Chapter 12

More about Regression

case study

Do Longer Drives Mean Lower Scores on the PGA Tour?

Recent advances in technology have led to golf balls that fly farther, clubs that generate more speed at impact, and swings that have been perfected through computer video analysis. Moreover, today's professional golfers are fitter than ever. The net result is many more players who routinely hit drives traveling 300 yards or more. Does greater distance off the tee translate to better (lower) scores?

We collected data on mean drive distance (in yards) and mean score per round from an SRS of 19 of the 197 players on the Professional Golfers Association (PGA) Tour in a recent year. Figure 12.1 is a scatterplot of the data with results from a least-squares regression analysis added. The graph shows that there is a moderately weak negative linear relationship between mean drive distance and mean score for the 19 players in the sample.[1]

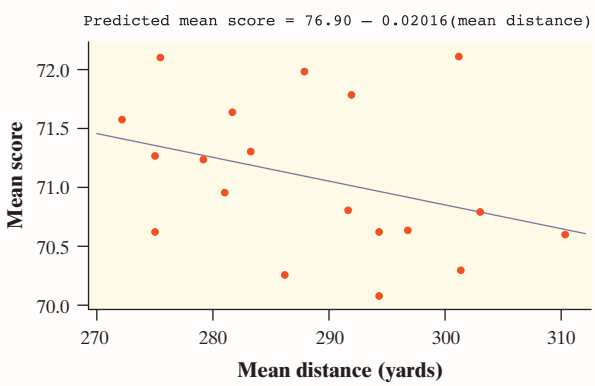


FIGURE 12.1 Scatterplot and least-squares regression line of mean score versus mean drive distance for a random sample of 19 players on the PGA Tour.

The slope of the least-squares regression line for the sample data is about −0.02. This line predicts a 0.02 decrease in mean score per round for each 1-yard increase in mean driving distance. But a slope of −0.02 is very close to 0. Do these data give convincing evidence that the slope of the true regression line for *all* 197 PGA Tour golfers is negative? By the end of this chapter, you'll have developed the tools you need to answer this question.

Introduction

When a scatterplot shows a linear relationship between a quantitative explanatory variable x and a quantitative response variable y, we can use the least-squares line calculated from the data to predict y for a given value of x. If the data are a random sample from a larger population, we need statistical inference to answer questions like these:

- Is there really a linear relationship between x and y in the population, or could the pattern we see in the scatterplot plausibly happen just by chance?
- In the population, how much will the predicted value of y change for each increase of 1 unit in x? What's the margin of error for this estimate?

If the data come from a randomized experiment, the values of the explanatory variable correspond to the levels of some factor that is being manipulated by the researchers. For instance, researchers might want to investigate how temperature affects the life span of mosquitoes. They could set up several tanks at each of several different temperatures and then randomly assign hundreds of mosquitoes to each of the tanks. The response variable of interest is the average time (in days) from hatching to death. Suppose that a scatterplot of average life span versus temperature has a linear form. We need statistical inference to decide if it's plausible that there is no linear relationship between the variables, and that the pattern observed in the graph is due simply to the chance involved in the random assignment.

It is conventional to refer to a scatterplot of the points (x, y) as a graph of y versus x. So a scatterplot of life span versus temperature uses life span as the response variable and temperature as the explanatory variable.

In Section 12.1, we will show you how to estimate and test claims about the slope of the population (true) regression line that describes the relationship between two quantitative variables. The following Activity gives you a preview of inference for linear regression.

ACTIVITY The Helicopter Experiment

MATERIALS:

50 copies of the helicopter template from the *Teacher's Resource Materials*, scissors, tape measures, stopwatches

Is there a linear relationship between the height from which a paper helicopter is released and the time it takes to hit the ground? In this Activity, your class will perform an experiment to investigate this question.[2]

1. Follow the directions provided with the template to construct 50 long-rotor helicopters.
2. Randomly assign 10 helicopters to each of five different drop heights. (The experiment works best for drop heights of 5 feet or more.)
3. Work in teams to release the helicopters from their assigned drop heights and record the descent times.
4. Make a scatterplot of the data in Step 3. Find the least-squares regression line for predicting descent time from drop height.
5. Interpret the slope of the regression line from Step 4 in context. What is your best guess for the increase in descent time for each additional foot of drop height?
6. Does it seem plausible that there is really no linear relationship between descent time and drop height and that the observed slope happened just by chance due to the random assignment? Discuss this as a class.

Sometimes a scatterplot reveals that the relationship between two quantitative variables has a strong curved form. One strategy is to transform one or both variables so that the graph shows a linear pattern. Then we can use least-squares regression to fit a linear model to the data. Section 12.2 examines methods of transforming data to achieve linearity.

12.1 Inference for Linear Regression

WHAT YOU WILL LEARN By the end of the section, you should be able to:

- Check the conditions for performing inference about the slope β of the population (true) regression line.
- Interpret the values of a, b, s, SE_b, and r^2 in context, and determine these values from computer output.
- Construct and interpret a confidence interval for the slope β of the population (true) regression line.
- Perform a significance test about the slope β of the population (true) regression line.

In Chapter 3, we examined data on eruptions of the Old Faithful geyser. Figure 12.2 is a scatterplot of the duration and interval of time until the next eruption for all 222 recorded eruptions in a single month. The least-squares regression line for this population of data has been added to the graph. Its equation is

$$\text{predicted interval} = 33.97 + 10.36\,(\text{duration})$$

We call this the **population regression line** (or *true regression line*) because it uses all the observations that month.

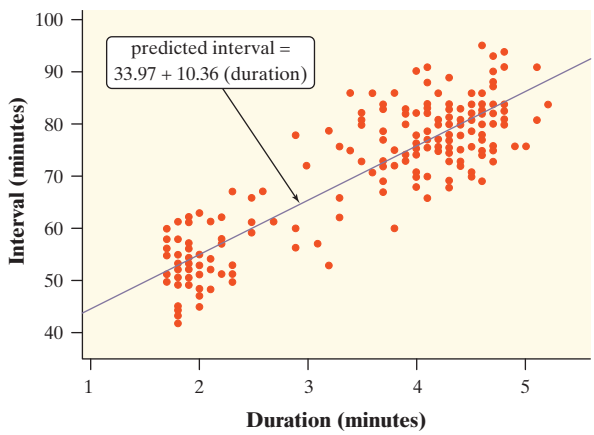
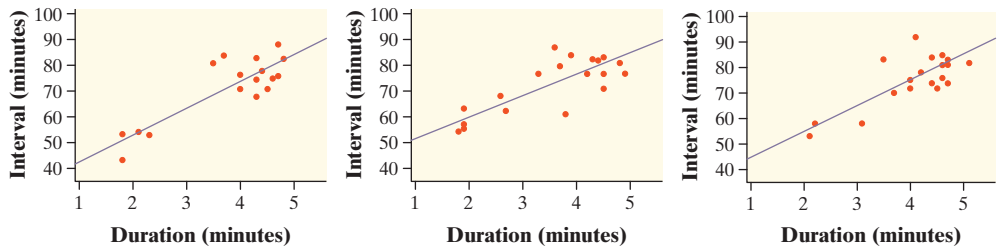


FIGURE 12.2 Scatterplot of the duration and interval between eruptions of Old Faithful for all 222 eruptions in a single month. The population least-squares line is shown in blue.

The sample regression line is sometimes called the *estimated regression line.*

Suppose we take an SRS of 20 eruptions from the population and calculate the least-squares regression line $\hat{y} = a + bx$ for the sample data. How does the slope b of the **sample regression line** relate to the slope of the population regression line? Figure 12.3 on the next page shows the results of taking three different SRSs of 20 Old Faithful eruptions in this month. Each graph displays the selected points and the least-squares regression line for that sample. Notice that the slopes of the sample regression lines (10.2, 7.7, and 9.5) vary quite a bit from the slope of the population regression line, 10.36. The pattern of variation in the slope b is described by its sampling distribution.



(a) Sample 1: $\hat{y} = 32.8 + 10.2x$ (b) Sample 2: $\hat{y} = 44.0 + 7.7x$ (c) Sample 3: $\hat{y} = 36.0 + 9.5x$

FIGURE 12.3 Scatterplots and least-squares regression lines for three different SRSs of 20 Old Faithful eruptions.

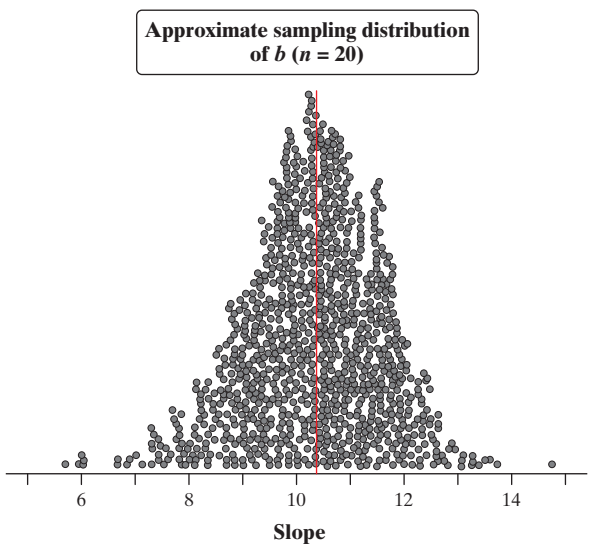


FIGURE 12.4 Dotplot of the slope b of the least-squares regression line in 1000 simulated SRSs by Fathom software.

Sampling Distribution of b

Confidence intervals and significance tests about the slope of the population regression line are based on the sampling distribution of b, the slope of the sample regression line. We used Fathom software to simulate choosing 1000 SRSs of $n = 20$ from the Old Faithful data, each time calculating the equation $\hat{y} = a + bx$ of the least-squares regression line for the sample. Figure 12.4 displays the values of the slope b for the 1000 sample regression lines. We have added a vertical line at 10.36 corresponding to the slope of the population regression line. Let's describe this approximate sampling distribution of b.

Shape: We can see that the distribution of b-values is roughly symmetric and unimodal. Figure 12.5(a) is a Normal probability plot of these sample regression line slopes. The strong linear pattern in the graph tells us that the approximate sampling distribution of b is close to Normal.

Center: The mean of the 1000 b-values is 10.35. This value is quite close to the slope of the population (true) regression line, 10.36.

Spread: The standard deviation of the 1000 b-values is 1.29. Soon, we will see that the standard deviation of the sampling distribution of b is actually 1.27.

Figure 12.5(b) is a histogram of the b-values from the 1000 simulated SRSs. We have superimposed the density curve for a Normal distribution with mean 10.36 and standard deviation 1.27. This curve models the approximate sampling distribution of the slope quite well.

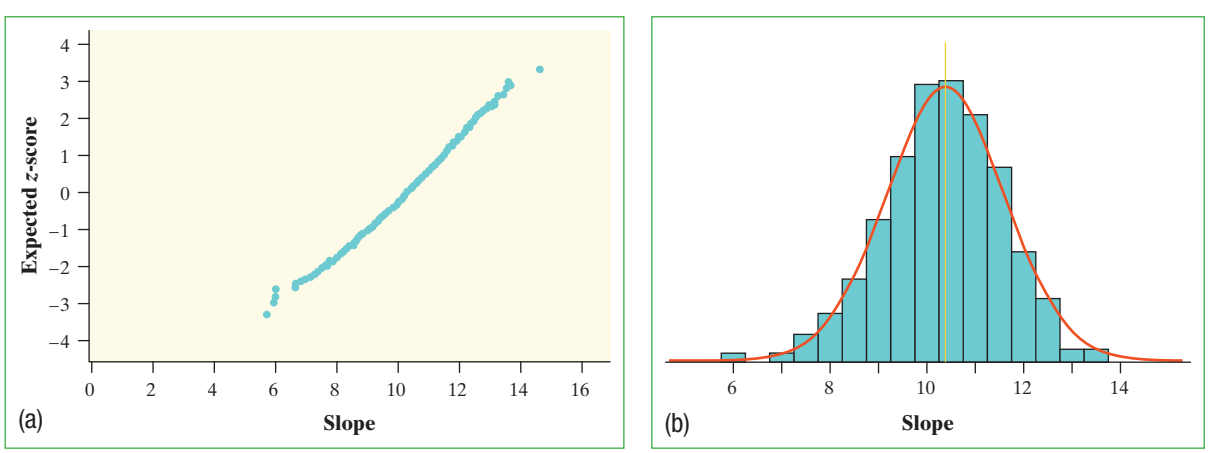
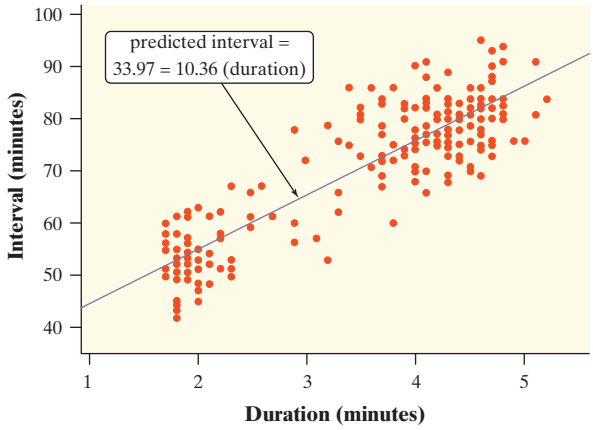


FIGURE 12.5 (a) Normal probability plot and (b) histogram of the 1000 sample regression line slopes from Figure 12.4. The red density curve in Figure 12.5(b) is for a Normal distribution with mean 10.36 (marked by the yellow line) and standard deviation 1.27.



Let's do a quick recap. For all 222 eruptions of Old Faithful in a single month, the population regression line is: predicted interval = 33.97 + 10.36 (duration). We use the symbols $\alpha = 33.97$ and $\beta = 10.36$ to represent the y intercept and slope parameters. The standard deviation of the residuals for this line is the parameter $\sigma = 6.131$.

Figure 12.5(b) shows the approximate sampling distribution of the slope b of the sample regression line for samples of 20 eruptions. If we take *all* possible SRSs of size $n = 20$ from the population, we get the sampling distribution of b. Can you guess its shape, center, and spread?

Shape: Approximately Normal

Center: $\mu_b = \beta = 10.36$ (b is an unbiased estimator of β)

Spread: $\sigma_b = \dfrac{\sigma}{\sigma_x\sqrt{n}} = \dfrac{6.131}{1.0815\sqrt{20}} = 1.27$ where σ_x is the standard deviation of the 222 eruption durations.

We interpret σ_b just like any other standard deviation: the slopes of the sample regression lines typically differ from the slope of the population regression line by about 1.27.

Here's a summary of the important facts about the sampling distribution of b.

Note that the symbols α and β here refer to the intercept and slope, respectively, of the population regression line. They are in no way related to Type I and Type II error probabilities, which are sometimes designated by these same symbols.

SAMPLING DISTRIBUTION OF A SLOPE

Choose an SRS of n observations (x, y) from a population of size N with least-squares regression line

$$\text{predicted } y = \alpha + \beta x$$

Let b be the slope of the sample regression line. Then:

- The **mean** of the sampling distribution of b is $\mu_b = \beta$.
- The **standard deviation** of the sampling distribution of b is

$$\sigma_b = \frac{\sigma}{\sigma_x\sqrt{n}}$$

as long as the *10% condition* is satisfied: $n \le \dfrac{1}{10}N$.

- The sampling distribution of b will be **approximately Normal** if the values of the response variable y follow a Normal distribution for each value of the explanatory variable x (the *Normal condition*).

We'll say more about the Normal condition in a moment.

THINK ABOUT IT

What's with that formula for σ_b? There are three factors that affect the standard deviation of the sampling distribution of b:

- σ, the standard deviation of the residuals for the population regression line. Because σ is in the numerator of the formula, when σ is larger, so is σ_b. When the points are more spread out around the population (true) regression line, we should expect more variability in the slopes b of sample regression lines from repeated random sampling or random assignment.

- σ_x, the standard deviation of the explanatory variable. Because σ_x is in the denominator of the formula, when σ_x is larger, σ_b is smaller. More variability in the values of the explanatory variable leads to a more precise estimate of the slope of the true regression line.
- n, the sample size. Just like every other formula for the standard deviation of a statistic, the variability of the statistic gets smaller as the sample size increases. A larger sample size will lead to a more precise estimate of the true slope.

Conditions for Regression Inference

We can fit a least-squares line to any data relating two quantitative variables, but the results are useful only if the scatterplot shows a linear pattern. Inference about regression involves more detailed conditions. Figure 12.6 shows the regression model when the conditions are met in picture form. The regression model requires that for each possible value of the explanatory variable x:

1. The mean value of the response variable μ_y falls on the population (true) regression line $\mu_y = \alpha + \beta x$.
2. The values of the response variable y follow a Normal distribution with common standard deviation σ.

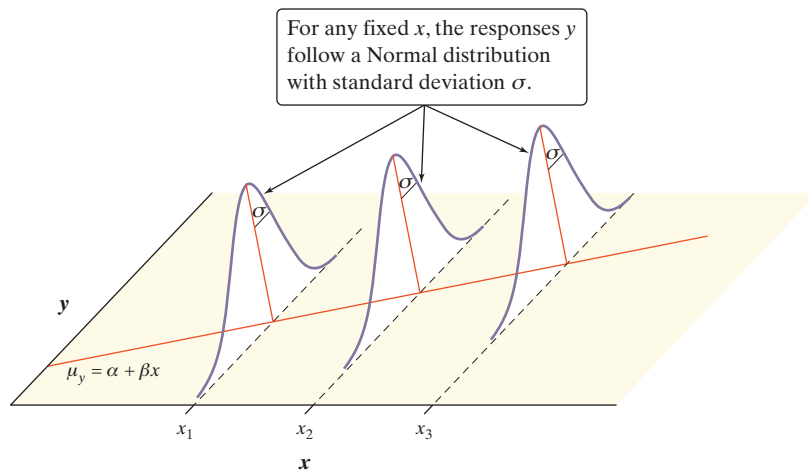


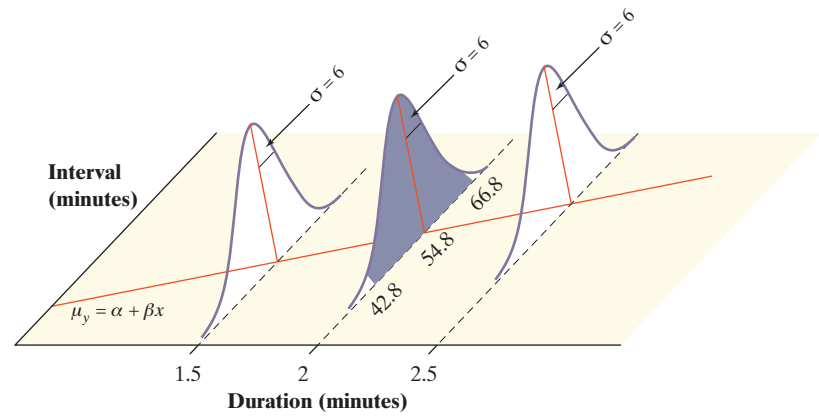
FIGURE 12.6 The regression model when the conditions for inference are met. The line is the population (true) regression line, which shows how the mean response μ_y changes as the explanatory variable x changes. For any fixed value of x, the observed response y varies according to a Normal distribution having mean μ_y and standard deviation σ.

