12.13. We now concentrate on the analysis of variance part of the output.



	A	B	C	D	E	F	G
1	SUMMARY OUTPUT						
2							
3	*Regression Statistics*						
4	Multiple R	0.239444323					
5	R Square	0.057333584					
6	Adjusted R Square	0.047714539					
7	Standard Error	1.114599592					
8	Observations	100					
9							
10	ANOVA						
11		*df*	*SS*	*MS*	*F*	*Significance F*	
12	Regression	1	7.404826509	7.404827	5.960424	0.016424076	
13	Residual	98	121.7485605	1.242332			
14	Total	99	129.153387				
15							
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
17	Intercept	8.254643317	0.622482517	13.26084	1.35E-23	7.019347022	9.489939612
18	Educ	0.112587853	0.046116142	2.441398	0.016424	0.021071869	0.204103836

Excel uses the names "Total," "Regression," and "Residual" for these three sources of variation. Other software (see Figure 12.6, page 579) may use other row names.

The proportion of variation in log incomes explained by regressing on years of education is

$$\begin{aligned} r^2 &= \frac{\text{SSR}}{\text{SST}} \\ &= \frac{7.4048}{129.1534} = 0.0573 \end{aligned}$$

This agrees with the "R Square" value in the output. Only about 6% of the variation in log incomes is explained by the linear relationship between log income and years of education. The rest is variation in log incomes among entrepreneurs with the same level of education. ■

degrees of freedom, p. 33

The remaining columns of the ANOVA table are similar to those described in Chapter 9. For the degrees of freedom column, the total degrees of freedom (DFT) are $n - 1 = 99$ (the degrees of freedom for the variance of $n = 100$ observations). We know that the degrees of freedom for the residuals (DFE) and for t statistics in simple linear regression are $n - 2$. Therefore, it is no surprise that the degrees of freedom for the residual sum of squares are also $n - 2 = 98$. That leaves just 1 degree of freedom for regression (DFR), because degrees of freedom in ANOVA are added as follows:

$$\begin{array}{ccccc} \text{DFT} & = & \text{DFR} & + & \text{DFE} \\ n-1 & = & 1 & + & n-2 \end{array}$$

mean squares, p. 472

The next column reports the mean squares, obtained by dividing each sum of squares by its degrees of freedom. The total mean square (not given in the output) is just the variance of the responses y_i. The residual mean square (MSE) is the square of our old friend, the regression standard error:

$$\begin{aligned}\text{MSE} &= \frac{\text{SSE}}{\text{DFE}}\\ &= \frac{\sum(y_i - \hat{y}_i)^2}{n-2}\\ &= s^2\end{aligned}$$

This is why the JMP output in Figure 12.6 labels s as the "Root Mean Square Error."

The next column is named "F" and is the ratio of the two calculated mean squares:

ANOVA F statistic, p. 473

$$F = \frac{\text{MSR}}{\text{MSE}}$$

F distribution, p. 473

This ANOVA F statistic provides a different way to test for the overall significance of the regression. Its P-value is displayed in the last column named "Significance F." This is computed from an F distribution with 1 and $n - 2 = 98$ degrees of freedom. Table E in the back of the book contains the F critical values to compare the ANOVA F statistic against.

Recall that if regression on x has no value for predicting y, we expect the slope of the population regression line to be close to zero. That is, the null hypothesis of "no linear relationship" is H_0: $\beta_1 = 0$. To test H_0, we standardize the slope of the least-squares line to get a t statistic. The ANOVA approach starts instead with sums of squares. If regression on x has no value for predicting y, we expect the SSR to be only a small part of the SST, most of which will be made up of the SSE. That, in turn, means we expect F to be small.

For simple linear regression, the ANOVA F statistic always equals the square of the t statistic for testing H_0: $\beta_1 = 0$. That is, the two tests amount to the same thing. Let's verify that relationship using Case 12.1.