THINK ABOUT IT

What does the regression model in Figure 12.6 tell us? Consider the population of all eruptions of the Old Faithful geyser in a given year. For each eruption, let x be the duration (in minutes) and y be the interval of time (in minutes) until the next eruption. Suppose that the conditions for regression inference are met for this data set, that the population regression line is $\mu_y = 34 + 10.4x$, and that the spread around the line is given by $\sigma = 6$. Let's focus on the eruptions that lasted $x = 2$ minutes. For this "subpopulation":

- The average amount of time until the next eruption is $\mu_y = 34 + 10.4(2) = 54.8$ minutes.

- The amounts of time until the next eruption follow a Normal distribution with mean 54.8 minutes and standard deviation 6 minutes.
- For about 95% of these eruptions, the amount of time y until the next eruption is between $54.8 - 2(6) = 42.8$ minutes and $54.8 + 2(6) = 66.8$ minutes. That is, if the previous eruption lasted 2 minutes, 95% of the time the next eruption will occur in 42.8 to 66.8 minutes.



Here are the conditions for performing inference about the linear regression model.

CONDITIONS FOR REGRESSION INFERENCE

Suppose we have n observations on an explanatory variable x and a response variable y. Our goal is to study or predict the behavior of y for given values of x.

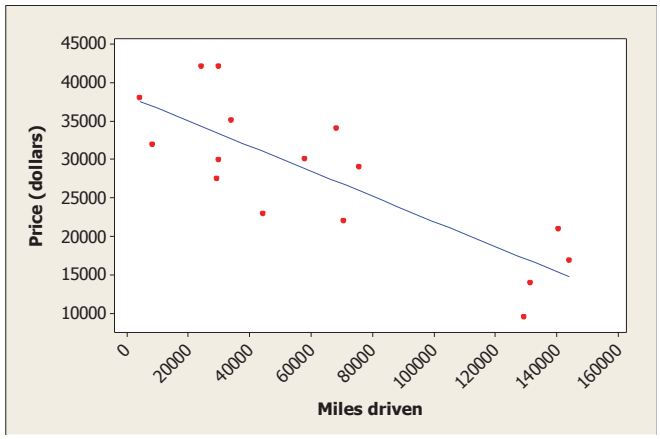
- **Linear:** The actual relationship between x and y is linear. For any fixed value of x, the mean response μ_y falls on the population (true) regression line $\mu_y = \alpha + \beta x$.
- **Independent:** Individual observations are independent of each other. When sampling without replacement, check the *10% condition*.
- **Normal:** For any fixed value of x, the response y varies according to a Normal distribution.
- **Equal SD:** The standard deviation of y (call it σ) is the same for all values of x.
- **Random:** The data come from a well-designed random sample or randomized experiment.

The acronym LINER should help you remember the conditions for inference about regression.

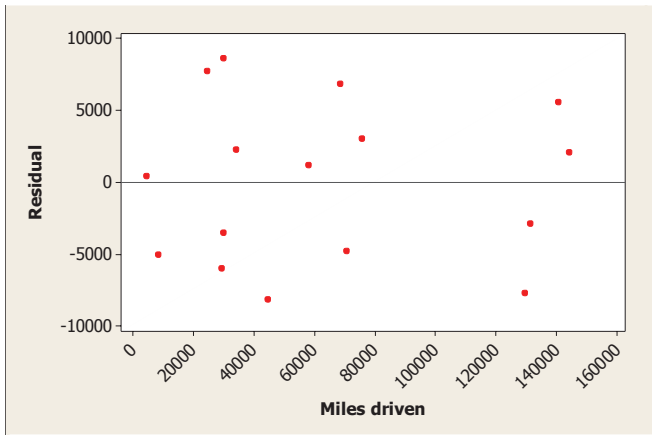
Although the conditions for regression inference are a bit complicated, it is not hard to check for major violations. Most of the conditions involve the population (true) regression line and the deviations of responses from this line. We usually can't observe the population line, but the sample regression line estimates it. The residuals from the sample regression line estimate the deviations from the population line. We can check several of the conditions for regression inference by looking at graphs of the residuals. Start by making a residual plot and a histogram or Normal probability plot of the residuals.

Here's a summary of how to check the conditions one by one.

- **Linear:** Examine the scatterplot to see if the overall pattern is roughly linear. Make sure there are no curved patterns in the residual plot. Check to see that the residuals center on the "residual = 0" line at each x-value in the residual plot.

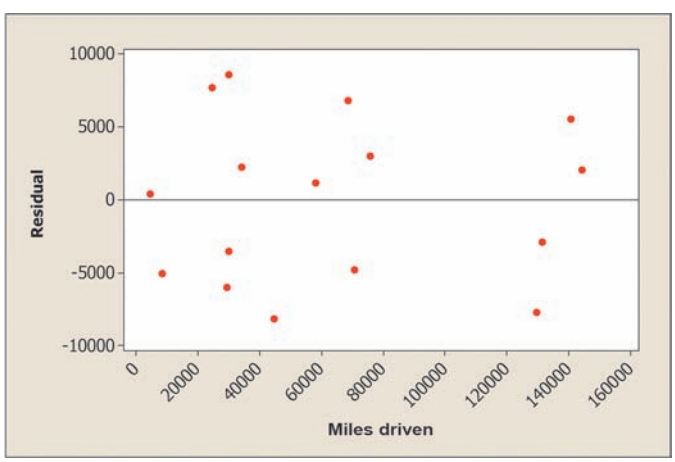


Good: Scatterplot has a linear form.

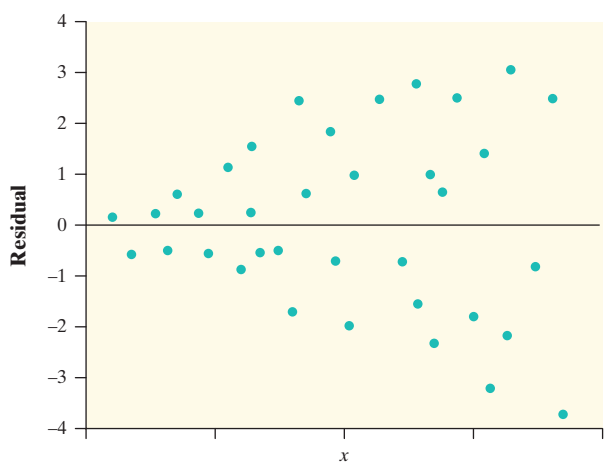


Bad: Residual plot shows a curved pattern.

- **Independent:** Look at how the data were produced. Random sampling and random assignment help ensure the independence of individual observations. If sampling is done without replacement, remember to check that the population is at least 10 times as large as the sample (*10% condition*). But there are other issues that can lead to a lack of independence. One example is measuring the same variable at intervals over time, yielding what is known as *time-series data*. Knowing that a young girl's height at age 6 is 48 inches would definitely give you additional information about her height at age 7. You should avoid doing inference about the regression model for time-series data.
- **Normal:** Make a stemplot, histogram, or Normal probability plot of the residuals and check for clear skewness or other major departures from Normality. Ideally, we would check the distribution of residuals for Normality at each possible value of x. Because we rarely have enough observations at each x-value, however, we make one graph of all the residuals to check for Normality.
- **Equal SD:** Look at the scatter of the residuals above and below the "residual = 0" line in the residual plot. The vertical spread of the residuals should be roughly the same from the smallest to the largest x-value.



Good: Residuals have roughly equal spread at all x-values in the data set.



Bad: The response variable y has greater spread for larger values of the explanatory variable x.

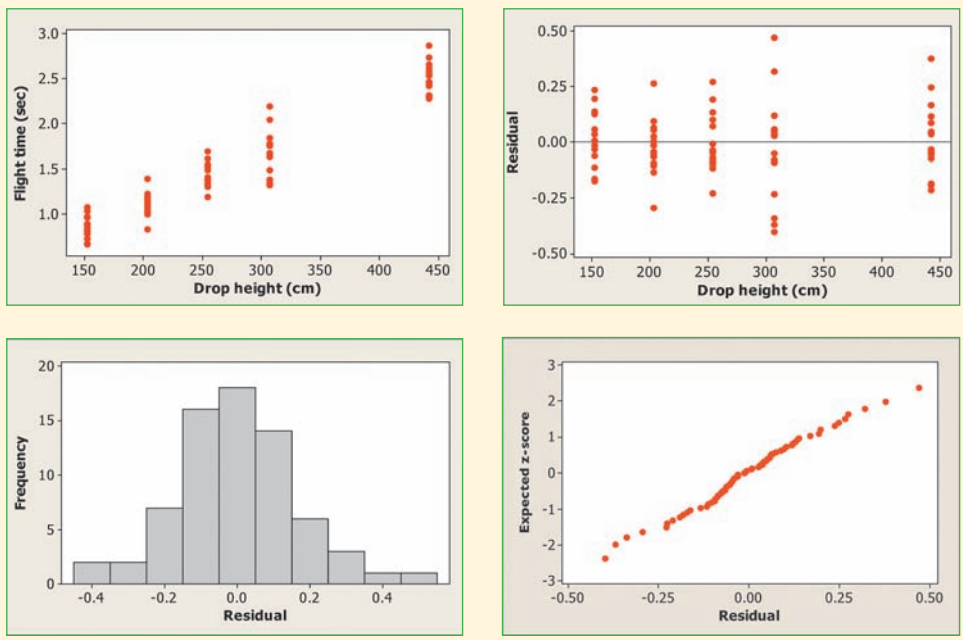
- **Random:** See if the data came from a well-designed random sample or randomized experiment. If not, we can't make inferences about a larger population or about cause and effect.

Let's look at an example that illustrates the process of checking conditions.

EXAMPLE The Helicopter Experiment

Checking conditions

Mrs. Barrett's class did a variation of the helicopter experiment on page 738. Students randomly assigned 14 helicopters to each of five drop heights: 152 centimeters (cm), 203 cm, 254 cm, 307 cm, and 442 cm. Teams of students released the 70 helicopters in a predetermined random order and measured the flight times in seconds. The class used Minitab to carry out a least-squares regression analysis for these data. A scatterplot and residual plot, plus a histogram and Normal probability plot of the residuals are shown below.



PROBLEM: Check whether the conditions for performing inference about the regression model are met.

SOLUTION: We'll use our LINER acronym!

- *Linear:* The scatterplot shows a clear linear form. The residual plot shows a random scatter about the horizontal line. For each drop height used in the experiment, the residuals are centered on the horizontal line at 0.
- *Independent:* Because the helicopters were released in a random order and no helicopter was used twice, knowing the result of one observation should not help us predict the value of another observation.
- *Normal:* The histogram of the residuals is single-peaked and somewhat bell-shaped. In addition, the Normal probability plot is very close to linear.

Note that we do not have to check the 10% condition here because there was no random sampling.

- *Equal SD:* The residual plot shows a similar amount of scatter about the residual = 0 line for the 152, 203, 254, and 442 cm drop heights. Flight times (and the corresponding residuals) seem to vary a little more for the helicopters that were dropped from a height of 307 cm.
- *Random:* The helicopters were randomly assigned to the five possible drop heights.

Except for a slight concern about the equal-SD condition, we should be safe performing inference about the regression model in this setting.

For Practice *Try Exercise* 3

You will always see some irregularity when you look for Normality and equal standard deviation in the residuals, especially when you have few observations. Don't overreact to minor issues in the graphs when checking these two conditions.

AP® EXAM TIP The AP® exam formula sheet gives $\hat{y} = b_0 + b_1x$ for the equation of the sample regression line. We will stick with our simpler notation, $\hat{y} = a + bx$, which is also used by TI calculators. Just remember: the coefficient of x is always the slope, no matter what symbol is used.

Estimating the Parameters

When the conditions are met, we can do inference about the regression model $\mu_y = \alpha + \beta x$. The first step is to estimate the unknown parameters. If we calculate the sample regression line $\hat{y} = a + bx$, the slope b is an unbiased estimator of the true slope β, and the y intercept a is an unbiased estimator of the true y intercept α. The remaining parameter is the standard deviation σ, which describes the variability of the response y about the population (true) regression line.

The least-squares regression line computed from the sample data estimates the population (true) regression line. So the residuals estimate how much y varies about the population line. Because σ is the standard deviation of responses about the population (true) regression line, we estimate it by the standard deviation of the residuals

$$s = \sqrt{\frac{\Sigma \text{residuals}^2}{n-2}} = \sqrt{\frac{\Sigma(y_i - \hat{y}_i)^2}{n-2}}$$

Because s is estimated from data, it is sometimes called the *regression standard error* or the *root mean squared error*.

Recall from Chapter 3 that s describes the size of a "typical" prediction error.

It is possible to do inference about any of the three parameters in the regression model: α, β, or σ. However, the slope β of the population (true) regression line is usually the most important parameter in a regression problem. So we'll restrict our attention to inference about the slope.

When the conditions are met, the sampling distribution of the slope b is approximately Normal with mean $\mu_b = \beta$ and standard deviation

$$\sigma_b = \frac{\sigma}{\sigma_x\sqrt{n}}$$

In practice, we don't know σ for the true regression line. So we estimate it with the standard deviation of the residuals, s. We also don't know the standard deviation σ_x for the population of x-values. For reasons beyond the scope of this text, we replace the denominator with $s_x\sqrt{n-1}$. So we estimate the spread of the sampling distribution of b with the **standard error of the slope**

$$\text{SE}_b = \frac{s}{s_x\sqrt{n-1}}$$

What happens if we transform the values of b by standardizing? Because the sampling distribution of b is approximately Normal, the statistic

$$z = \frac{b - \beta}{\sigma_b}$$

is modeled well by the standard Normal distribution. Replacing the standard deviation σ_b of the sampling distribution with its standard error gives the statistic

$$t = \frac{b - \beta}{\text{SE}_b}$$

which has a t distribution with $n - 2$ degrees of freedom. (The explanation of why df $= n - 2$ is beyond the scope of this book.)

Let's return to the Old Faithful eruption data. Figure 12.7(a) displays the simulated sampling distribution of the slope b from 1000 SRSs of $n = 20$ eruptions. Figure 12.7(b) shows the result of standardizing the b-values from these 1000 samples. The superimposed curve is a t distribution with df = 20 – 2 = 18.

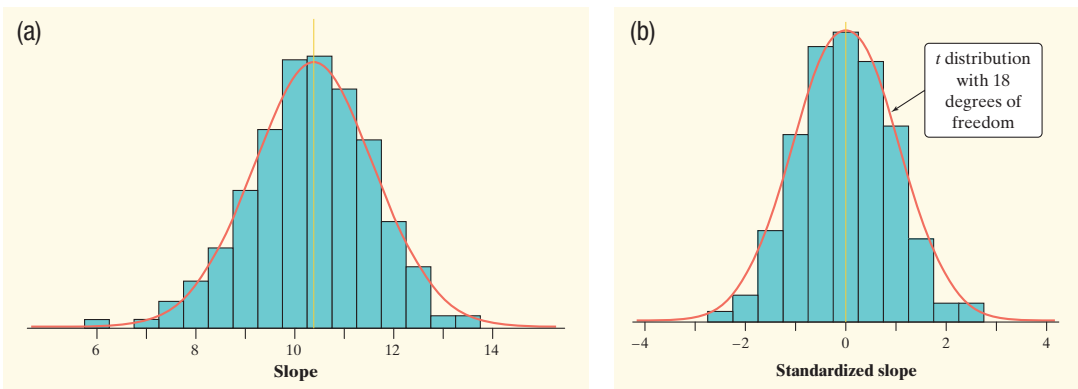


FIGURE 12.7 (a) The approximate sampling distribution of the slope b for samples of size $n = 20$ eruptions. This distribution has a roughly Normal shape with mean about 10.36 and standard deviation about 1.27. (b) The sampling distribution of the standardized slope values has approximately a t distribution with df $= n - 2$.

Constructing a Confidence Interval for the Slope

In a regression setting, we often want to estimate the slope β of the population (true) regression line. The slope b of the sample regression line is our point estimate for β. A confidence interval is more useful than the point estimate because it gives a set of plausible values for β.

The confidence interval for β has the familiar form

$$\text{statistic} \pm (\text{critical value}) \cdot (\text{standard deviation of statistic})$$

Because we use the statistic b as our point estimate, the confidence interval is

$$b \pm t^* \text{SE}_b$$

We call this a ***t* interval for the slope**. Here are the details.

AP® EXAM TIP The AP® exam formula sheet gives the formula for the standard error of the slope as

$$s_{b_1} = \frac{\sqrt{\frac{\Sigma(y_i - \hat{y}_i)^2}{n-2}}}{\sqrt{\Sigma(x_i - \bar{x})^2}}$$

The numerator is just a fancy way of writing the standard deviation of the residuals s. Can you show that the denominator of this formula is the same as ours?

t INTERVAL FOR THE SLOPE

When the conditions for regression inference are met, a $C\%$ confidence interval for the slope β of the population (true) regression line is

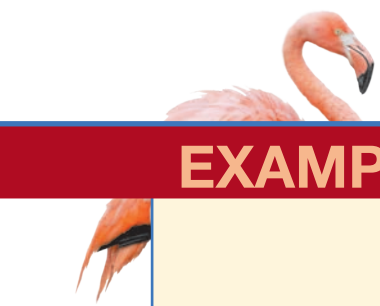
$$b \pm t^* \text{SE}_b$$

In this formula, the standard error of the slope is

$$\text{SE}_b = \frac{s}{s_x\sqrt{n-1}}$$

and t^* is the critical value for the t distribution with df $= n - 2$ having $C\%$ of its area between $-t^*$ and t^*.

Although we give the formula for the standard error of b, you should rarely have to calculate it by hand. Computer output gives the standard error SE_b along with b itself. However we get it, SE_b estimates how much the slope of the sample regression line typically varies from the slope of the population (true) regression line if we repeat the data production process many times.



EXAMPLE

The Helicopter Experiment

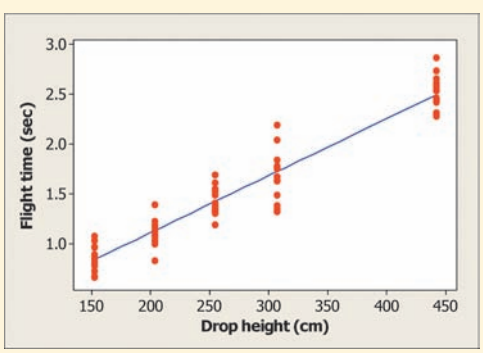
A confidence interval for β

Earlier, we used Minitab to perform a least-squares regression analysis on the helicopter data for Mrs. Barrett's class. Recall that the data came from dropping 70 paper helicopters from various heights and measuring the flight times. Some computer output from this regression is shown below. We checked conditions for performing inference earlier.

Regression Analysis: Flight time versus Drop height

Predictor	Coef	SE Coef	T	P
Constant	−0.03761	0.05838	−0.64	0.522
Drop height (cm)	0.0057244	0.0002018	28.37	0.000

S = 0.168181 R-Sq = 92.2% R-Sq(adj) = 92.1%



PROBLEM:

(a) Give the standard error of the slope, SE_b. Interpret this value in context.

(b) Find the critical value for a 95% confidence interval for the slope of the true regression line. Then calculate the confidence interval. Show your work.

(c) Interpret the interval from part (b) in context.

(d) Explain the meaning of "95% confident" in context.

When we compute the least-squares regression line based on a random sample of data, we can think about doing inference for the population regression line. When our least-squares regression line is based on data from a randomized experiment, as in this example, the resulting inference is about the *true* regression line relating the explanatory and response variables. We'll follow this convention from now on.

SOLUTION:

(a) We got the value of the standard error of the slope, 0.0002018, from the "SE Coef" column in the computer output. If we repeated the random assignment many times, the slope of the sample regression line would typically vary by about 0.0002 from the slope of the true regression line for predicting flight time from drop height.

(b) Because the conditions are met, we can calculate a t interval for the slope β based on a t distribution with $df = n - 2 = 70 - 2 = 68$. Using the more conservative $df = 60$ from Table B gives $t^* = 2.000$. The 95% confidence interval is

$$b \pm t^* SE_b = 0.0057244 \pm 2.000(0.0002018) = 0.0057244 \pm 0.0004036$$
$$= (0.0053208, 0.0061280)$$

Using technology: From `invT(.025,68)`, we get $t^* = 1.995$. The resulting 95% confidence interval is

$$0.0057244 \pm 1.995(0.0002018) = 0.0057244 \pm 0.0004026$$
$$= (0.0053218, 0.0061270)$$

This interval is slightly narrower due to the more precise t^* critical value.

(c) We are 95% confident that the interval from 0.0053218 to 0.0061270 seconds per cm captures the slope of the true regression line relating the flight time y and drop height x of paper helicopters.

(d) If we repeat the experiment many, many times, and use the method in part (b) to construct a confidence interval each time, about 95% of the resulting intervals will capture the slope of the true regression line relating flight time y and drop height x of paper helicopters.

For Practice *Try Exercise* 7

The values of t given in the computer regression output are *not* the critical values for a confidence interval. They come from carrying out a significance test about the y intercept or slope of the population (true) regression line. We'll discuss tests in more detail shortly.

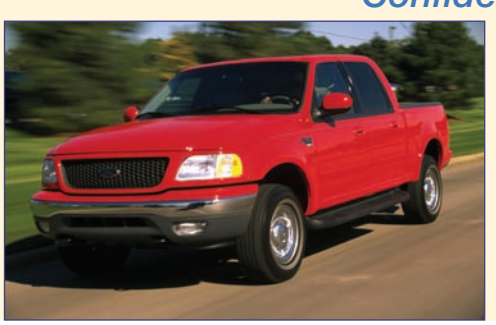
You can find a confidence interval for the y intercept α of the population (true) regression line in the same way, using a and SE_a from the "Constant" row of the Minitab output. However, we are usually interested only in the point estimate for α that's provided in the computer output.

Here is an example using a familiar context that illustrates the four-step process for calculating and interpreting a confidence interval for the slope.

EXAMPLE How Much Is That Truck Worth?

STEP 4

Confidence interval for a slope

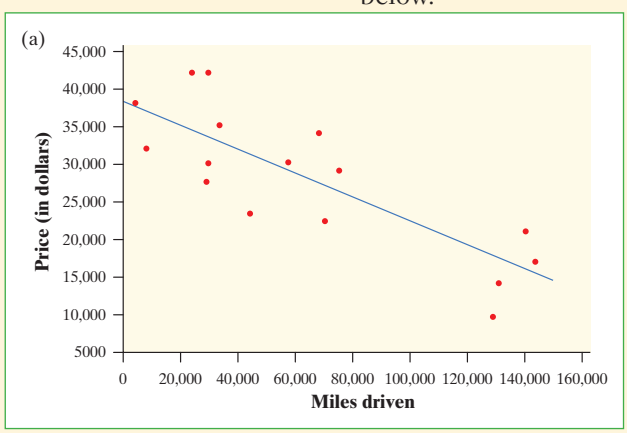


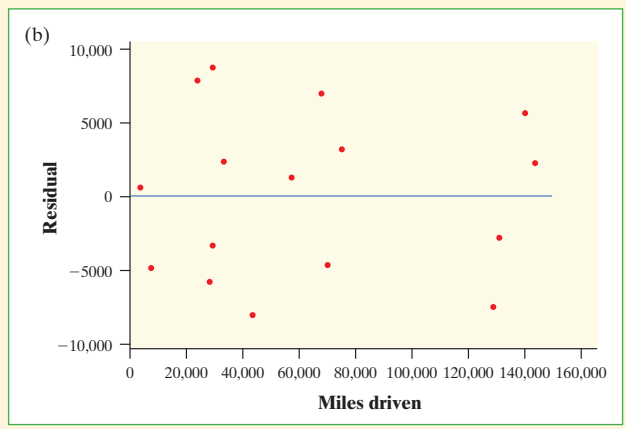
Everyone knows that cars and trucks lose value the more they are driven. Can we predict the price of a used Ford F-150 SuperCrew 4 × 4 if we know how many miles it has on the odometer? A random sample of 16 used Ford F-150 SuperCrew 4 × 4s was selected from among those listed for sale on autotrader.com. The number of miles driven and price (in dollars) were recorded for each of the trucks.[3]

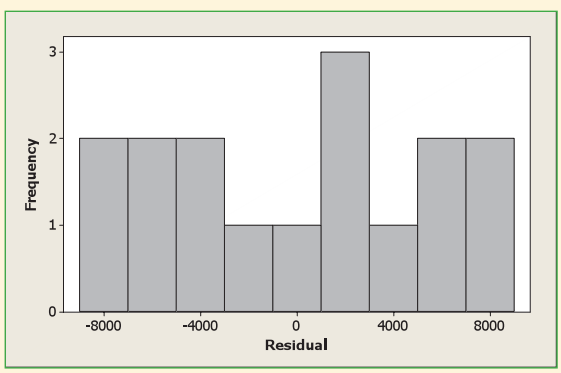
Here are the data:

Miles driven:	70,583	129,484	29,932	29,953	24,495	75,678	8359	4447
Price (in dollars):	21,994	9500	29,875	41,995	41,995	28,986	31,891	37,991
Miles driven:	34,077	58,023	44,447	68,474	144,162	140,776	29,397	131,385
Price (in dollars):	34,995	29,988	22,896	33,961	16,883	20,897	27,495	13,997

Minitab output from a least-squares regression analysis for these data is shown below.







Regression Analysis: Price (dollars) versus Miles driven

Predictor	Coef	SE Coef	T	P
Constant	38257	2446	15.64	0.000
Miles driven	−0.16292	0.03096	−5.26	0.000

S = 5740.13 R-Sq = 66.4% R-Sq(adj) = 64.0%

PROBLEM: Construct and interpret a 90% confidence interval for the slope of the population regression line.

SOLUTION: We will follow the familiar four-step process.

STATE: We want to estimate the slope β of the population regression line relating miles driven to price with 90% confidence.

PLAN: If the conditions are met, we will use a t interval for the slope of a regression line.

- *Linear:* The scatterplot shows a clear linear pattern. Also, the residual plot shows a random scatter of points about the residual = 0 line.
- *Independent:* Because we sampled without replacement to get the data, there have to be at least 10(16) = 160 used Ford F-150 SuperCrew 4 × 4s listed for sale on autotrader.com. This seems reasonable to believe.
- *Normal:* The histogram of the residuals is roughly symmetric and single-peaked, so there are no obvious departures from Normality.
- *Equal SD:* The scatter of points around the residual = 0 line appears to be about the same at all x-values.
- *Random:* We randomly selected the 16 pickup trucks in the sample.

DO: We use the t distribution with $16 - 2 = 14$ degrees of freedom to find the critical value. For a 90% confidence level, the critical value is $t^* = 1.761$. So the 90% confidence interval for β is

$$b \pm t^* SE_b = -0.16292 \pm 1.761(0.03096) = -0.16292 \pm 0.05452$$
$$= (-0.21744, -0.10840)$$

Using technology: Refer to the Technology Corner that follows the example. The calculator's LinRegTInt gives $(-0.2173, -0.1084)$ using $df = 14$.

CONCLUDE: We are 90% confident that the interval from -0.2173 to -0.1084 captures the slope of the population regression line relating price to miles driven for used Ford F-150 SuperCrew 4 × 4s listed for sale on autotrader.com.

For Practice *Try Exercise* 9

The predicted change in price of a used Ford F-150 is quite small for a 1-mile increase in miles driven. What if miles driven increased by 1000 miles? We can just multiply both endpoints of the confidence interval in the example by 1000 to get a 90% confidence interval for the corresponding predicted change in average price. The resulting interval is $(-217.3, -108.4)$. That is, the population regression line predicts a decrease in price of between \$108.40 and \$217.30 for every additional 1000 miles driven.

So far, we have used computer regression output when performing inference about the slope of a population (true) regression line. The TI-83/84, TI-89, and TI-Nspire can do the calculations for inference when the sample data are provided.

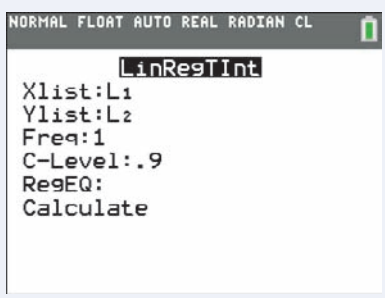
28. TECHNOLOGY CORNER CONFIDENCE INTERVAL FOR SLOPE ON THE CALCULATOR

TI-Nspire instructions in Appendix B; HP Prime instructions on the book's Web site.

Let's use the data from the previous example to construct a confidence interval for the slope of a population (true) regression line on the TI-83/84 and TI-89. Enter the x-values (miles driven) into L1/list1 and the y-values (price) into L2/list2.

TI-83/84 with recent OS

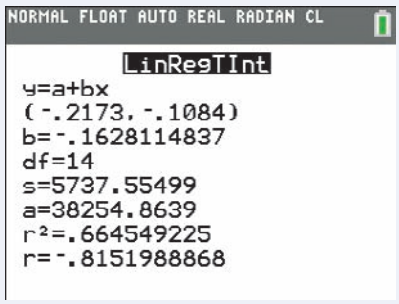
- Press STAT, then choose TESTS and `LinRegTInt. . . .`
- In the LinRegTInt screen, adjust the inputs as shown. Then highlight "Calculate" and press ENTER.



TI-89

- Press 2nd F2 ([F7]) and choose `LinRegTInt. . . .`
- In the LinRegTInt screen, adjust the inputs as shown and press ENTER.

- The linear regression *t* interval results are shown below. The TI-84 Plus C fits the results on one screen. The TI-83/84 and TI-89 require you to arrow down to see the rest of the output.



Note that *s* is the standard deviation of the residuals, *not* the standard error of the slope.

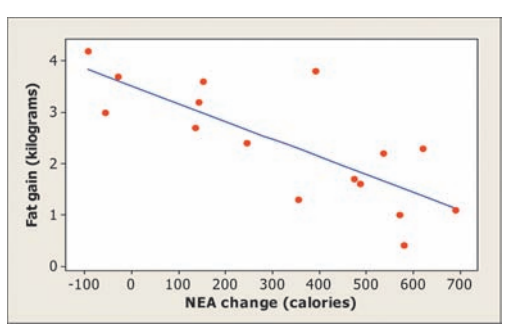
AP® EXAM TIP The formula for the *t* interval for the slope of a population (true) regression line often leads to calculation errors by students. As a result, we recommend using the calculator's `LinRegTInt` feature to compute the confidence interval on the AP® Exam. Be sure to name the procedure (*t* interval for slope) and to give the interval (–0.217, –0.108) and df (14) as part of the "Do" step.

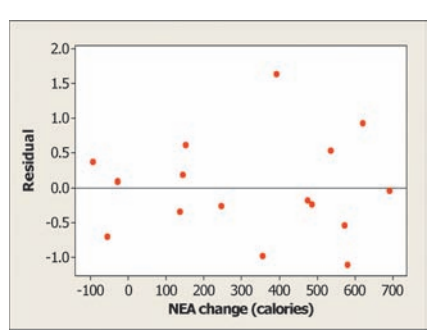
CHECK YOUR UNDERSTANDING

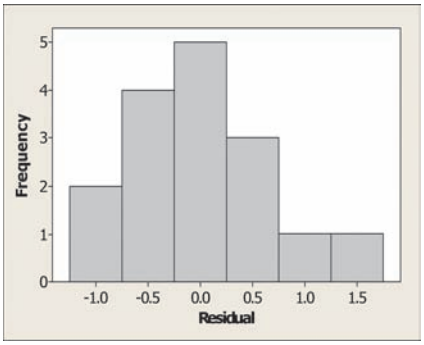
Does fidgeting keep you slim? Some people don't gain weight even when they overeat. Perhaps fidgeting and other "nonexercise activity" (NEA) explain why—some people may spontaneously increase nonexercise activity when fed more. Researchers deliberately overfed a random sample of 16 healthy young adults for 8 weeks. They measured fat gain (in kilograms) as the response variable and change in energy use (in calories) from activity other than deliberate exercise—fidgeting, daily living, and the like—as the explanatory variable. Here are the data:[4]

NEA change (cal):	–94	–57	–29	135	143	151	245	355
Fat gain (kg):	4.2	3.0	3.7	2.7	3.2	3.6	2.4	1.3
NEA change (cal):	392	473	486	535	571	580	620	690
Fat gain (kg):	3.8	1.7	1.6	2.2	1.0	0.4	2.3	1.1

Minitab output from a least-squares regression analysis for these data is shown below.







Regression Analysis: Fat gain versus NEA change

Predictor	Coef	SE Coef	T	P
Constant	3.5051	0.03036	11.54	0.000
NEA change	−0.0034415	0.0007414	−4.64	0.000

S = 0.739853 R-Sq = 60.6% R-Sq(adj) = 57.8%

Construct and interpret a 95% confidence interval for the slope of the population (true) regression line.

Performing a Significance Test for the Slope

When the conditions for inference are met, we can use the slope b of the sample regression line to construct a confidence interval for the slope β of the population (true) regression line. We can also perform a significance test to determine whether a specified value of β is plausible. The null hypothesis has the general form $H_0: \beta = \beta_0$. To do a test, standardize b to get the test statistic:

$$\text{test statistic} = \frac{\text{statistic} - \text{parameter}}{\text{standard deviation of statistic}}$$

$$t = \frac{b - \beta_0}{SE_b}$$

To find the P-value, use a t distribution with $n - 2$ degrees of freedom. Here are the details for the ***t* test for the slope**.

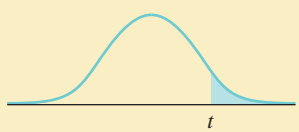
t TEST FOR THE SLOPE

Suppose the conditions for inference are met. To test the hypothesis $H_0: \beta = \beta_0$, compute the test statistic

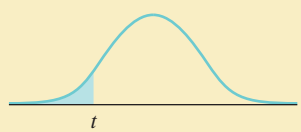
$$t = \frac{b - \beta_0}{SE_b}$$

Find the P-value by calculating the probability of getting a t statistic this large or larger in the direction specified by the alternative hypothesis H_a. Use the t distribution with df $= n - 2$.

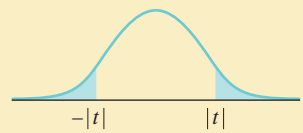
$H_a: \beta >$ hypothesized value



$H_a: \beta <$ hypothesized value



$H_a: \beta \neq$ hypothesized value



If sample data suggest a linear relationship between two variables, how can we determine whether this happened just by chance or whether there is actually a linear relationship between x and y in the population? By performing a test of $H_0: \beta = 0$. A regression line with slope 0 is horizontal. That is, the mean of y does not change at all when x changes. So $H_0: \beta = 0$ says that there is *no linear relationship* between x and y in the population. Put another way, H_0 says that *linear regression of y on x is of no value for predicting y.*

Most technology will only do a test with $H_0 : \beta = 0$.

Regression output from statistical software usually gives t and its two-sided P-value for a test of $H_0 : \beta = 0$. For a one-sided test in the proper direction, just divide the P-value in the output by 2. The following example shows what we mean.

EXAMPLE

Crying and IQ

STEP 4

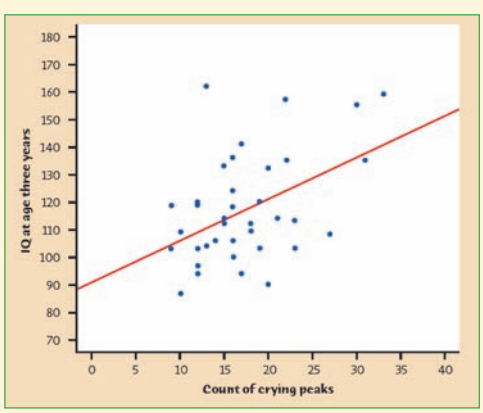
Significance test for β

Infants who cry easily may be more easily stimulated than others. This may be a sign of higher IQ. Child development researchers explored the relationship between the crying of infants 4 to 10 days old and their later IQ test scores. A snap of a rubber band on the sole of the foot caused the infants to cry. The researchers recorded the crying and measured its intensity by the number of peaks in the most active 20 seconds. They later measured the children's IQ at age three years using the Stanford-Binet IQ test. The table below contains data from a random sample of 38 infants.[5]

AP® EXAM TIP When you see a list of data values on an exam question, don't just start typing the data into your calculator. Read the question first. Often, additional information is provided that makes it unnecessary for you to enter the data at all. This can save you valuable time on the AP® exam.

Crycount	IQ	Crycount	IQ	Crycount	IQ	Crycount	IQ
10	87	20	90	17	94	12	94
12	97	16	100	19	103	12	103
9	103	23	103	13	104	14	106
16	106	27	108	18	109	10	109
18	109	15	112	18	112	23	113
15	114	21	114	16	118	9	119
12	119	12	120	19	120	16	124
20	132	15	133	22	135	31	135
16	136	17	141	30	155	22	157
33	159	13	162				

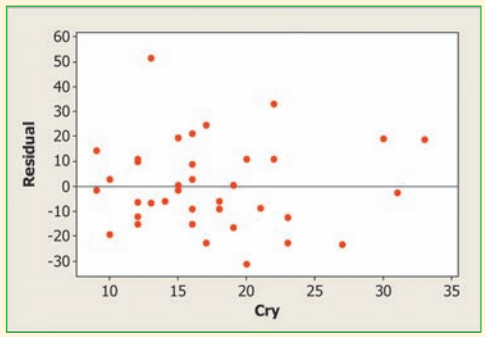
Some computer output from a least-squares regression analysis on these data is shown below.