EXAMPLE 12.13

CASE 12.1 **ANOVA for Entrepreneur Income Study, Continued** The Excel output in Figure 12.19 contains the values for the analysis of variance equation for sums of squares and also the corresponding degrees of freedom. The residual mean square is

$$\begin{aligned}\text{MSE} &= \frac{\text{SSE}}{\text{DFE}}\\ &= \frac{121.7486}{98} = 1.2423\end{aligned}$$

The square root of the residual MS is $\sqrt{1.2423} = 1.1146$. This is the regression standard error s, as claimed. The ANOVA F statistic is

$$\begin{aligned}F &= \frac{\text{MSR}}{\text{MSE}}\\ &= \frac{7.4048}{1.2423} = 5.9604\end{aligned}$$

The square root of F is $\sqrt{5.9604} = 2.441$. Sure enough, this is the value of the t statistic for testing the significance of the regression, which also appears in the Excel output. The P-value for F, $P = 0.0164$, is the same as the two-sided P-value for t. ■

We have now explained almost all the results that appear in a typical regression output such as Figure 12.19. ANOVA shows exactly what r^2 means in regression. Aside from this, ANOVA seems redundant; it repeats in less clear form information that is found elsewhere in the output. This is true in simple linear regression, but ANOVA comes into its own in *multiple regression,* the topic of the next chapter.

APPLY YOUR KNOWLEDGE

T-bills and inflation. *Figure 12.10 (page 586) gives Excel output for the regression of the rate of return on Treasury bills against the rate of inflation during the same year. Exercises 12.57 through 12.59 use this output.*

12.57 A significant relationship? The output reports *two* tests of the null hypothesis that regressing on inflation does *not* help to explain the return on T-bills. State the hypotheses carefully, give the two test statistics, show how they are related, and give the common P-value.

12.58 The ANOVA table. Use the numerical results in the Excel output to verify each of these relationships.

(a) The ANOVA equation for sums of squares.

(b) How to obtain the total degrees of freedom and the residual degrees of freedom from the number of observations.

(c) How to obtain each mean square from a sum of squares and its degrees of freedom.

(d) How to obtain the F statistic from the mean squares.

12.59 ANOVA by-products.

(a) The output gives $r^2 = 0.5020$. How can you obtain this from the ANOVA table?

(b) The output gives the regression standard error as $s = 2.1857$. How can you obtain this from the ANOVA table?

SECTION 12.3 SUMMARY

- The **analysis of variance (ANOVA) equation** for simple linear regression expresses the total variation in the responses as the sum of two sources: the linear relationship of y with x and the residual variation in responses for the same x. The equation is expressed in terms of **sums of squares.**

- Each sum of squares has a **degrees of freedom.** A sum of squares divided by its degrees of freedom is a **mean square.** The residual mean square is the square of the regression standard error.

- The **ANOVA table** gives the degrees of freedom, sums of squares, and mean squares for total, regression, and residual variation. The **ANOVA *F* statistic** is the ratio $F =$ Regression MS/Residual MS. In simple linear regression, F is the square of the t statistic for the hypothesis that regression on x does not help explain y.

- The **square of the sample correlation** can be expressed as

$$r^2 = \frac{\text{Regression SS}}{\text{Total SS}} = \frac{\text{SSR}}{\text{SST}}$$

and is interpreted as the proportion of the variability in the response variable y that is explained by the explanatory variable x in the linear regression.

SECTION 12.3 EXERCISES

For Exercises 12.55 and 12.56, see page 606; and for 12.57 to 12.59, see page 610.

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

(a) In simple linear regression, the standard error for a future observation is s, the measure of spread about the regression line.

(b) In an ANOVA table, SSE is the sum of the deviations.

(c) There is a close connection between the correlation r and the intercept of the regression line.

(d) The squared correlation r^2 is equal to MSR/MST.

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

(a) In simple linear regression, the null hypothesis of the ANOVA F test is H_0: $\beta_0 = 0$.

(b) In an ANOVA table, the mean squares add; in other words, MST = MSR + MSE.

(c) The smaller the P-value for the ANOVA F test, the greater the explanatory power of the model.

(d) The total degrees of freedom in an ANOVA table are equal to the number of observations n.