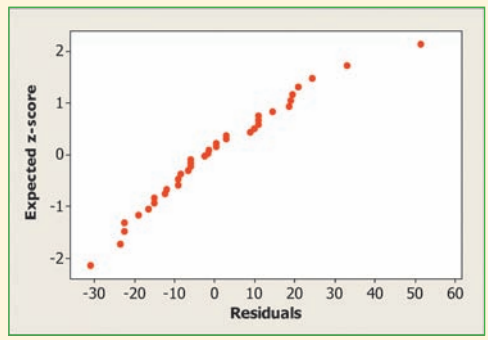


Regression Analysis: IQ versus Crycount

Predictor	Coef	SE Coef	T	P
Constant	91.268	8.934	10.22	0.000
Crycount	1.4929	0.4870	3.07	0.004

S = 17.50 R-Sq = 20.7% R-Sq(adj) = 18.5%





PROBLEM:

(a) What is the equation of the least-squares regression line for predicting IQ at age 3 from the number of crying peaks (crycount)? Interpret the slope and y intercept of the regression line in context.

(b) Explain what the value of s means in this setting.

(c) Do these data provide convincing evidence of a positive linear relationship between crying counts and IQ in the population of infants?

SOLUTION:

(a) The equation of the least-squares line is

$$\text{predicted IQ score} = 91.268 + 1.4929\,(\text{crycount})$$

Because there were no infants who recorded fewer than 9 crying peaks in their most active 20 seconds, it is a risky extrapolation to use this line to predict the value of y when $x = 0$.

Slope: For each additional crying peak in the most active 20 seconds, the regression line predicts an increase of about 1.5 IQ points. *y intercept:* The model predicts that an infant who doesn't cry when flicked with a rubber band will have a later IQ score of about 91.

(b) The size of a typical prediction error when using the regression line in part (a) is 17.50 IQ points.

(c) We'll follow the four-step process.

STATE: We want to perform a test of

$$H_0: \beta = 0$$
$$H_a: \beta > 0$$

where β is the slope of the population regression line relating crying count to IQ score. No significance level was given, so we'll use $\alpha = 0.05$.

PLAN: If the conditions are met, we will do a t test for the slope β.

- *Linear:* The scatterplot suggests a moderately weak positive linear relationship between crying peaks and IQ. The residual plot shows a random scatter of points about the residual $= 0$ line.
- *Independent:* Due to sampling without replacement, there have to be at least $10(38) = 380$ infants in the population from which these children were selected.
- *Normal:* The Normal probability plot of the residuals shows slight curvature, but no strong skewness or obvious outliers that would prevent use of t procedures.
- *Equal SD:* The residual plot shows a fairly equal amount of scatter around the horizontal line at 0 for all x-values.
- *Random:* We are told that these 38 infants were randomly selected.

Our usual formula for the test statistic confirms the value in the computer output:

$$t = \frac{b - \beta_0}{SE_b} = \frac{1.4929 - 0}{0.4870} = 3.07$$

DO: We can get the test statistic and *P*-value from the Minitab output.

- *Test statistic:* $t = 3.07$ (look in the "T" column of the computer output across from "Crycount")
- *P-value:* Figure 12.8 displays the *P*-value for this one-sided test as an area under the *t* distribution curve with $38 - 2 = 36$ degrees of freedom. The Minitab output gives $P = 0.004$ as the *P*-value for a two-sided test. The *P*-value for the one-sided test is half of this, $P = 0.002$.

Using technology: Refer to the Technology Corner that follows the example. The calculator's LinRegTTest gives $t = 3.065$ and *P*-value $= 0.002$ using $df = 36$.

CONCLUDE: Because the *P*-value, 0.002, is less than $\alpha = 0.05$, we reject H_0. There is convincing evidence of a positive linear relationship between intensity of crying and IQ score in the population of infants.

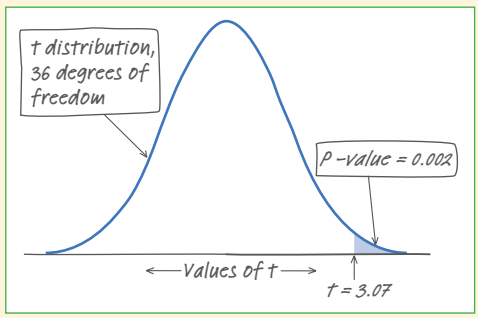


FIGURE 12.8 The *P*-value for the one-sided test.

For Practice *Try Exercise* 13

Based on the results of the crying and IQ study, should we ask doctors and parents to make infants cry more so that they'll be smarter later in life? Hardly. This observational study gives statistically significant evidence of a positive linear relationship between the two variables. However, we can't conclude that more intense crying as an infant *causes* an increase in IQ. Maybe infants who cry more are more alert to begin with and tend to score higher on intelligence tests.

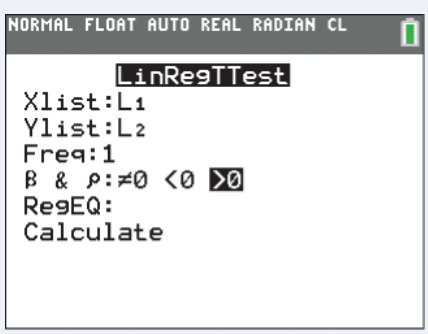
29. TECHNOLOGY CORNER SIGNIFICANCE TEST FOR SLOPE ON THE CALCULATOR

TI-Nspire instructions in Appendix B; HP Prime instructions on the book's Web site.

Let's use the data from the crying and IQ study to perform a significance test for the slope of the population regression line on the TI-83/84 and TI-89. Enter the *x*-values (crying count) into L1/list1 and the *y*-values (IQ score) into L2/list2.

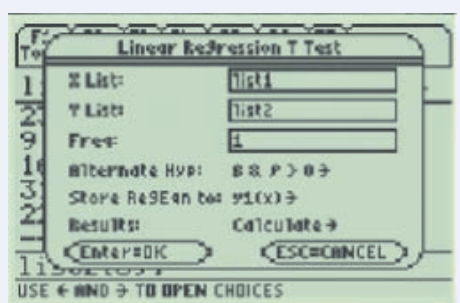
TI-83/84

- Press STAT, then choose TESTS and `LinRegTTest. . . .`
- In the LinRegTTest screen, adjust the inputs as shown. Then highlight "Calculate" and press ENTER.

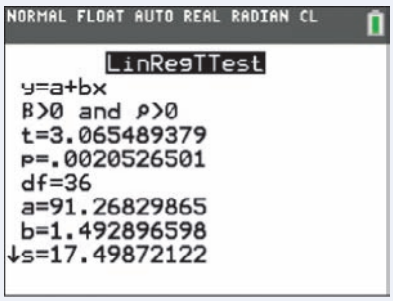


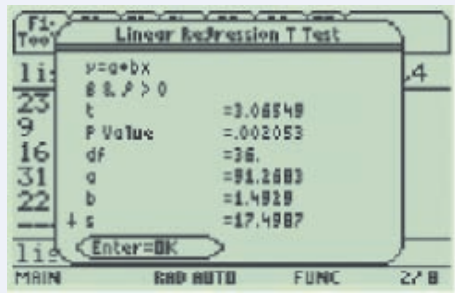
TI-89

- Press 2nd F1 ([F6]) and choose `LinRegTTest. . . .`
- In the LinRegTTest screen, adjust the inputs as shown and press ENTER.



- The linear regression t test results take two screens to present. We show only the first screen.





AP® EXAM TIP The formula for the test statistic in a t test for the slope of a population (true) regression line often leads to calculation errors by students. As a result, we recommend using the calculator's `LinRegTTest` feature to perform calculations on the AP® exam. Be sure to name the procedure (t test for slope) and to report the test statistic ($t = 3.065$), P-value (0.002), and df (36) as part of the "Do" step.

THINK ABOUT IT

What's with that $\rho > 0$ in the LinRegTTest screen? The slope b of the least-squares regression line is closely related to the correlation r between the explanatory and response variables x and y. (Recall that $b = r\frac{s_y}{s_x}$). In the same way, the slope β of the population regression line is closely related to the correlation ρ (the lowercase Greek letter rho) between x and y in the population. In particular, the slope is 0 when the correlation is 0.

Testing the null hypothesis $H_0: \beta = 0$ is, therefore, exactly the same as testing that there is *no correlation* between x and y in the population from which we drew our data. You can use the test for zero slope to test the hypothesis $H_0: \rho = 0$ of zero correlation between any two quantitative variables. That's a useful trick. Because correlation also makes sense when there is no explanatory-response distinction, it is handy to be able to test correlation without doing regression.

CHECK YOUR UNDERSTANDING

The previous Check Your Understanding (page 752) described some results from a study of nonexercise activity (NEA) and fat gain. Here, again, is the Minitab output from a least-squares regression analysis for these data.

Regression Analysis: Fat gain versus NEA change

Predictor	Coef	SE Coef	T	P
Constant	3.5051	0.3036	11.54	0.000
NEA change	−0.0034415	0.0007414	−4.64	0.000

S = 0.739853 R-Sq = 60.6% R-Sq(adj) = 57.8%

Do these data provide convincing evidence at the $\alpha = 0.05$ significance level of a negative linear relationship between fat gain and NEA change in the population of healthy young adults? Assume that the conditions for regression inference are met.

Section 12.1 Summary

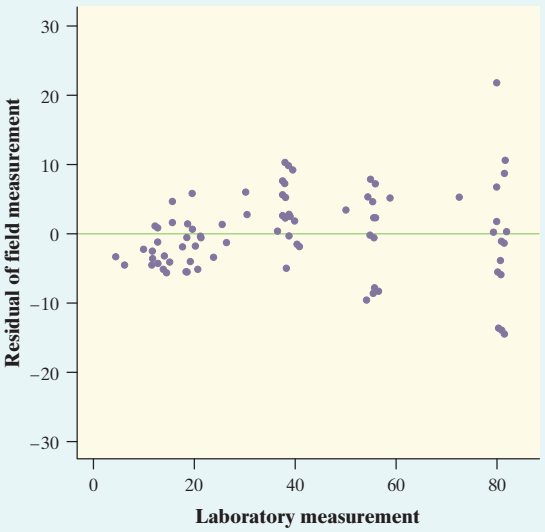
- Least-squares regression fits a straight line of the form $\hat{y} = a + bx$ to data to predict a response variable y from an explanatory variable x. Inference in this setting uses the **sample regression line** to estimate or test a claim about the **population (true) regression line**.
- The conditions for regression inference are
  - **Linear:** The actual relationship between x and y is linear. For any fixed value of x, the mean response μ_y falls on the population (true) regression line $\mu_y = \alpha + \beta x$.
  - **Independent:** Individual observations are independent. When sampling is done without replacement, check the *10% condition*.
  - **Normal:** For any fixed value of x, the response y varies according to a Normal distribution.
  - **Equal SD:** The standard deviation of y (call it σ) is the same for all values of x.
  - **Random:** The data are produced from a well-designed random sample or randomized experiment.
- The slope b and intercept a of the sample regression line estimate the slope β and intercept α of the population (true) regression line. Use the standard deviation of the residuals, s, to estimate σ.
- Confidence intervals and significance tests for the slope β of the population regression line are based on a t distribution with $n - 2$ degrees of freedom.
- The ***t* interval for the slope** β has the form $b \pm t^*\text{SE}_b$, where the **standard error of the slope** is $\text{SE}_b = \dfrac{s}{s_x\sqrt{n-1}}$.
- To test the null hypothesis $H_0: \beta = \beta_0$, carry out a ***t* test for the slope.** This test uses the statistic $t = \dfrac{b - \beta_0}{\text{SE}_b}$. The most common null hypothesis is $H_0: \beta = 0$, which says that there is no linear relationship between x and y in the population.

12.1 TECHNOLOGY CORNERS

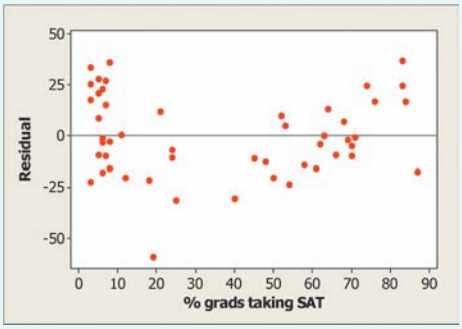
TI-Nspire Instructions in Appendix B; HP Prime instructions on the book's Web site.

Section 12.1 Exercises

1. **Oil and residuals** Exercise 53 on page 194 (Chapter 3) examined data on the depth of small defects in the Trans-Alaska Oil Pipeline. Researchers compared the results of measurements on 100 defects made in the field with measurements of the same defects made in the laboratory.[6] The figure below shows a residual plot for the least-squares regression line based on these data. Explain why the conditions for performing inference about the slope β of the population regression line are not met.

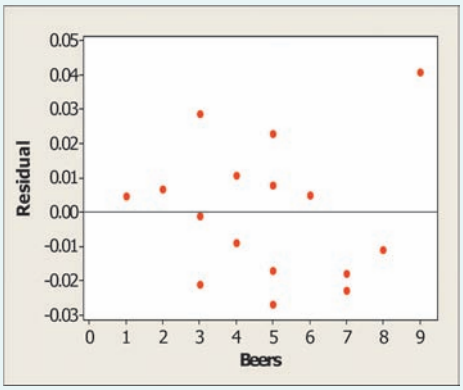


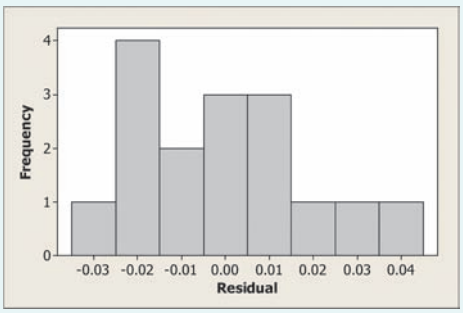
2. **SAT Math scores** In Chapter 3, we examined data on the percent of high school graduates in each state who took the SAT and the state's mean SAT Math score in a recent year. The figure below shows a residual plot for the least-squares regression line based on these data. Explain why the conditions for performing inference about the slope β of the population regression line are not met.



3. **Beer and BAC** How well does the number of beers a person drinks predict his or her blood alcohol content (BAC)? Sixteen volunteers aged 21 or older with an initial BAC of 0 took part in a study to find out. Each volunteer drank a randomly assigned number of cans of beer. Thirty minutes later, a police officer measured their BAC. Least-squares regression was performed on the data. A residual plot and a histogram of the residuals are shown below. Check whether the conditions for performing inference about the regression model are met.

pg 745





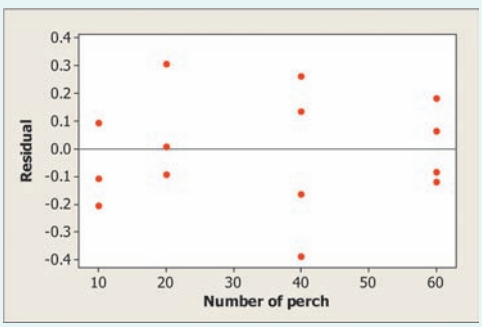
4. **Prey attracts predators** Here is one way in which nature regulates the size of animal populations: high population density attracts predators, which remove a higher proportion of the population than when the density of the prey is low. One study looked at kelp perch and their common predator, the kelp bass. The researcher set up four large circular pens on sandy ocean bottoms off the coast of southern California. He chose young perch at random from a large group and placed 10, 20, 40, and 60 perch in the four pens. Then he dropped the nets protecting the pens, allowing bass to swarm in, and counted the

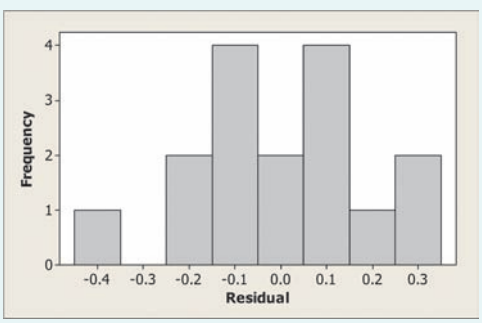
perch left after two hours. Here are data on the proportions of perch eaten in four repetitions of this setup:[7]

Number of Perch	Proportion Killed			
10	0.0	0.1	0.3	0.3
20	0.2	0.3	0.3	0.6
40	0.075	0.3	0.6	0.725
60	0.517	0.55	0.7	0.817

The explanatory variable is the number of perch (the prey) in a confined area. The response variable is the proportion of perch killed by bass (the predator) in two hours when the bass are allowed access to the perch. A scatterplot of the data shows a linear relationship.

We used Minitab software to carry out a least-squares regression analysis for these data. A residual plot and a histogram of the residuals are shown below. Check whether the conditions for performing inference about the regression model are met.





5. **Beer and BAC** Refer to Exercise 3. Computer output from the least-squares regression analysis on the beer and blood alcohol data is shown below.

```
Dependent variable is:    BAC
No Selector
R squared = 80.0%  R squared (adjusted) = 78.6%
s = 0.0204 with 16 − 2 = 14 degrees of freedom

Variable    Coefficient  s.e. of Coeff  t-ratio    prob
Constant    −0.012701      0.0126       −1.00      0.3320
Beers        0.017964      0.0024        7.84     ≤0.0001
```

The model for regression inference has three parameters: α, β, and σ. Explain what each parameter represents in context. Then provide an estimate for each.

6. **Prey attracts predators** Refer to Exercise 4. Computer output from the least-squares regression analysis on the perch data is shown below.

```
Predictor   Coef       Stdev.     t-ratio   p
Constant    0.12049    0.09269    1.30      0.215
Perch       0.008569   0.002456   3.49      0.004

S = 0.1886    R-Sq = 46.5%    R-Sq(adj) = 42.7%
```

The model for regression inference has three parameters: α, β, and σ. Explain what each parameter represents in context. Then provide an estimate for each.

7. **Beer and BAC** Refer to Exercise 5.

pg 748

(a) Give the standard error of the slope, SE_b. Interpret this value in context.

(b) Find the critical value for a 99% confidence interval for the slope of the true regression line. Then calculate the confidence interval. Show your work.

(c) Interpret the interval from part (b) in context.

(d) Explain the meaning of "99% confident" in context.

8. **Prey attracts predators** Refer to Exercise 6.

(a) Give the standard error of the slope, SE_b. Interpret this value in context.

(b) Find the critical value for a 90% confidence interval for the slope of the true regression line. Then calculate the confidence interval. Show your work.

(c) Interpret the interval from part (b) in context.

(d) Explain the meaning of "90% confident" in context.

9. **Beavers and beetles** Do beavers benefit beetles? Researchers laid out 23 circular plots, each 4 meters in diameter, at random in an area where beavers were cutting down cottonwood trees. In each plot, they counted the number of stumps from trees cut by beavers and the number of clusters of beetle larvae. Ecologists think that the new sprouts from stumps are more tender than other cottonwood growth, so that beetles prefer them. If so, more stumps should produce more beetle larvae.[8]

pg 749

Minitab output for a regression analysis on these data is shown below. Construct and interpret a 99% confidence interval for the slope of the population regression line. Assume that the conditions for performing inference are met.

Regression Analysis: Beetle larvae versus Stumps

```
Predictor   Coef      SE Coef   T       P
Constant    −1.286    2.853     −0.45   0.657
Stumps      11.894    1.136     10.47   0.000

S = 6.41939   R-Sq = 83.9%   R-Sq(adj) = 83.1%
```

10. Ideal proportions The students in Mr. Shenk's class measured the arm spans and heights (in inches) of a random sample of 18 students from their large high school. Some computer output from a least-squares regression analysis on these data is shown below. Construct and interpret a 90% confidence interval for the slope of the population regression line. Assume that the conditions for performing inference are met.