U.S. versus overseas stock returns. *How are returns on common stocks in overseas markets related to returns in U.S. markets? Consider measuring U.S. returns by the annual rate of return on the Standard & Poor's 500 stock index and overseas returns by the annual rate of return on the Morgan Stanley Europe, Australasia, Far East (EAFE) index. Both are recorded in percents. Here is part of the Minitab output for regressing the EAFE returns on the S&P 500 returns for the 29 years 1989 to 2017.*

The regression equation is
Eafe = −3.11 + 0.820 S&P

Analysis of Variance

Source	DF	SS	MS	F
Regression		5821.3		
Residual Error				
Total	28	10454.7		

Exercises 12.62 through 12.66 use this output. EAFE

12.62 The ANOVA table. Complete the analysis of variance table by filling in the "Residual Error" row and the other missing items in the DF, MS, and F columns.

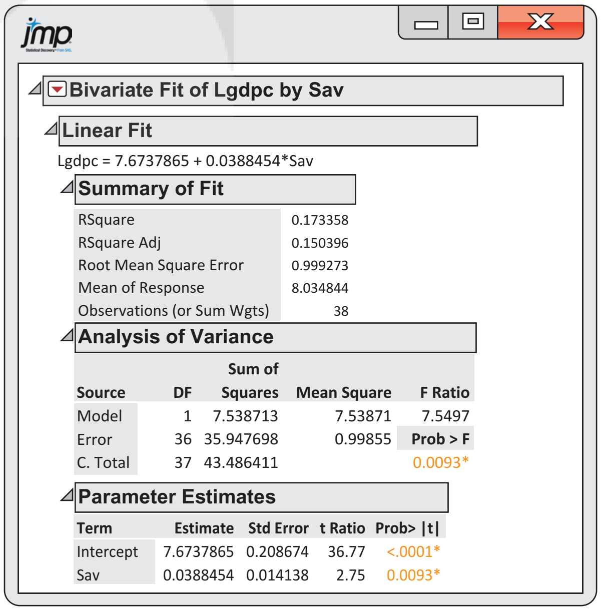
12.63 s and r^2. What are the values of the regression standard error s and the squared correlation r^2?

12.64 Estimating the standard error of the slope. The standard deviation of the S&P 500 returns for these years is 17.58%. From this and your work in the previous exercise, find the standard error for the least-squares slope b_1. Give a 90% confidence interval for the slope β_1 of the population regression line.

12.65 Inference for the intercept? The mean of the S&P 500 returns for these years is 12.01. From this and information from the previous exercises, find the standard error for the least-squares intercept b_0. Use this to construct a 95% confidence interval. Finally, explain why the intercept β_0 is meaningful in this example.

12.66 Predicting the return for a future year. Suppose the S&P annual return for a future year is 0%. Using the information from the previous four exercises, construct the appropriate 95% interval. Also, explain why this interval is or is not the same interval constructed in Exercise 12.65.

Gross domestic product per capita and net savings. *The gross domestic product (GDP) measures the aggregate amount of good and services produced in an economy. Growing GDP is a primary focus of policymakers. A random sample of 38 emerging economies (countries) was taken to assess whether there was a positive linear relationship between GDP per capita and a country's adjusted net savings.*[17] *Adjusted net savings is defined as a country's net savings plus expenditures on education minus depletion of a country's air, minerals, and forests. It is reported as a percent of the gross national income. Figure 12.20 contains JMP output for the regression of the logarithm of GDP per capita (Lgdpc) on adjusted net savings (Sav). Exercises 12.67 through 12.74 concern this analysis. You can take it as given that an examination of the data shows no serious violations of the conditions required for regression inference.*



Bivariate Fit of Lgdpc by Sav

Linear Fit

Lgdpc = 7.6737865 + 0.0388454*Sav

Summary of Fit

RSquare	0.173358
RSquare Adj	0.150396
Root Mean Square Error	0.999273
Mean of Response	8.034844
Observations (or Sum Wgts)	38

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	7.538713	7.53871	7.5497
Error	36	35.947698	0.99855	Prob > F
C. Total	37	43.486411		0.0093*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob> |t|
Intercept	7.6737865	0.208674	36.77	<.0001*
Sav	0.0388454	0.014138	2.75	0.0093*

FIGURE 12.20 JMP output for the regression of GDP per capita of 38 emerging economies on the country's adjusted net savings, for Exercises 12.67 to 12.74.

12.67 Significance in two senses.

(a) Provide an explanation for using the logarithm of GDP per capita as the response variable rather than GDP per capita.

(b) Is there good evidence that adjusted net savings helps explain Lgdpc? (State the hypotheses, give a test statistic and *P*-value, and state a conclusion.)

(c) What percent of the variation in Lgdpc among these economies is explained by a regression on adjusted net savings?

(d) Use your findings in parts (a) and (b) as the basis for a short description of the distinction between statistical significance and practical significance.

12.68 Estimating the slope. Explain clearly what the slope β_1 of the population regression line tells us in this setting. Give a 95% confidence interval for this slope.

12.69 Predicting Lgdpc. An additional calculation shows that the variance of the adjusted net savings for these 38 countries is $s_x^2 = 135.03$. JMP labels the regression standard error s as "Root Mean Square Error" and the sample mean of the responses $\bar{y}$ as "Mean of Response." Starting from these facts, give a 95% confidence interval for the mean Lgdpc for all countries with an adjusted net savings $x = 15.0$. [*Hint:* The least-squares regression line always goes through $(\bar{x}, \bar{y})$.]

12.70 Predicting Lgdpc. Will a 95% prediction interval for a country's Lgdpc when $x = 15.0$ be wider or narrower than the confidence interval found in the previous exercise? Explain why we should expect this result. Then give the 95% prediction interval.

12.71 *F* versus *t*. How do the ANOVA *F* statistic and its *P*-value relate to the *t* statistic for the slope and its *P*-value? Identify these results on the output and verify their relationship (up to roundoff error).

12.72 The regression standard error. How can you obtain s from the ANOVA table? Do this, and verify that your result agrees with the value that JMP reports for s.

12.73 Squared correlation. JMP gives the squared correlation r^2 as "RSquare." How can you obtain r^2 from the ANOVA table? Do this, and verify that your result agrees with the RSquare reported by JMP.

12.74 Correlation. The regression in Figure 12.20 takes adjusted net savings as explaining Lgdpc. As an alternative, we could take Lgdpc as explaining adjusted net savings. We would then reverse the roles of the variables, regressing Sav on Lgdpc. Both regressions lead to the same conclusions about the correlation between Lgdpc and Sav. What is this correlation r? Is there good evidence that it is positive?

CHAPTER 12 REVIEW EXERCISES

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

(a) The slope describes the change in x for a unit change in y.

(b) The population regression line is $y = b_0 + b_1 x$.

(c) A 95% confidence interval for the mean response is the same width regardless of x.

(d) The residual for the ith observation is $\hat{y}_i - y_i$

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

(a) The parameters of the simple linear regression model are b_0, b_1, and s.

(b) To test H_0: $b_1 = 0$, you would use a t test.

(c) For any value of the explanatory variable x, the confidence interval for the mean response will be wider than the prediction interval for a future observation.

(d) The least-squares line is the line that maximizes the sum of the squares of the residuals.

12.77 Interpreting a residual plot. Figure 12.21 shows four plots of residuals versus x. For each plot, comment on the regression model conditions necessary for inference. Which plots suggest a reasonable fit to the linear regression model? What actions might you take to remedy the problems in each of the other plots?

12.78 Are the results consistent? A researcher surveyed $n = 214$ hotel managers to assess the relationship between customer-relationship management (CRM) and organizational culture.[18] Each variable was an average of more than twenty-five 5-point Likert survey responses and, therefore, was treated as a quantitative variable. The researcher reports a sample correlation of $r = 0.74$ and an ANOVA *F* statistic of 60.35 for a simple linear regression of CRM on organizational culture. Using the relationship between testing H_0: $\rho = 0$ and testing H_0: $\beta_1 = 0$, show that these two results are not consistent. (*Hint:* It's far more likely that there was a typo and $r = 0.47$.)

12.79 College debt versus adjusted in-state costs. Kiplinger's "Best Values in Public Colleges" provides a ranking of U.S. public colleges based on a combination of various measures of academics and affordability.[19] We'll consider a random collection of 40 colleges from Kiplinger's 2018 report and focus on the average debt in dollars at graduation (AveDebt) and the in-state cost per year after need-based aid (InCostAid). BESTVAL

(a) A scatterplot of these two variables with a smoothed curve is shown in Figure 12.22. Describe the relationship. Are there any possible outliers or unusual values? Does a linear relationship between InCostAid and AveDebt seem reasonable?