Predictor	Coef	Stdev	t-ratio	p
Constant	11.547	5.600	2.06	0.056
Armspan	0.84042	0.08091	10.39	0.000

S = 1.613 R-Sq = 87.1% R-Sq(adj) = 86.3%

11. Beavers and beetles Refer to Exercise 9.

(a) How many clusters of beetle larvae would you predict in a circular plot with 5 tree stumps cut by beavers? Show your work.

(b) About how far off do you expect the prediction in part (a) to be from the actual number of clusters of beetle larvae? Justify your answer.

12. Ideal proportions Refer to Exercise 10.

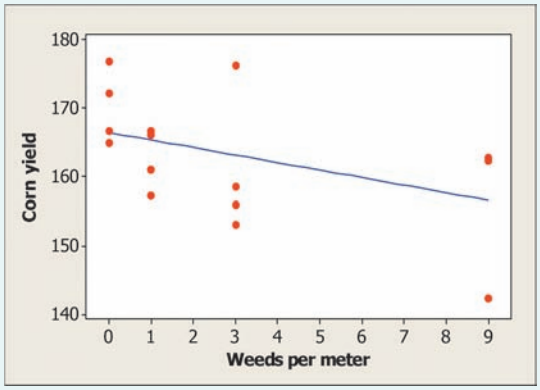
(a) What height would you predict for a student with an arm span of 76 inches? Show your work.

(b) About how far off do you expect the prediction in part (a) to be from the student's actual height? Justify your answer.

13. Weeds among the corn Lamb's-quarter is a common weed that interferes with the growth of corn. An agriculture researcher planted corn at the same rate in 16 small plots of ground and then weeded the plots by hand to allow a fixed number of lamb's-quarter plants to grow in each meter of corn row. The decision of how many of these plants to leave in each plot was made at random. No other weeds were allowed to grow. Here are the yields of corn (bushels per acre) in each of the plots:[9]

pg 754

Some computer output from a least-squares regression analysis on these data is shown below.



Predictor	Coef	SE Coef	T	P
Constant	166.483	2.725	61.11	0.000
Weeds per meter	−1.0987	0.5712	−1.92	0.075

S = 7.97665 R-Sq = 20.9% R-Sq(adj) = 15.3%

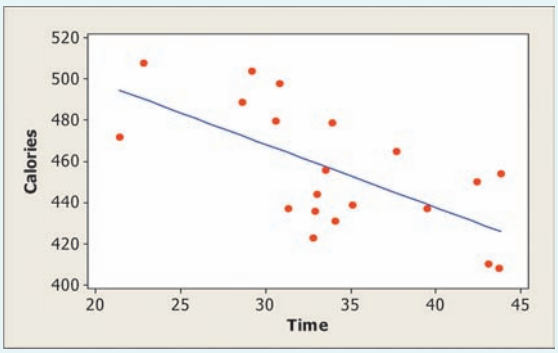
(a) What is the equation of the least-squares regression line for predicting corn yield from the number of lamb's quarter plants per meter? Interpret the slope and y intercept of the regression line in context.

(b) Explain what the value of s means in this settting.

(c) Do these data provide convincing evidence at the $\alpha = 0.05$ level that more weeds reduce corn yield? Assume that the conditions for performing inference are met.

14. Time at the table Does how long young children remain at the lunch table help predict how much they eat? Here are data on a random sample of 20 toddlers observed over several months.[10] "Time" is the average number of minutes a child spent at the table when lunch was served. "Calories" is the average number of calories the child consumed during lunch, calculated from careful observation of what the child ate each day.

Some computer output from a least-squares regression analysis on these data is shown below.



Predictor	Coef	SE Coef	T	P
Constant	560.65	29.37	19.09	0.000
Time	−3.0771	0.8498	−3.62	0.002

S = 23.3980 R-Sq = 42.1% R-Sq(adj) = 38.9%

(a) What is the equation of the least-squares regression line for predicting calories consumed from time at the table? Interpret the slope of the regression line in context. Does it make sense to interpret the y intercept in this case? Why or why not?

(b) Explain what the value of s means in this setting.

(c) Do these data provide convincing evidence at the $\alpha = 0.01$ level of a linear relationship between time

at the table and calories consumed in the population of toddlers? Assume that the conditions for performing inference are met.

15. **Is wine good for your heart?** A researcher from the University of California, San Diego, collected data on average per capita wine consumption and heart disease death rate in a random sample of 19 countries for which data were available. The following table displays the data.[11]

Alcohol from wine (liters/year)	Heart disease death rate (per 100,000)	Alcohol from wine (liters/year)	Heart disease death rate (per 100,000)
2.5	211	7.9	107
3.9	167	1.8	167
2.9	131	1.9	266
2.4	191	0.8	227
2.9	220	6.5	86
0.8	297	1.6	207
9.1	71	5.8	115
2.7	172	1.3	285
0.8	211	1.2	199
0.7	300		

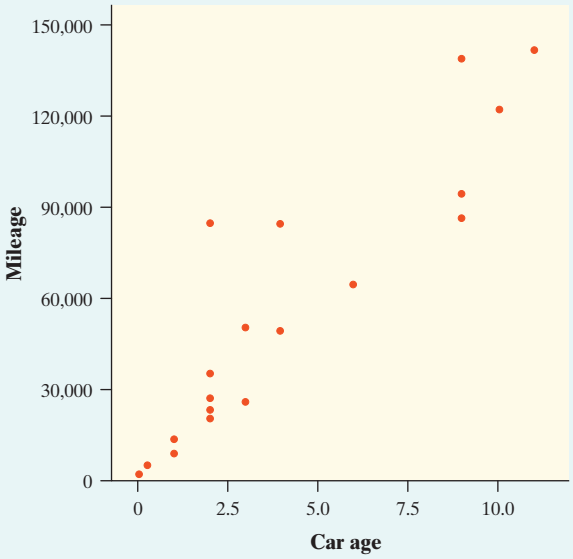
Is there statistically significant evidence of a negative linear relationship between wine consumption and heart disease deaths in the population of countries? Carry out an appropriate significance test at the $\alpha = 0.05$ level.

16. **The professor swims** Here are data on the time (in minutes) Professor Moore takes to swim 2000 yards and his pulse rate (beats per minute) after swimming on a random sample of 23 days:

Time:	34.12	35.72	34.72	34.05	34.13	35.72
Pulse:	152	124	140	152	146	128
Time:	36.17	35.57	35.37	35.57	35.43	36.05
Pulse:	136	144	148	144	136	124
Time:	34.85	34.70	34.75	33.93	34.60	34.00
Pulse:	148	144	140	156	136	148
Time:	34.35	35.62	35.68	35.28	35.97	
Pulse:	148	132	124	132	139	

Is there statistically significant evidence of a negative linear relationship between Professor Moore's swim time and his pulse rate in the population of days on which he swims 2000 yards? Carry out an appropriate significance test at the $\alpha = 0.05$ level.

17. **Stats teachers' cars** A random sample of AP® Statistics teachers was asked to report the age (in years) and mileage of their primary vehicles. A scatterplot of the data is shown at top right.



Computer output from a least-squares regression analysis of these data is shown below (df = 19). Assume that the conditions for regression inference are met.

```
Variable  Coef      SE Coef  t-ratio  prob
Constant   7288.54  6591      1.11     0.2826
Car age   11630.6   1249      9.31    <0.0001
S = 19280    R-Sq = 82.0%    RSq(adj) = 81.1%
```

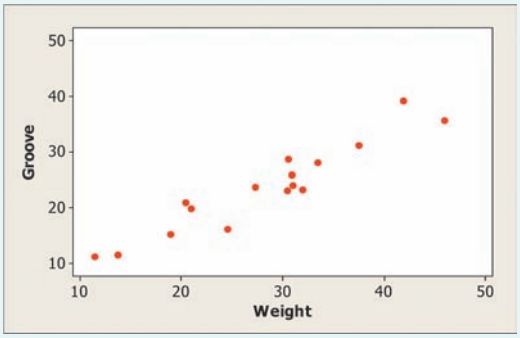
(a) Verify that the 95% confidence interval for the slope of the population regression line is (9016.4, 14,244.8).

(b) A national automotive group claims that the typical driver puts 15,000 miles per year on his or her main vehicle. We want to test whether AP® Statistics teachers are typical drivers. Explain why an appropriate pair of hypotheses for this test is $H_0: \beta = 15{,}000$ versus $H_a: \beta \neq 15{,}000$.

(c) Compute the test statistic and *P*-value for the test in part (b). What conclusion would you draw at the $\alpha = 0.05$ significance level?

(d) Does the confidence interval in part (a) lead to the same conclusion as the test in part (c)? Explain.

18. **Paired tires** Exercise 71 in Chapter 8 (page 529) compared two methods for estimating tire wear. The first method used the amount of weight lost by a tire. The second method used the amount of wear in the grooves of the tire. A random sample of 16 tires was obtained. Both methods were used to estimate the total distance traveled by each tire. The following scatterplot displays the two estimates (in thousands of miles) for each tire.[12]



Computer output from a least-squares regression analysis of these data is shown below. Assume that the conditions for regression inference are met.

```
Predictor  Coef      SE Coef   T       P
Constant   1.351     2.105      0.64   0.531
Weight     0.79021   0.07104   11.12   0.000

S = 2.62078   R-Sq = 89.8%   R-Sq(adj) = 89.1%
```

(a) Verify that the 99% confidence interval for the slope of the population regression line is (0.5787, 1.0017).

(b) Researchers want to test whether there is a difference in the two methods of estimating tire wear. Explain why the researchers might think that an appropriate pair of hypotheses for this test is $H_0: \beta = 1$ versus $H_a: \beta \neq 1$.

(c) Compute the test statistic and P-value for the test in part (b). What conclusion would you draw at the $\alpha = 0.01$ significance level?

(d) Does the confidence interval in part (a) lead to the same conclusion as the test in part (c)? Explain.

Multiple choice: Select the best answer for Exercises 19 to 24, which are based on the following information.
To determine property taxes, Florida reappraises real estate every year, and the county appraiser's Web site lists the current "fair market value" of each piece of property. Property usually sells for somewhat more than the appraised market value. We collected data on the appraised market values x and actual selling prices y (in thousands of dollars) of a random sample of 16 condominium units in Florida. We checked that the conditions for inference about the slope of the population regression line are met. Here is part of the Minitab output from a least-squares regression analysis using these data.[13]

```
Predictor   Coef       SE Coef   T      P
Constant    127.27     79.49     1.60   0.132
Appraisal     1.0466    0.1126   9.29   0.000

S = 69.7299   R-Sq = 86.1%   R-Sq(adj) = 85.1%
```

19. The equation of the least-squares regression line for predicting selling price from appraised value is

(a) $\widehat{\text{price}} = 79.49 + 0.1126$ (appraised value).

(b) $\widehat{\text{price}} = 0.1126 + 1.0466$ (appraised value).

(c) $\widehat{\text{price}} = 127.27 + 1.0466$ (appraised value).

(d) $\widehat{\text{price}} = 1.0466 + 127.27$ (appraised value).

(e) $\widehat{\text{price}} = 1.0466 + 69.7299$ (appraised value).

20. The slope β of the population regression line describes

(a) the exact increase in the selling price of an individual unit when its appraised value increases by $1000.

(b) the average increase in the appraised value in a population of units when selling price increases by $1000.

(c) the average increase in selling price in a population of units when appraised value increases by $1000.

(d) the average increase in the appraised value in the sample of units when selling price increases by $1000.

(e) the average increase in selling price in the sample of units when the appraised value increases by $1000.

21. Is there convincing evidence that selling price increases as appraised value increases? To answer this question, test the hypotheses

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

(b) $H_0: \beta = 0$ versus $H_a: \beta < 0$.

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

(d) $H_0: \beta > 0$ versus $H_a: \beta = 0$.

(e) $H_0: \beta = 1$ versus $H_a: \beta > 1$.

22. Which of the following is the best interpretation for the value 0.1126 in the computer output?

(a) For each increase of $1000 in appraised value, the average selling price increases by about 0.1126.

(b) When using this model to predict selling price, the predictions will typically be off by about 0.1126.

(c) 11.26% of the variation in selling price is accounted for by the linear relationship between selling price and appraised value.

(d) There is a weak, positive linear relationship between selling price and appraised value.

(e) In repeated samples of size 16, the sample slope will typically vary from the population slope by about 0.1126.

23. A 95% confidence interval for the population slope β is

(a) 1.0466 ± 1.046.

(b) 1.0466 ± 0.2415.

(c) 1.0466 ± 0.2387.

(d) 1.0466 ± 0.2207.

(e) 1.0466 ± 0.2400.

24. Which of the following would have resulted in a violation of the conditions for inference?

(a) If the entire sample was selected from one neighborhood

(b) If the sample size was cut in half

(c) If the scatterplot of x = appraised value and y = selling price did not show a perfect linear relationship

(d) If the histogram of selling prices had an outlier

(e) If the standard deviation of appraised values was different from the standard deviation of selling prices

Exercises 25 to 28 refer to the following setting. Does the color in which words are printed affect your ability to read them? Do the words themselves affect your ability to name the color in which they are printed? Mr. Starnes designed a study to investigate these questions using the 16 students in his AP® Statistics class as subjects. Each student performed two tasks in a random order while a partner timed: (1) read 32 words aloud as quickly as possible, and (2) say the color in which each of 32 words is printed as quickly as possible. Try both tasks for yourself using the word list below.

YELLOW	RED	BLUE	GREEN
RED	GREEN	YELLOW	YELLOW
GREEN	RED	BLUE	BLUE
YELLOW	BLUE	GREEN	RED
BLUE	YELLOW	RED	RED
RED	BLUE	YELLOW	GREEN
BLUE	GREEN	GREEN	BLUE
GREEN	YELLOW	RED	YELLOW

25. **Color words** (4.2) Let's review the design of the study.

(a) Explain why this was an experiment and not an observational study.

(b) Did Mr. Starnes use a completely randomized design or a randomized block design? Why do you think he chose this experimental design?

(c) Explain the purpose of the random assignment in the context of the study.

The data from Mr. Starnes's experiment are shown below. For each subject, the time to perform the two tasks is given to the nearest second.

Subject	Words	Colors	Subject	Words	Colors
1	13	20	9	10	16
2	10	21	10	9	13
3	15	22	11	11	11
4	12	25	12	17	26
5	13	17	13	15	20
6	11	13	14	15	15
7	14	32	15	12	18
8	16	21	16	10	18

26. **Color words** (1.3) Do the data provide evidence of a difference in the average time required to perform the two tasks? Include an appropriate graph and numerical summaries in your answer.

27. **Color words** (9.3) Explain why it is not safe to use paired t procedures to do inference about the difference in the mean time to complete the two tasks.

28. **Color words** (3.1, 3.2, 12.1) Can we use a student's word task time to predict his or her color task time?

(a) Make an appropriate scatterplot to help answer this question. Describe what you see.

(b) Use your calculator to find the equation of the least-squares regression line. Define any symbols you use.

(c) Find and interpret the residual for the student who completed the word task in 9 seconds.

(d) Assume that the conditions for performing inference about the slope of the true regression line are met. The P-value for a test of $H_0: \beta = 0$ versus $H_a: \beta > 0$ is 0.0215. Explain what this value means in context.

Note: John Ridley Stroop is often credited with the discovery in 1935 of the fact that the color in which "color words" are printed interferes with people's ability to identify the color. The so-called Stroop Effect, though, was originally published by German researchers in 1929.

Exercises 29 and 30 refer to the following setting. Yellowstone National Park surveyed a random sample of 1526 winter visitors to the park. They asked each person whether he or she owned, rented, or had never used a snowmobile. Respondents were also asked whether they belonged to an environmental organization (like the Sierra Club). The two-way table summarizes the survey responses.

	Environmental Clubs		
	No	Yes	Total
Never used	445	212	**657**
Snowmobile renter	497	77	**574**
Snowmobile owner	279	16	**295**
Total	**1221**	**305**	**1526**

29. **Snowmobiles** (5.2, 5.3)

(a) If we choose a survey respondent at random, what's the probability that this individual

(i) is a snowmobile owner?

(ii) belongs to an environmental organization or owns a snowmobile?

(iii) has never used a snowmobile given that the person belongs to an environmental organization?

(b) Are the events "is a snowmobile owner" and "belongs to an environmental organization" independent for the members of the sample? Justify your answer.

(c) If we choose two survey respondents at random, what's the probability that

(i) both are snowmobile owners?

(ii) at least one of the two belongs to an environmental organization?

30. **Snowmobiles** (11.2) Do these data provide convincing evidence at the 5% significance level of an association between environmental club membership and snowmobile use for the population of visitors to Yellowstone National Park? Justify your answer.

12.2 Transforming to Achieve Linearity

WHAT YOU WILL LEARN By the end of the section, you should be able to:

- Use transformations involving powers and roots to find a power model that describes the relationship between two variables, and use the model to make predictions.
- Use transformations involving logarithms to find a power model or an exponential model that describes the relationship between two variables, and use the model to make predictions.
- Determine which of several transformations does a better job of producing a linear relationship.

In Chapter 3, we learned how to analyze relationships between two quantitative variables that showed a linear pattern. When two-variable data show a curved relationship, we must develop new techniques for finding an appropriate model. This section describes several simple *transformations* of data that can straighten a nonlinear pattern. Once the data have been transformed to achieve linearity, we can use least-squares regression to generate a useful model for making predictions. And if the conditions for regression inference are met, we can estimate or test a claim about the slope of the population (true) regression line using the transformed data.

EXAMPLE Health and Wealth

Straightening out a curved pattern

The Gapminder Web site, www.gapminder.org, provides loads of data on the health and well-being of the world's inhabitants. Figure 12.9 on the next page is a scatterplot of data from Gapminder.[14] The individuals are all the world's nations for which data are available. The explanatory variable is a measure of how rich a country is: income per person. The response variable is life expectancy at birth.

We expect people in richer countries to live longer because they have better access to medical care and typically lead healthier lives. The overall pattern of the scatterplot does show this, but the relationship is not linear. Life expectancy rises very quickly as income per person increases and then levels off. People in very rich countries such as the United States live no longer than people in poorer but not extremely poor nations. In some less wealthy countries, people live longer than in the United States.

Four African nations are outliers. Their life expectancies are similar to those of their neighbors, but their income per person is higher. Gabon and Equatorial Guinea produce oil, and South Africa and Botswana produce diamonds. It may be that income from mineral exports goes mainly to a few people and so pulls up income per person without much effect on either the income or the life expectancy of ordinary citizens. That is, income per person is a mean, and we know that mean income can be much higher than median income.

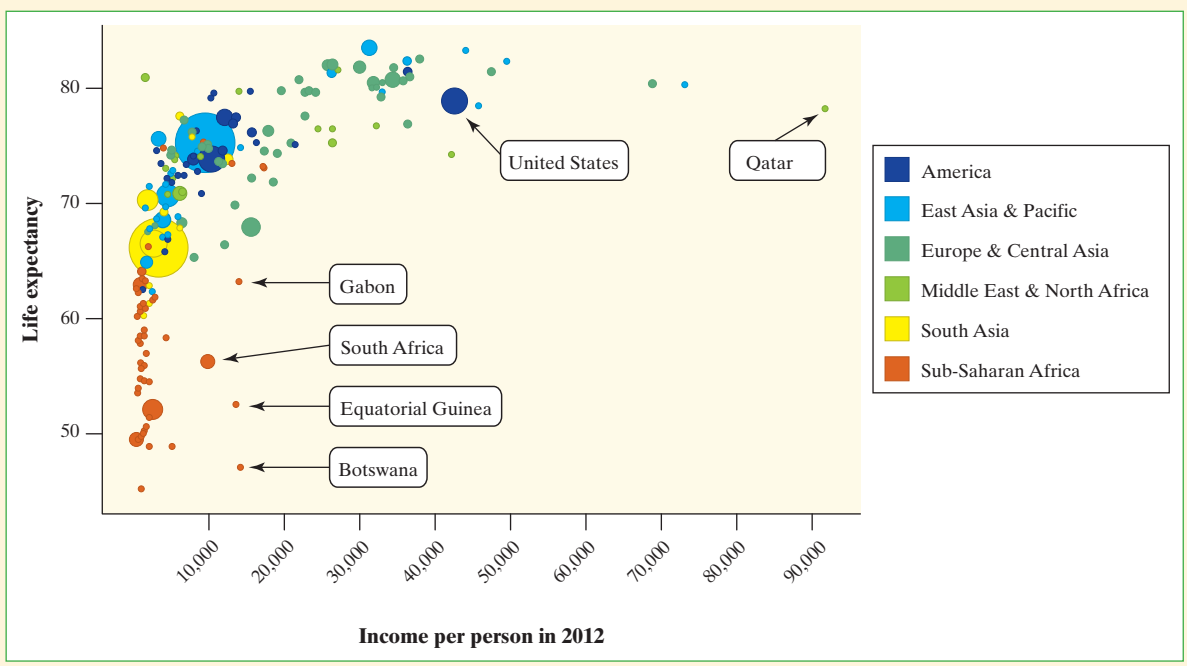


FIGURE 12.9 Scatterplot of the life expectancy of people in many nations against each nation's income per person. The color of each circle indicates the geographic region in which that country is located. The size of each circle is based on the population of the country—bigger circles indicate larger populations.

The scatterplot in Figure 12.9 shows a curved pattern. We can straighten things out using logarithms. Figure 12.10 (on the facing page) plots the logarithm of income per person against life expectancy for these same countries. The effect is almost magical. This graph has a clear, linear pattern.

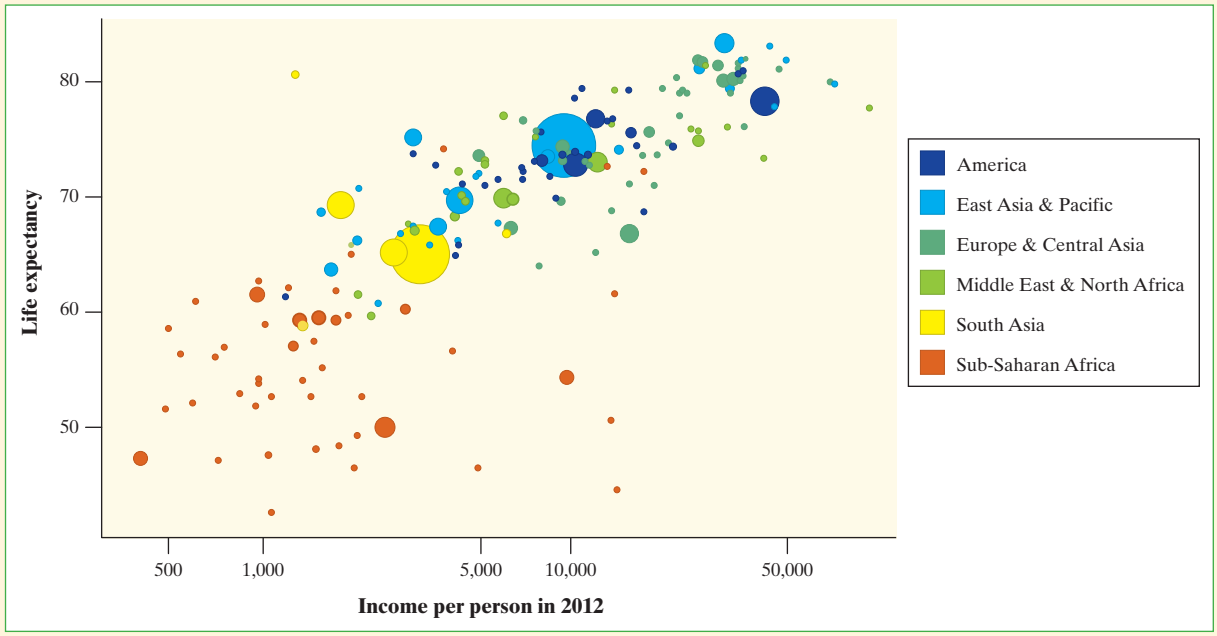


FIGURE 12.10 Scatterplot of life expectancy against income per person (on a logarithm scale) for many nations.

Applying a function such as the logarithm or square root to a quantitative variable is called **transforming** the data. We will see in this section that understanding how simple functions work helps us choose and use transformations to straighten nonlinear patterns.

Transforming data amounts to changing the scale of measurement that was used when the data were collected. We can choose to measure temperature in degrees Fahrenheit or in degrees Celsius, distance in miles or in kilometers. These changes of units are *linear transformations*, discussed in Chapter 2.

Linear transformations cannot straighten a curved relationship between two variables. To do that, we resort to functions that are not linear. The logarithm function, applied in the "Health and Wealth" example, is a nonlinear function. We'll return to transformations involving logarithms later.

Transforming with Powers and Roots

When you visit a pizza parlor, you order a pizza by its diameter—say, 10 inches, 12 inches, or 14 inches. But the amount you get to eat depends on the area of the pizza. The area of a circle is π times the square of its radius r. So the area of a round pizza with diameter x is

$$\text{area} = \pi r^2 = \pi\left(\frac{x}{2}\right)^2 = \pi\left(\frac{x^2}{4}\right) = \frac{\pi}{4}x^2$$

This is a **power model** of the form $y = ax^p$ with $a = \pi/4$ and $p = 2$.

When we are dealing with things of the same general form, whether circles or fish or people, we expect area to go up with the square of a dimension such as diameter or height. Volume should go up with the cube of a linear dimension. That is, geometry tells us to expect power models in some settings. There are other physical relationships between two variables that are described by power models. Here are some examples from science.

- The distance that an object dropped from a given height falls is related to time since release by the model

$$\text{distance} = a(\text{time})^2$$

- The time it takes a pendulum to complete one back-and-forth swing (its period) is related to its length by the model

$$\text{period} = a\sqrt{\text{length}} = a(\text{length})^{1/2}$$

- The intensity of a light bulb is related to distance from the bulb by the model

$$\text{intensity} = \frac{a}{\text{distance}^2} = a(\text{distance})^{-2}$$

Although a power model of the form $y = ax^p$ describes the relationship between x and y in each of these settings, there is a *linear* relationship between x^p and y. If we transform the values of the explanatory variable x by raising them to the p power, and graph the points (x^p, y), the scatterplot should have a linear form. The following example shows what we mean.

EXAMPLE Go Fish!

Transforming with powers

Imagine that you have been put in charge of organizing a fishing tournament in which prizes will be given for the heaviest Atlantic Ocean rockfish caught. You know that many of the fish caught during the tournament will be measured and released. You are also aware that using delicate scales to try to weigh a fish that is flopping around in a moving boat will probably not yield very accurate results. It would be much easier to measure the length of the fish while on the boat. What you need is a way to convert the length of the fish to its weight.

You contact the nearby marine research laboratory, and they provide reference data on the length (in centimeters) and weight (in grams) for Atlantic Ocean rockfish of several sizes.[15]

Length:	5.2	8.5	11.5	14.3	16.8	19.2	21.3	23.3	25.0	26.7
Weight:	2	8	21	38	69	117	148	190	264	293

Length:	28.2	29.6	30.8	32.0	33.0	34.0	34.9	36.4	37.1	37.7
Weight:	318	371	455	504	518	537	651	719	726	810

Figure 12.11 is a scatterplot of the data. Note the clear curved shape.

Because length is one-dimensional and weight (like volume) is three-dimensional, a power model of the form weight = a (length)^3 should describe the relationship. What happens if we cube the lengths in the data table and then graph weight versus length^3? Figure 12.12 gives us the answer. This transformation of the explanatory variable helps us produce a graph that is quite linear.

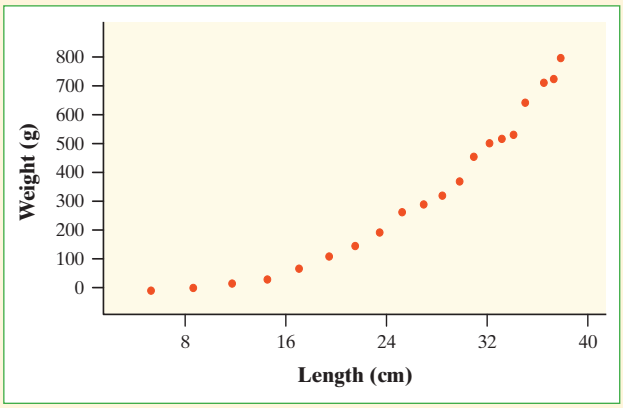


FIGURE 12.11 Scatterplot of Atlantic Ocean rockfish weight versus length.

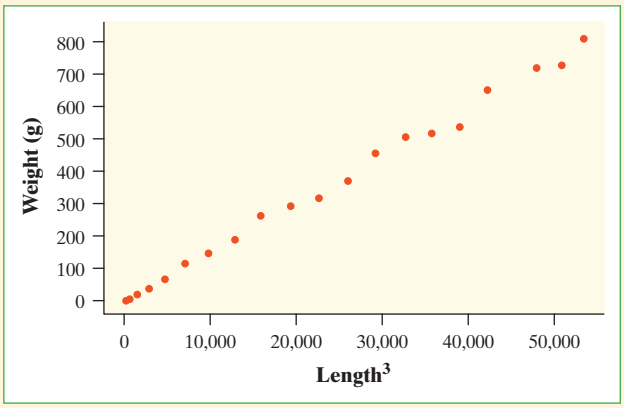


FIGURE 12.12 The scatterplot of weight versus length^3 is linear.

There's another way to transform the data in the example to achieve linearity. We can take the cube root of the weight values and graph $\sqrt[3]{\text{weight}}$ versus length. Figure 12.13 shows that the resulting scatterplot has a linear form. Why does this transformation work? Start with weight = a(length)^3 and take the cube root of both sides of the equation:

$$\sqrt[3]{\text{weight}} = \sqrt[3]{a(\text{length})^3}$$

$$\sqrt[3]{\text{weight}} = \sqrt[3]{a}(\text{length})$$

That is, there is a linear relationship between length and $\sqrt[3]{\text{weight}}$.

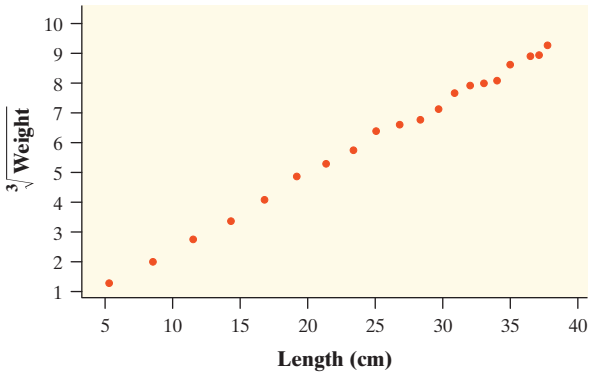


FIGURE 12.13 The scatterplot of $\sqrt[3]{\text{weight}}$ versus length is linear.

Once we straighten out the curved pattern in the original scatterplot, we fit a least-squares line to the transformed data. This linear model can be used to predict values of the response variable y. As in Chapter 3, a residual plot tells us if the linear model is appropriate. The values of s and r^2 tell us how well the regression line fits the data.

EXAMPLE Go Fish!

Transforming with Powers and Roots

Here is Minitab output from separate regression analyses of the two sets of transformed Atlantic Ocean rockfish data.

Transformation 1: (length^3, weight)

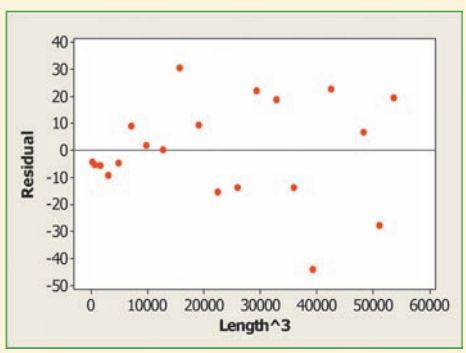
Predictor	Coef	SE Coef	T	P
Constant	4.066	6.902	0.59	0.563
Length^3	0.0146774	0.0002404	61.07	0.000

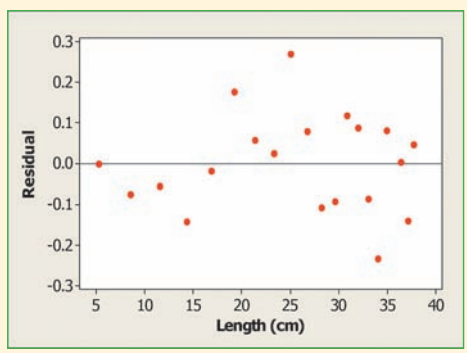
S = 18.8412 R-Sq = 99.5% R-Sq(adj) = 99.5%

Transformation 2: (length, $\sqrt[3]{\text{weight}}$)

Predictor	Coef	SE Coef	T	P
Constant	−0.02204	0.07762	−0.28	0.780
Length	0.246616	0.002868	86.00	0.000

S = 0.124161 R-Sq = 99.8% R-Sq(adj) = 99.7%





PROBLEM: Do each of the following for *both* transformations.

(a) Give the equation of the least-squares regression line. Define any variables you use.

(b) Suppose a contestant in the fishing tournament catches an Atlantic Ocean rockfish that's 36 centimeters long. Use the model from part (a) to predict the fish's weight. Show your work.

SOLUTION:

(a) Transformation 1: $\widehat{\text{weight}} = 4.066 + 0.0146774\,(\text{length}^3)$

Transformation 2: $\widehat{\sqrt[3]{\text{weight}}} = -0.02204 + 0.246616\,(\text{length})$

(b) Transformation 1: $\widehat{\text{weight}} = 4.066 + 0.0146774(36^3) = 688.9$ grams

Transformation 2: $\widehat{\sqrt[3]{\text{weight}}} = -0.02204 + 0.246616(36) = 8.856$

$$\widehat{\text{weight}} = 8.856^3 = 694.6 \text{ grams}$$

For Practice *Try Exercise* 33

When experience or theory suggests that the relationship between two variables is described by a power model of the form $y = ax^p$, you now have two strategies for transforming the data to achieve linearity.

1. Raise the values of the explanatory variable x to the p power and plot the points (x^p, y).
2. Take the pth root of the values of the response variable y and plot the points $(x, \sqrt[p]{y})$

What if you have no idea what power to choose? You could guess and test until you find a transformation that works. Some technology comes with built-in sliders that allow you to dynamically adjust the power and watch the scatterplot change shape as you do.

It turns out that there is a much more efficient method for linearizing a curved pattern in a scatterplot. Instead of transforming with powers and roots, we use logarithms. This more general method works when the data follow an unknown power model or any of several other common mathematical models.

Transforming with Logarithms

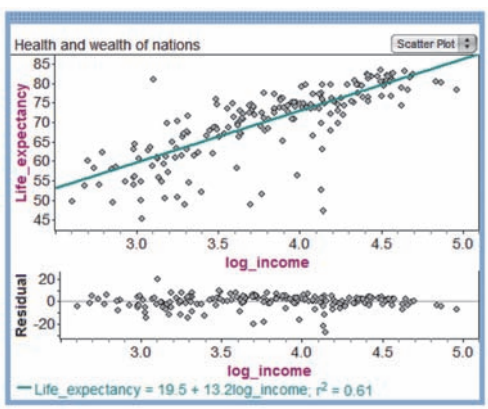


FIGURE 12.14 Scatterplot with least-squares line added and residual plot from Fathom for the transformed data about the health and wealth of nations.

Not all curved relationships are described by power models. For instance, in the "Health and Wealth" example (page 765), a graph of life expectancy versus the logarithm (base 10) of income per person showed a linear pattern. We used Fathom software to fit a least-squares regression line to the transformed data and to make a residual plot. Figure 12.14 shows the results.

The regression line is

$$\text{predicted life expectancy} = 19.5 + 13.2 \log(\text{income})$$

How well does this model fit the data? The residual plot shows a random scatter of prediction errors about the residual = 0 line. Also, because $r^2 = 0.61$, about 61% of the variation in life expectancy is accounted for by the linear model using log(income) as the explanatory variable.

The relationship between life expectancy and income per person is described by a logarithmic model of the form $y = a + b \log x$. We can use this model to predict how long a country's citizens will live from how much money they make. For the United States, which has income per person of $42,296.20,

$$\text{predicted life expectancy} = 19.5 + 13.2 \log(42{,}296.20) = 80.567 \text{ years}$$

The actual U.S. life expectancy in 2012 was 78.80 years.

Taking the logarithm of the income per person values straightened out the curved pattern in the original scatterplot. The logarithm transformation can also help achieve linearity when the relationship between two variables is described by a *power model* or an *exponential model*.

Power Models Biologists have found that many characteristics of living things are described quite closely by power models. There are more mice than elephants, and more flies than mice—the abundance of species follows a power model with body weight as the explanatory variable. So do pulse rate, length of life, the number of eggs a bird lays, and so on.

Sometimes the powers can be predicted from geometry, but sometimes they are mysterious. Why, for example, does the rate at which animals use energy go up as the 3/4 power of their body weight? Biologists call this relationship Kleiber's law. It has been found to work all the way from bacteria to whales. The search goes on for some physical or geometrical explanation for why life follows power laws.

To achieve linearity from a power model, we apply the logarithm transformation to *both* variables. Here are the details:

1. A power model has the form $y = ax^p$, where a and p are constants.
2. Take the logarithm of both sides of this equation. Using properties of logarithms, we get

$$\log y = \log(ax^p) = \log a + \log(x^p) = \log a + p \log x$$

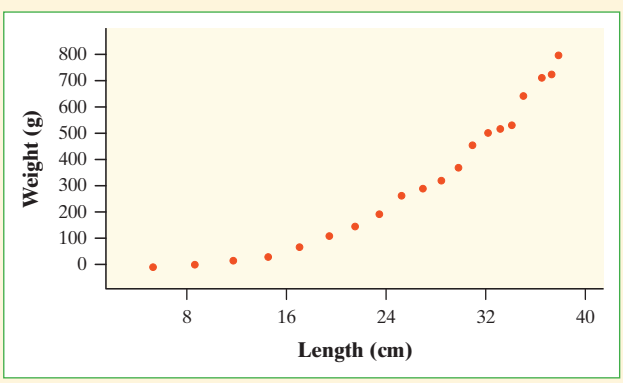
The equation $\log y = \log a + p \log x$ shows that taking the logarithm of both variables results in a linear relationship between $\log x$ and $\log y$.