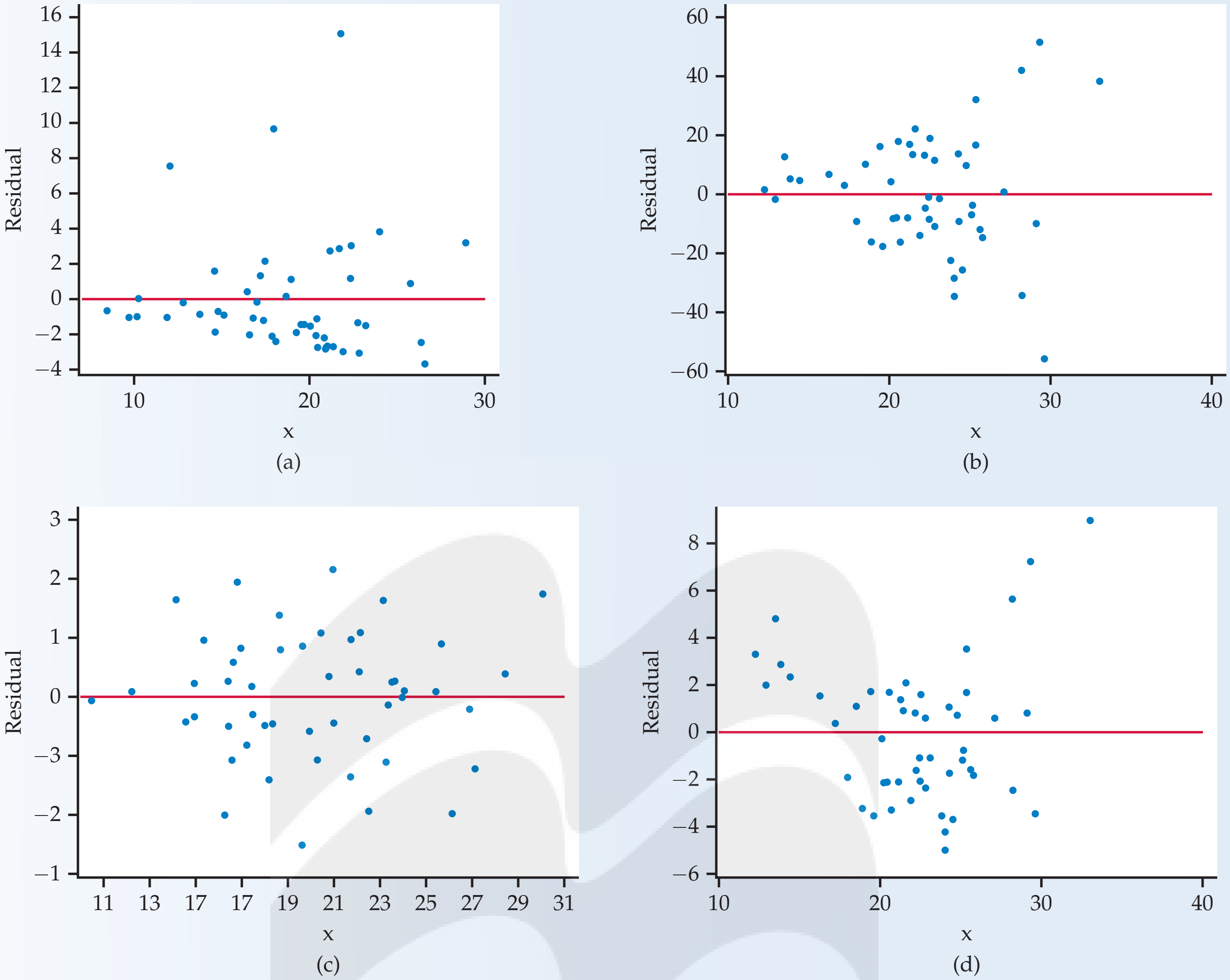


FIGURE 12.21 Four plots of regression residuals versus explanatory variable *x*, for Exercise 12.77.

(b) Based on the scatterplot, approximately how much does the average debt change for an additional $1000 of annual cost?