3. Look carefully: the *power* p in the power model becomes the *slope* of the straight line that links $\log y$ to $\log x$.

If a power model describes the relationship between two variables, a scatterplot of the logarithms of both variables should produce a linear pattern. Then we can fit a least-squares regression line to the transformed data and use the linear model to make predictions. Here's an example.

EXAMPLE Go Fish!

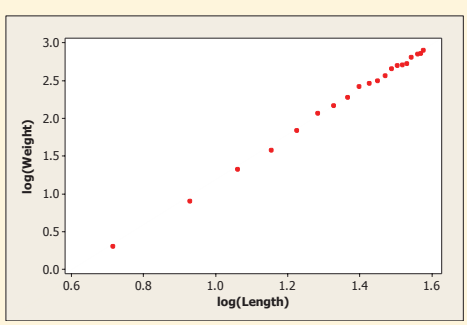
Transforming with logarithms

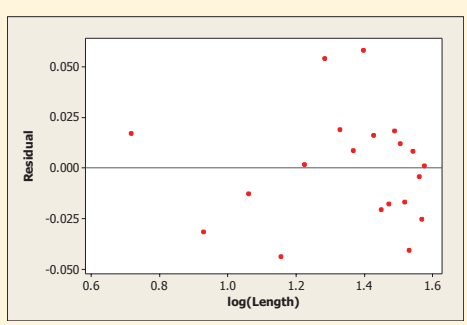


Let's return to the fishing tournament from the previous example. Our goal remains the same: to find a model for predicting the weight of an Atlantic Ocean rockfish from its length. We still expect a power model of the form weight = a(length)^3 based on geometry. Here once again is a scatterplot of the data from the local marine research lab.

Earlier, we transformed the data in two ways to try to achieve linearity: (1) cubing the length values and (2) taking the cube root of the weight values. This time we'll use logarithms.

We took the logarithm (base 10) of the values for both variables. Some computer output from a linear regression analysis on the transformed data is shown below.





Regression Analysis: log(Weight) versus log(Length)

Predictor	Coef	SE Coef	T	P
Constant	−1.89940	0.03799	−49.99	0.000
log(Length)	3.04942	0.02764	110.31	0.000

S = 0.0281823 R-Sq = 99.9% R-Sq(adj) = 99.8%

PROBLEM:

(a) Based on the output, explain why it would be reasonable to use a power model to describe the relationship between weight and length for Atlantic Ocean rockfish.

(b) Give the equation of the least-squares regression line. Be sure to define any variables you use.

SOLUTION:

(a) If a power model describes the relationship between two variables x and y, then a *linear* model should describe the relationship between log x and log y. The scatterplot of log(weight) versus log(length) has a linear form, and the residual plot shows a fairly random scatter of points about the residual = 0 line. So a power model seems reasonable here.

(b) $\widehat{\log(\text{weight})} = -1.89940 + 3.04942 \log(\text{length})$

For Practice *Try Exercise* 35

On the TI-83/84, you can "undo" the logarithm using the 2nd function keys. To solve log $y = 2$, press 2nd LOG 2 ENTER. To solve ln $y = 2$, press 2nd LN 2 ENTER.

If we fit a least-squares regression line to the transformed data, we can find the predicted value of the logarithm of y for any value of the explanatory variable x by substituting our x-value into the equation of the line. To obtain the corresponding prediction for the response variable y, we have to "undo" the logarithm transformation to return to the original units of measurement. One way of doing this is to use the definition of a logarithm as an exponent:

$$\log_b a = x \Rightarrow b^x = a$$

For instance, if we have $\log y = 2$, then

$$\log y = 2 \Rightarrow \log_{10} y = 2 \Rightarrow 10^2 = y \Rightarrow 100 = y$$

If instead we have $\ln y = 2$, then

$$\ln y = 2 \Rightarrow \log_e y = 2 \Rightarrow e^2 = y \Rightarrow 7.389 = y$$

EXAMPLE Go Fish!

Making predictions

PROBLEM: Suppose a contestant in the fishing tournament catches an Atlantic Ocean rockfish that's 36 centimeters long. Use the model from part (b) of the previous example to predict the fish's weight. Show your work.

SOLUTION: For a length of 36 centimeters, we have

$$\widehat{\log(\text{weight})} = -1.89940 + 3.04942 \log(36) = 2.8464$$

To find the predicted weight, we use the definition of a logarithm as an exponent:

$$\widehat{\log_{10}(\text{weight})} = 2.8464$$

$$\widehat{\text{weight}} = 10^{2.8464} \approx 702.1$$

This model predicts that a 36-centimeter-long rockfish will weigh about 702 grams.

For Practice *Try Exercise* 37

Your calculator and most statistical software will calculate the logarithms of all the values of a variable with a single command. The important thing to remember is this: if the relationship between two variables is described by a power model, then we can linearize the relationship by taking the logarithm of *both* the explanatory and response variables.

THINK ABOUT IT

How do we find the power model for predicting *y* from *x*? The least-squares line for the transformed rockfish data is

$$\widehat{\log(\text{weight})} = -1.89940 + 3.04942 \log(\text{length})$$

If we use the definition of the logarithm as an exponent, we can rewrite this equation as

$$\widehat{\text{weight}} = 10^{-1.89940 + 3.04942\log(\text{length})}$$

Using properties of exponents, we can simplify this as follows:

$$\widehat{\text{weight}} = 10^{-1.89940} \cdot 10^{3.04942\log(\text{length})} \quad \text{using the fact that } b^m b^n = b^{m+n}$$

$$\widehat{\text{weight}} = 10^{-1.89940} \cdot 10^{\log(\text{length})^{3.04942}} \quad \text{using the fact that } p \log x = \log x^p$$

$$\widehat{\text{weight}} = 0.0126(\text{length})^{3.04942} \quad \text{using the fact that } 10^{\log x} = x$$

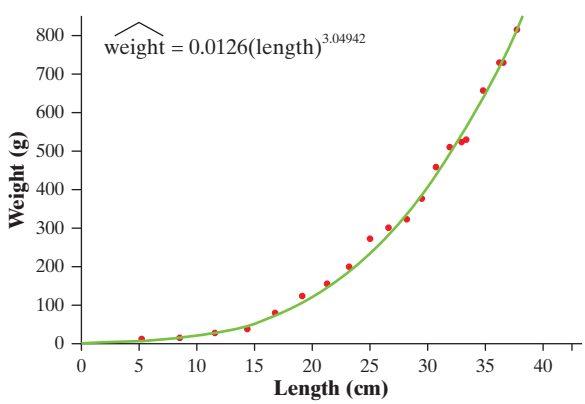


FIGURE 12.15 Rockfish data with power model.

This equation is now in the familiar form of a power model $y = ax^b$ with $a = 0.0126$ and $b = 3.04942$. Notice how close the power is to 3, as expected from geometry.

We could use the power model to predict the weight of a 36-centimeter-long Atlantic Ocean rockfish:

$$\widehat{\text{weight}} = 0.0126(36)^{3.04942} \approx 701.76 \text{ grams}$$

This is the same prediction we got earlier (up to rounding). The scatterplot of the original rockfish data with the power model added appears in Figure 12.15. Note how well this model fits the data!

Exponential Models A linear model has the form $y = a + bx$. The value of y increases (or decreases) at a constant rate as x increases. The slope b describes the constant rate of change of a linear model. That is, for each 1 unit increase in x, the model predicts an increase of b units in y. You can think of a linear model as describing the repeated addition of a constant amount. Sometimes the relationship between y and x is based on repeated *multiplication* by a constant factor. That

is, each time x increases by 1 unit, the value of y is multiplied by b. An **exponential model** of the form $y = ab^x$ describes such multiplicative growth.

Populations of living things tend to grow exponentially if not restrained by outside limits such as lack of food or space. More pleasantly (unless we're talking about credit card debt!), money also displays exponential growth when interest is compounded each time period. Compounding means that last period's income earns income in the next period.

EXAMPLE Money, Money, Money

Understanding exponential growth

Suppose that you invest \$100 in a savings account that pays 6% interest compounded annually. After a year, you will have earned \$100(0.06) = \$6.00 in interest. Your new account balance is the initial deposit plus the interest earned: \$100 + (\$100)(0.06), or \$106. We can rewrite this as \$100(1 + 0.06), or more simply as \$100(1.06). That is, 6% annual interest means that any amount on deposit for the entire year is multiplied by 1.06.

If you leave the money invested for a second year, your new balance will be $[\$100(1.06)](1.06) = \$100(1.06)^2 = \$112.36$. Notice that you earn \$6.36 in interest during the second year. That's another \$6 in interest from your initial \$100 deposit plus the interest on your \$6 interest earned for Year 1. After x years, your account balance y is given by the exponential model $y = 100(1.06)^x$.

The table below shows the balance in your savings account at the end of each of the first six years. Figure 12.16 shows the growth in your investment over 100 years. It is characteristic of exponential growth that the increase appears slow for a long period and then seems to explode.

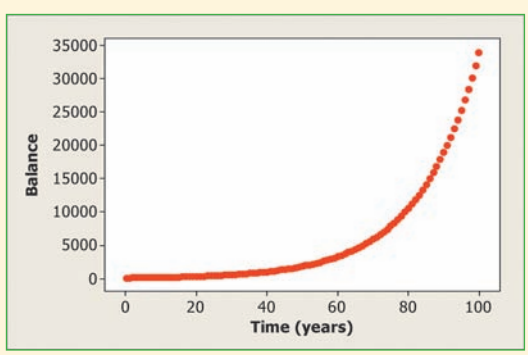


FIGURE 12.16 Scatterplot of the growth of a \$100 investment in a savings account paying 6% interest, compounded annually.

Time x (years)	Account balance y
0	\$100.00
1	\$106.00
2	\$112.36
3	\$119.10
4	\$126.25
5	\$133.82
6	\$141.85
…	…

If an exponential model of the form $y = ab^x$ describes the relationship between x and y, we can use logarithms to transform the data to produce a linear relationship. Start by taking the logarithm (we'll use base 10, but the natural logarithm ln

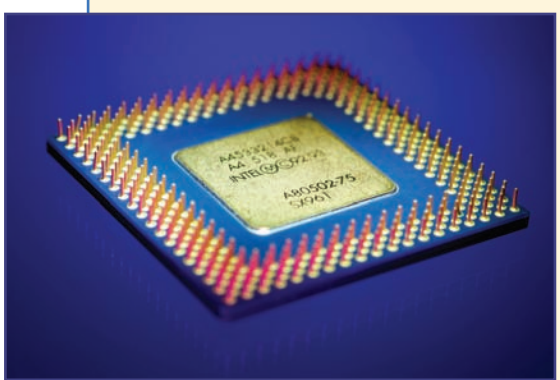
using base e would work just as well). Then use algebraic properties of logarithms to simplify the resulting expressions. Here are the details:

$$\log y = \log (ab^x) \qquad \text{taking the logarithm of both sides}$$
$$\log y = \log a + \log (b^x) \qquad \text{using the property } \log(mn) = \log m + \log n$$
$$\log y = \log a + x \log b \qquad \text{using the property } \log m^p = p \log m$$

We can then rearrange the final equation as $\log y = \log a + (\log b)x$. Notice that $\log a$ and $\log b$ are constants because a and b are constants. So the equation gives a linear model relating the explanatory variable x to the transformed variable $\log y$. Thus, if the relationship between two variables follows an exponential model, and we plot the logarithm (base 10 or base e) of y against x, we should observe a *straight-line pattern* in the transformed data.

EXAMPLE Moore's Law and Computer Chips

Logarithm transformations and exponential models



Gordon Moore, one of the founders of Intel Corporation, predicted in 1965 that the number of transistors on an integrated circuit chip would double every 18 months. This is Moore's law, one way to measure the revolution in computing. Here are data on the dates and number of transistors for Intel microprocessors:[16]

Processor	Date	Transistors
4004	1971	2,250
8008	1972	2,500
8080	1974	5,000
8086	1978	29,000
286	1982	120,000
386	1985	275,000
486 DX	1989	1,180,000
Pentium	1993	3,100,000
Pentium II	1997	7,500,000
Pentium III	1999	24,000,000
Pentium 4	2000	42,000,000
Itanium 2	2003	220,000,000
Itanium 2 w/9MB cache	2004	592,000,000
Dual-core Itanium 2	2006	1,700,000,000
Six-core Xeon 7400	2008	1,900,000,000
8-core Xeon Nehalem-EX	2010	2,300,000,000

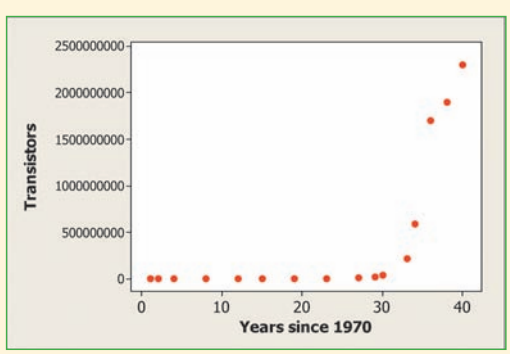
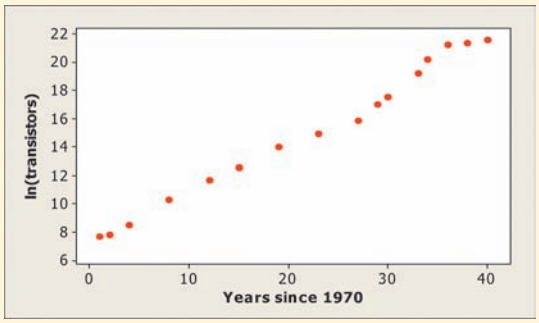


FIGURE 12.17 Scatterplot of the number of transistors on a computer chip from 1971 to 2010.

Figure 12.17 shows the growth in the number of transistors on a computer chip from 1971 to 2010. Notice that we used "years since 1970" as the explanatory variable. We'll explain this later. If Moore's law is correct, then an exponential model should describe the relationship between the variables.



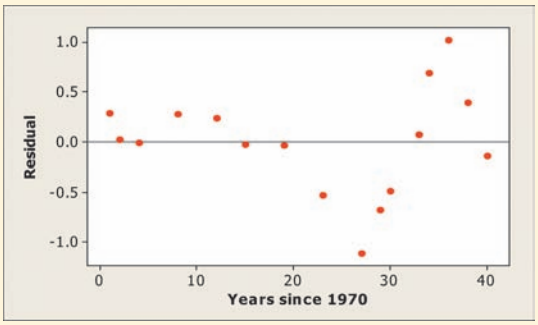
PROBLEM:

(a) A scatterplot of the natural logarithm (log base e or ln) of the number of transistors on a computer chip versus years since 1970 is shown. Based on this graph, explain why it would be reasonable to use an exponential model to describe the relationship between number of transistors and years since 1970.

(b) Minitab output from a linear regression analysis on the transformed data is shown below. Give the equation of the least-squares regression line. Be sure to define any variables you use.

Predictor	Coef	SE Coef	T	P
Constant	7.0647	0.2672	26.44	0.000
Years since 1970	0.36583	0.01048	34.91	0.000

S = 0.544467 R-Sq = 98.9% R-Sq(adj) = 98.8%



(c) Use your model from part (b) to predict the number of transistors on an Intel computer chip in 2020. Show your work.

(d) A residual plot for the linear regression in part (b) is shown at left. Discuss what this graph tells you about the appropriateness of the model.

SOLUTION:

(a) If an exponential model describes the relationship between two variables x and y, then we expect a scatterplot of $(x, \ln y)$ to be roughly linear. The scatterplot of ln(transistors) versus years since 1970 has a fairly linear pattern, especially through the year 2000. So an exponential model seems reasonable here.

(b) $\widehat{\ln(\text{transistors})} = 7.0647 + 0.36583(\text{years since } 1970)$

(c) Because 2020 is 50 years since 1970, we have

$$\widehat{\ln(\text{transistors})} = 7.0647 + 0.36583(50) = 25.3562$$

To find the predicted number of transistors, we use the definition of a logarithm as an exponent:

$$\widehat{\ln(\text{transistors})} = 25.3562 \Rightarrow \widehat{\log_e(\text{transistors})} = 25.3562$$

$$\widehat{\text{transistors}} = e^{25.362} \approx 1.028 \cdot 10^{11}$$

This model predicts that an Intel chip made in 2020 will have about 100 billion transistors.

(d) The residual plot shows a distinct pattern, with the residuals going from positive to negative to positive as we move from left to right. But the residuals are small in size relative to the transformed y-values. Also, the scatterplot of the transformed data is much more linear than the original scatterplot. We feel reasonably comfortable using this model to make predictions about the number of transistors on a computer chip.

For Practice *Try Exercise* 41

Make sure that you understand the big idea here. The necessary transformation is carried out by taking the logarithm of the response variable. The crucial property of the logarithm for our purposes is that *if a variable grows exponentially, its logarithm grows linearly.*

THINK ABOUT IT

How do we find the exponential model for predicting y from x?

The least-squares line for the transformed data in the computer chip example is

$$\widehat{\ln(\text{transistors})} = 7.0647 + 0.36583\,(\text{years since } 1970)$$

If we use the definition of the logarithm as an exponent, we can rewrite this equation as

$$\widehat{\text{transistors}} = e^{7.0647 + 0.36583(\text{years since } 1970)}$$

Using properties of exponents, we can simplify this as follows:

$$\widehat{\text{transistors}} = e^{7.0647} \cdot e^{0.36583(\text{years since } 1970)} \quad \text{using the fact that } b^m b^n = b^{m+n}$$

$$\widehat{\text{transistors}} = e^{7.0647} \cdot (e^{0.36583})^{(\text{years since } 1970)} \quad \text{using the fact that } (b^m)^n = b^{mn}$$

$$\widehat{\text{transistors}} = 1169.93 \cdot (1.44171)^{(\text{years since } 1970)} \quad \text{simplifying}$$

This equation is now in the familiar form of an exponential model $y = ab^x$ with $a = 1169.93$ and $b = 1.44171$.

We could use the exponential model to predict the number of transistors on an Intel chip in 2020: $\widehat{\text{transistors}} = 1169.93(1.44171)^{50} \approx 1.0281 \cdot 10^{11}$. This is the same prediction we got earlier. How does this compare with the prediction from Moore's law? Suppose the number of transistors on an Intel computer chip doubles every 18 months (1.5 years). Then in the 49 years from 1971 to 2020, the number of transistors would double $49/1.5 = 32.67$ times. So the predicted number of transistors on an Intel chip in 2020 would be

$$\widehat{\text{transistors}} = 2250(2)^{32.67} = 1.54 \cdot 10^{13}$$

Moore's law predicts more rapid exponential growth than our model does.