(c) The University of North Carolina at Chapel Hill is a school with an adjusted in-state cost of $4843. Discuss the appropriateness of using this data set to predict the average debt at graduation for students attending this school.

12.80 Can we consider this an SRS? Refer to the previous exercise. The report states that Kiplinger's rankings focus on traditional four-year public colleges with broad-based curricula and on-campus housing. Each year, the researchers start with more than 500 schools and then narrow the list down to roughly 120 based on academic quality before ranking them. The data set in the previous exercise is an SRS from Kiplinger's published list of 100 schools. When investigating the relationship between the average debt and the in-state

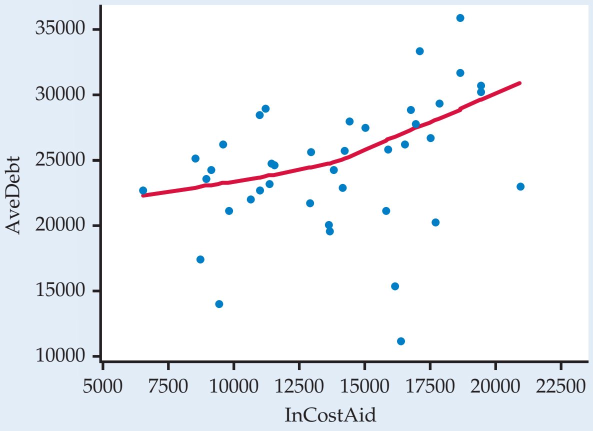


FIGURE 12.22 Scatterplot of average debt (in dollars) at graduation versus the in-state cost per year (in dollars) after need-based aid, for Exercise 12.79.

cost after adjusting for need-based aid, is it reasonable to consider this to be an SRS from the population of interest? Write a short paragraph explaining your answer.

12.81 Predicting college debt. Refer to Exercise 12.79. Figure 12.23 contains JMP output for the simple linear regression of AveDebt on InCostAid. BESTVAL

(a) State the least-squares regression line.

(b) The University of California at Irvine is one school in this sample. It has an in-state cost of $9621 and an average debt at graduation of $20,466. What is the residual?

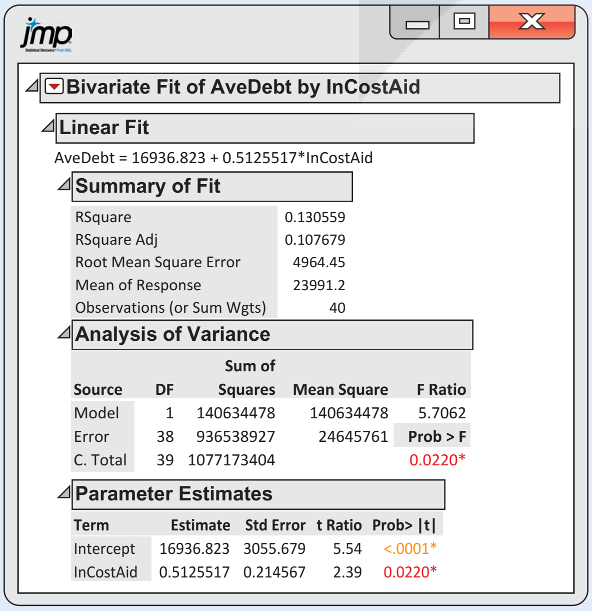
(c) Construct a 95% confidence interval for the slope. What does this interval tell you about the change in average debt for a $500 change in the in-state cost?

12.82 More on predicting college debt. Refer to the previous exercise. Purdue University has an in-state cost of $7788 and an average debt at graduation of $27,530. Texas A&M University has an in-state cost of $11,396 and an average debt at graduation of $24,072. BESTVAL

(a) Using your answer to part (a) of the previous exercise, what is the predicted average debt at graduation for a student attending Purdue University?

(b) What is the predicted average debt at graduation for a student attending Texas A&M University?

(c) Without doing any calculations, would the standard error for the estimated average debt be larger for Purdue University or the Texas A&M University? Explain your answer.



Bivariate Fit of AveDebt by InCostAid

Linear Fit

AveDebt = 16936.823 + 0.5125517*InCostAid

Summary of Fit

RSquare	0.130559
RSquare Adj	0.107679
Root Mean Square Error	4964.45
Mean of Response	23991.2
Observations (or Sum Wgts)	40

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	140634478	140634478	5.7062
Error	38	936538927	24645761	Prob > F
C. Total	39	1077173404		0.0220*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob> |t|
Intercept	16936.823	3055.679	5.54	<.0001*
InCostAid	0.5125517	0.214567	2.39	0.0220*

FIGURE 12.23 JMP output for the regression of average debt (in dollars) at graduation on the in-state cost (in dollars) per year, for Exercise 12.81.

12.83 Predicting college debt: Other measures. Refer to Exercise 12.79. Let's now look at AveDebt and its relationship with all six measures available in the data set. In addition to the in-state cost after aid (InCostAid), we have the admittance rate (Admit), the four-year graduation rate (Grad4Rate), the in-state cost before aid (TotCostIn), the out-of-state cost before aid (TotCostOut), and the out-of-state cost after aid (OutCostAid). BESTVAL

(a) Generate scatterplots of each explanatory variable and AveDebt. Do all these relationships look linear? Describe what you see.

(b) Fit each of the explanatory variables separately and create a table that lists the explanatory variable, regression standard error s, and the P-value for the test of a linear association.

(c) Which variable appears to be the best single explanatory variable of average debt at graduation? Explain your answer.

12.84 Yearly number of tornadoes. The Storm Prediction Center of the National Oceanic and Atmospheric Administration maintains a database of tornadoes, floods, and other weather phenomena. Table 12.6 summarizes the annual number of tornadoes in the United States between 1953 and 2017.[20] (Note: These are time series data with a very weak correlation, so simple linear regression is reasonable here. See Chapter 14 for methods designed specifically for use with time series.) TWISTER

(a) Make a plot of the total number of tornadoes by year. Does a linear trend over years appear reasonable? Are there any outliers or unusual patterns? Explain your answer.

(b) Run the simple linear regression and summarize the results, making sure to construct a 95% confidence interval for the average annual increase in the number of tornadoes.

(c) Obtain the residuals and plot them versus year. Is there anything unusual in the plot?

(d) Are the residuals Normal? Justify your answer.

(e) The number of tornadoes in 2004 is much larger than expected under this linear model. Remove this observation and rerun the simple linear regression. Compare these results with the results in part (b). Do you think this observation should be considered an outlier and removed? Explain your answer.

12.85 Plot indicates model assumptions. Construct a plot with data and a regression line that fits the simple linear regression model framework. Then construct another plot that has the same slope and intercept but a much smaller value of the regression standard error s.

12.86 Significance tests and confidence intervals. The significance test for the slope in a simple linear regression gave a value $t = 2.08$ with 18 degrees of freedom. Would the 95% confidence interval for the slope include the value zero? Give a reason for your answer.

TABLE 12.6 Annual number of tornadoes in the United States between 1953 and 2017

Year	Number of tornadoes	Year	Number of tornadoes	Year	Number of tornadoes	Year	Number of tornadoes
1953	422	1970	653	1987	656	2004	1817
1954	550	1971	889	1988	702	2005	1265
1955	593	1972	741	1989	856	2006	1103
1956	504	1973	1102	1990	1133	2007	1096
1957	858	1974	945	1991	1132	2008	1692
1958	564	1975	919	1992	1297	2009	1156
1959	604	1976	834	1993	1173	2010	1282
1960	616	1977	852	1994	1082	2011	1692
1961	697	1978	789	1995	1235	2012	936
1962	657	1979	855	1996	1173	2013	891
1963	463	1980	866	1997	1148	2014	881
1964	704	1981	782	1998	1424	2015	1183
1965	897	1982	1047	1999	1339	2016	985
1966	585	1983	931	2000	1075	2017	1406
1967	926	1984	907	2001	1215		
1968	660	1985	684	2002	934		
1969	608	1986	765	2003	1374		

12.87 Predicting college debt: One last measure. Refer to Exercises 12.79, 12.81, and 12.83. Given the in-state cost of attending a college prior to and after aid, another measure of potential college debt is the average amount of need-based aid. Create this new variable by subtracting these two costs, and investigate its relationship with average debt at graduation. Write a short paragraph summarizing your findings. BESTVAL

12.88 Brand equity and sales. Brand equity is one of the most important assets of a business. It includes brand loyalty, brand awareness, perceived quality, and brand image. One study examined the relationship between brand equity and sales using simple linear regression analysis.[21] The correlation between brand equity and sales was reported to be 0.757, with a significance level of 0.001.

(a) Explain in simple language the meaning of these results.

(b) The study examined quick-service restaurants in Korea and was based on 394 usable surveys from a total of 950 that were distributed to shoppers at a mall. Write a short narrative commenting on the design of the study and how well you think the results would apply to other settings.

12.89 Hotel sizes and numbers of employees. A human resources study of hotels collected data on the hotel size, measured by number of rooms, and the number of employees for 14 hotels in Canada.[22] Here are the data: HOTSIZE

Employees	Rooms	Employees	Rooms
1200	1388	275	424
180	348	105	240
350	294	435	601
250	413	585	1590
415	346	560	380
139	353	166	297
121	191	228	108

(a) To what extent can the number of employees be predicted by the size of the hotel? Plot the data and summarize the relationship.

(b) Is this the type of relationship that you would expect to see before examining the data? Explain why or why not.

(c) Calculate the least-squares regression line and add it to the plot.

(d) Give the results of the significance test for the regression slope with your conclusion.

(e) Find a 95% confidence interval for the slope.

12.90 How can we use the results? Refer to the previous exercise.

(a) If one hotel had 100 more rooms than another hotel, how many additional employees would you expect the first hotel to have? HOTSIZE

(b) Give a 95% confidence interval for your answer in part (a).

(c) The study collected these data from 14 hotels in Toronto. Discuss how well you think the results can be generalized to other hotels in Toronto, to hotels in Canada, and to hotels in other countries.

12.91 Check the outliers. The plot that you generated in Exercise 12.89 has two observations that appear to be outliers. HOTSIZE

(a) Identify these points on a plot of the data.

(b) Rerun the analysis with the other 12 hotels, and summarize the effect of the two possible outliers on the results that you gave in Exercise 12.89.

12.92 Selling a large house. Among the houses for which we have data in Table 12.5 (page 595), just four have floor areas of 1800 square feet or more. Give a 90% confidence interval for the mean selling price of houses with floor areas of 1800 square feet or more. HSIZE

12.93 Agricultural productivity. Few sectors of the economy have increased their productivity as rapidly as agriculture. Let's describe this increase. Productivity is defined as output per unit input. "Total factor productivity" (TFP) takes all inputs (labor, capital, fuels, and so on) into account. The data set AGPROD contains TFP for the years 1948–2015.[23] The TFP entries are index numbers; that is, they give each year's TFP as a percent of the value for 1948. AGPROD

(a) Plot TFP against year. It appears that around 1980, the rate of increase in TFP changed. How is this apparent from the plot? What was the nature of the change?

(b) Regress TFP on year using only the data for the years 1948–1980. Add the least-squares line to your scatterplot. The line makes the finding in part (a) clearer.

(c) Give a 95% confidence interval for the annual rate of change in TFP during the period 1948–1980.

(d) Regress TFP on year for the years 1981–2015. Add this line to your plot. Give a 95% confidence interval for the annual rate of improvement in TFP during these years.

(e) Write a brief report on trends in U.S. farm productivity since 1948, making use of your analysis in parts (a) through (d).

12.94 CEO pay and gross profits. Starting in 2018, publicly traded companies must disclose their workers' median pay and the compensation ratio between a worker and the company's CEO. Does this ratio say something about the performance of the company? CNBC collected this ratio and the gross profits per employee from a variety of companies.[24] CNBC

(a) Generate a scatterplot of the gross profit per employee (Profit) versus the CEO pay ratio (Ratio). Describe the relationship.

(b) To compensate for the severe right skewness of both variables, take the logarithm of each variable. Generate a scatterplot and describe the relationship between these transformed variables.

(c) Fit a simple linear regression for log Profits versus log Ratio.

(d) Examine the residuals. Are the model conditions approximately satisfied? Explain your answer.

(e) Construct a 95% confidence interval for β_1 and interpret the result in terms of a percent change in y for a percent change in x.