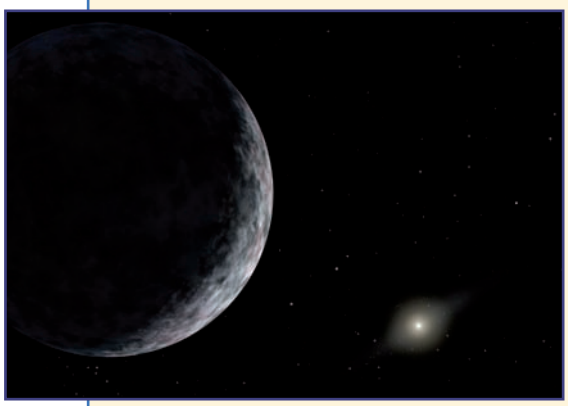
The calculation at the end of the Think about It feature might give you some idea of why we used years since 1970 as the explanatory variable in the example. To make a prediction, we substituted the value $x = 50$ into the equation for the exponential model. This value is the exponent in our calculation. If we had used years as the explanatory variable, our exponent would have been 2020. Such a large exponent can lead to overflow errors on a calculator.

Putting It All Together: Which Transformation Should We Choose?

Suppose that a scatterplot shows a curved relationship between two quantitative variables x and y. How can we decide whether a power model or an exponential model better describes the relationship? The following example shows the strategy we should use.

EXAMPLE What's a Planet, Anyway?

Power models and logarithm transformations



On July 31, 2005, a team of astronomers announced that they had discovered what appeared to be a new planet in our solar system. They had first observed this object almost two years earlier using a telescope at Caltech's Palomar Observatory in California. Originally named UB313, the potential planet is bigger than Pluto and has an average distance of about 9.5 billion miles from the sun. (For reference, Earth is about 93 million miles from the sun.) Could this new astronomical body, now called Eris, be a new planet?

At the time of the discovery, there were nine known planets in our solar system. Here are data on the distance from the sun and period of revolution of those planets. Note that distance is measured in astronomical units (AU), the number of Earth distances the object is from the sun.[17]

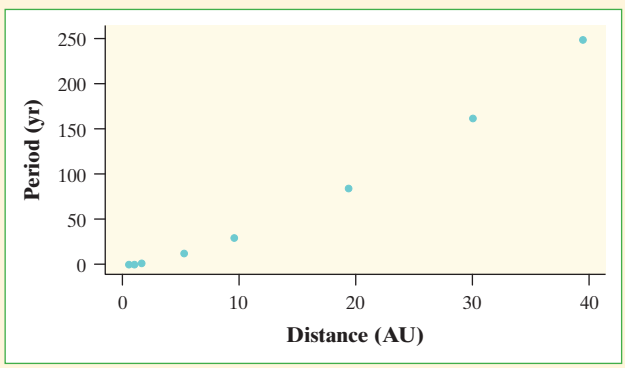


FIGURE 12.18 Scatterplot of planetary distance from the sun and period of revolution.

Planet	Distance from sun (astronomical units)	Period of revolution (Earth years)
Mercury	0.387	0.241
Venus	0.723	0.615
Earth	1.000	1.000
Mars	1.524	1.881
Jupiter	5.203	11.862
Saturn	9.539	29.456
Uranus	19.191	84.070
Neptune	30.061	164.810
Pluto	39.529	248.530

Figure 12.18 is a scatterplot of the planetary data. There appears to be a strong curved relationship between distance from the sun and period of revolution.

In August 2006, the International Astronomical Union agreed on a new definition of "planet." Both Pluto and Eris were classified as "dwarf planets."

PROBLEM: **The graphs below show the results of two different transformations of the data. Figure 12.19(a) plots the natural logarithm of period against distance from the sun for all nine planets. Figure 12.19(b) plots the natural logarithm of period against the natural logarithm of distance from the sun for the nine planets.**

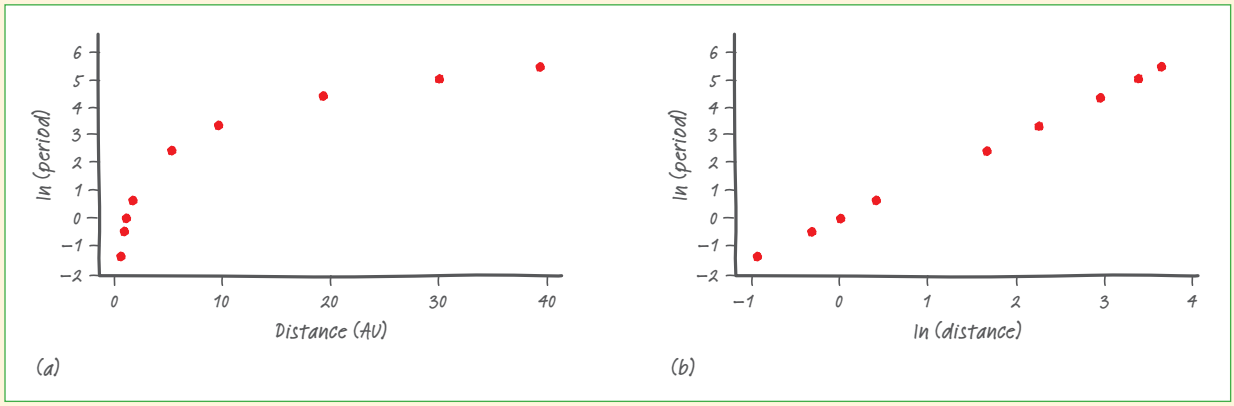


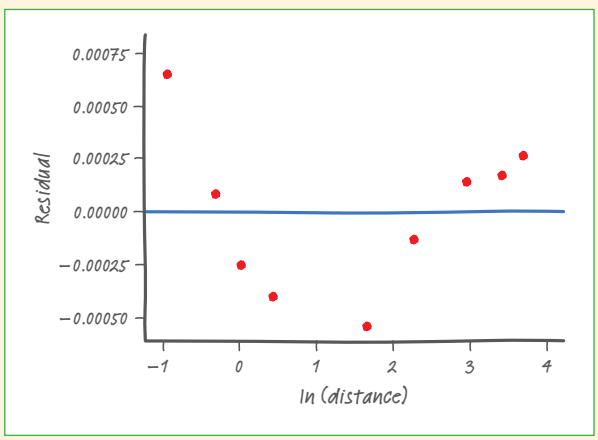
FIGURE 12.19 (a) A scatterplot of ln(period) versus distance. (b) A scatterplot of ln(period) versus ln(distance).

(a) Explain why a power model would provide a more appropriate description of the relationship between period of revolution and distance from the sun than an exponential model.

(b) Minitab output from a linear regression analysis on the transformed data in Figure 12.19(b) is shown below. Give the equation of the least-squares regression line. Be sure to define any variables you use.

Predictor	Coef	SE Coef	T	P
Constant	0.0002544	0.0001759	1.45	0.191
ln(distance)	1.49986	0.00008	18598.27	0.000

S = 0.000393364 R-Sq = 100.0% R-Sq(adj) = 100.0%



(c) Use your model from part (b) to predict the period of revolution for Eris, which is 9,500,000,000/93,000,000 = 102.15 AU from the sun. Show your work.

(d) A residual plot for the linear regression in part (b) is shown at left. Do you expect your prediction in part (c) to be too high, too low, or about right? Justify your answer.

SOLUTION:

(a) The scatterplot of ln(period) versus distance is clearly curved, so an exponential model would not be appropriate. However, the graph of ln(period) versus ln(distance) has a strong linear pattern, indicating that a power model would be more appropriate.

(b) $\widehat{\ln(\text{period})} = 0.0002544 + 1.49986 \ln(\text{distance})$

(c) Eris's average distance from the sun is 102.15 AU. Using this value for distance in our model from part (b) gives

$$\widehat{\ln(\text{period})} = 0.0002544 + 1.49986 \ln(102.15) = 6.939$$

To predict the period, we have to undo the logarithm transformation:

$$\widehat{\text{period}} = e^{6.939} \approx 1032 \text{ years}$$

We wouldn't want to wait for Eris to make a full revolution to see if our prediction is accurate!

(d) Eris's value for ln(distance) is ln(102.15) = 4.626, which would fall at the far right of the residual plot, where all the residuals are positive. Because residual = actual y − predicted y seems likely to be positive, we would expect our prediction to be too low.

For Practice *Try Exercise* 43

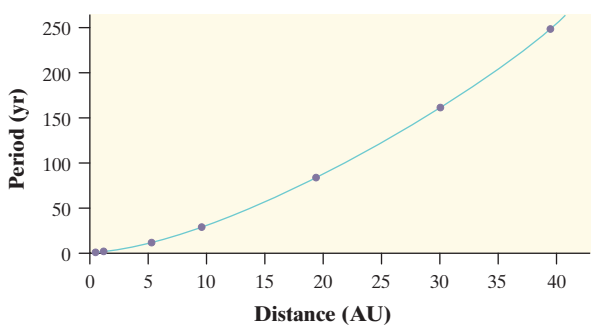


FIGURE 12.20 Planetary data with power model.

The scatterplot of the original data with the power model added appears in Figure 12.20. It seems remarkable that period of revolution is closely related to the 1.5 power of distance from the sun. Johannes Kepler made this fascinating discovery about 400 years ago without the aid of modern technology—a result known as Kepler's third law.

What if the scatterplots of (log x, log y) and (x, log y) *both* look linear? Fit a least-squares regression line to both sets of transformed data. Then compare residual plots and look for the one with the most random scatter. If the residual plots look roughly the same, use the values of s and r^2 to decide whether a power model or an exponential model is a better choice.

We have used statistical software to do all the transformations and linear regression analysis in this section so far. Now let's look at how the process works on a graphing calculator.

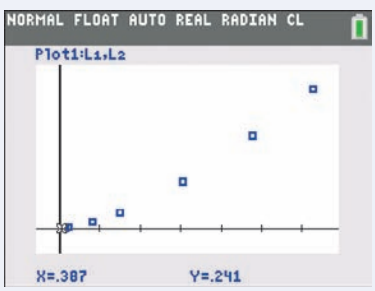
30. TECHNOLOGY CORNER TRANSFORMING TO ACHIEVE LINEARITY ON THE CALCULATOR

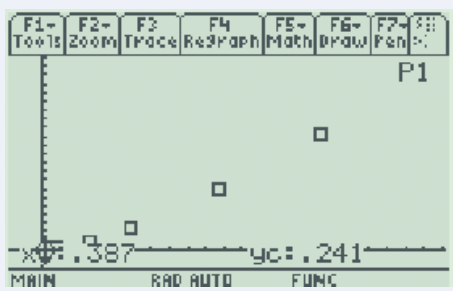
TI-Nspire instructions in Appendix B; HP Prime instructions on the book's Web site.

We'll use the planet data to illustrate a general strategy for performing transformations with logarithms on the TI-83/84 and TI-89. A similar approach could be used for transforming data with powers and roots.

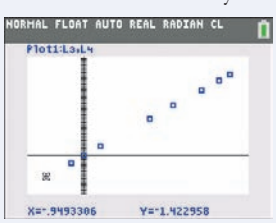
TI-83/84 TI-89

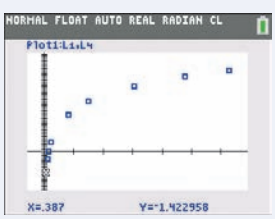
- Enter the values of the explanatory variable in L1/list1 and the values of the response variable in L2/list2. Make a scatterplot of y versus x and confirm that there is a curved pattern.

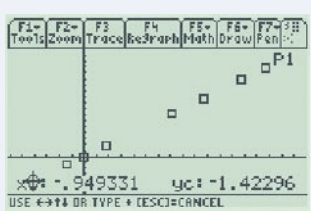


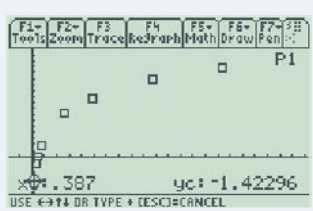


- Define L3/list3 to be the natural logarithm (ln) of L1/list1 and L4/list4 to be the natural logarithm of L2/list2. To see whether a power model fits the original data, make a plot of ln y (L4/list4) versus ln x (L3/list3) and look for linearity. To see whether an exponential model fits the original data, make a plot of ln y (L4/list4) versus x (L1/list1) and look for linearity.

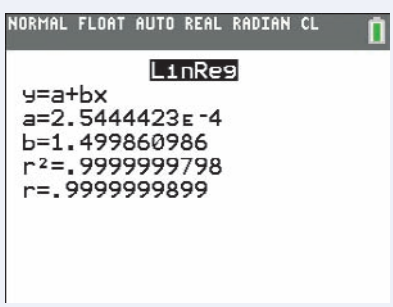


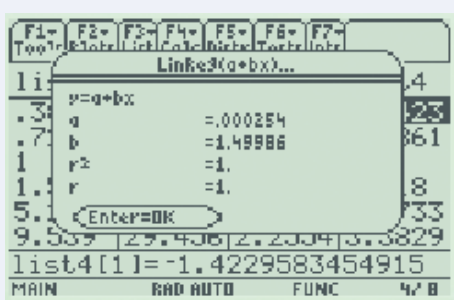




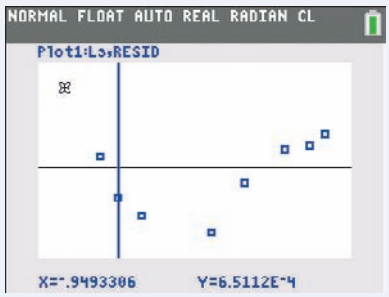


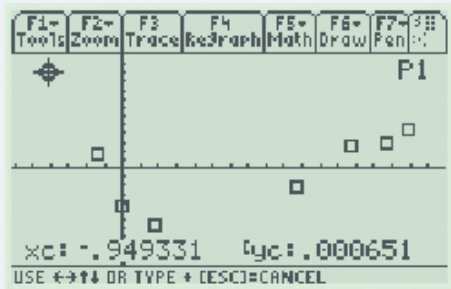
- If a linear pattern is present, calculate the equation of the least-squares regression line and store it in Y1. For the planet data, we executed the command `LinReg(a+bx)L3,L4,Y1`.



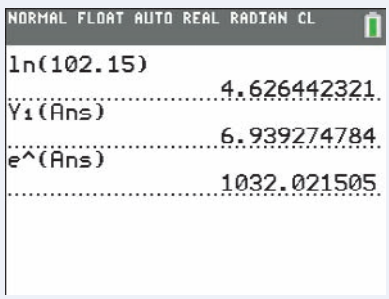


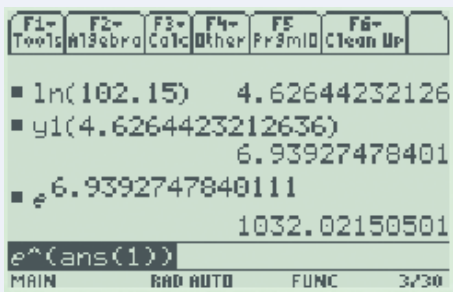
Construct a residual plot to look for any departures from the linear pattern. For Xlist, enter the list you used as the explanatory variable in the linear regression calculation. For Ylist, use the RESID list stored in the calculator. For the planet data, we used L3/list3 as the Xlist.





- To make a prediction for a specific value of the explanatory variable, compute log x or ln x, if appropriate. Then use Y1(k) to obtain the predicted value of log y or ln y. To get the predicted value of y, use 10^Ans or e^Ans to undo the logarithm transformation. Here's our prediction of the period of revolution for Eris, which is at a distance of 102.15 AU from the sun:





CHECK YOUR UNDERSTANDING

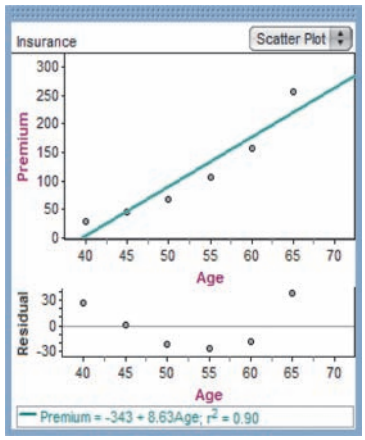
One sad fact about life is that we'll all die someday. Many adults plan ahead for their eventual passing by purchasing life insurance. Many different types of life insurance policies are available. Some provide coverage throughout an individual's life (whole life), while others last only for a specified number of years (term life). The policyholder makes regular payments (premiums) to the insurance company in return for the coverage. When the insured person dies, a payment is made to designated family members or other beneficiaries.

How do insurance companies decide how much to charge for life insurance? They rely on a staff of highly trained actuaries—people with expertise in probability, statistics, and advanced mathematics—to establish premiums. For an individual who wants to buy life insurance, the premium will depend on the type and amount of the policy as well as on personal characteristics like age, sex, and health status.

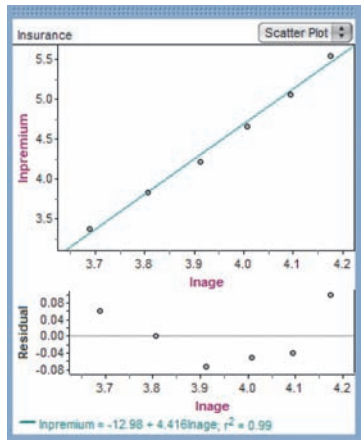
The table shows monthly premiums for a 10-year term-life insurance policy worth $1,000,000.[18]

Age (years)	Monthly premium
40	$29
45	$46
50	$68
55	$106
60	$157
65	$257

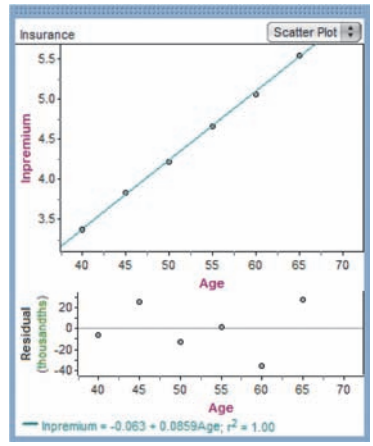
The Fathom screen shots below show three possible models for predicting monthly premium from age. Option 1 is based on the original data, while Options 2 and 3 involve transformations of the original data. Each screen shot includes a scatterplot with a least-squares regression line added and a residual plot.



OPTION 1



OPTION 2



OPTION 3

1. Use each model to predict how much a 58-year-old would pay for such a policy. Show your work.

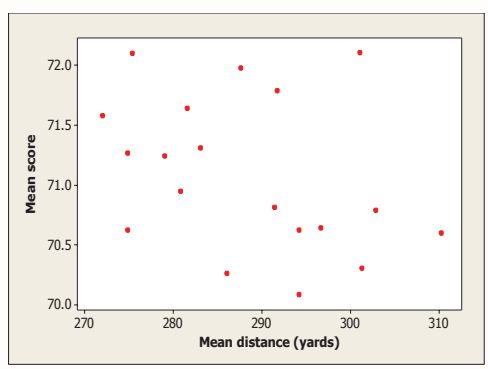
2. What type of function—linear, power, or exponential—best describes the relationship between age and monthly premium? Explain.

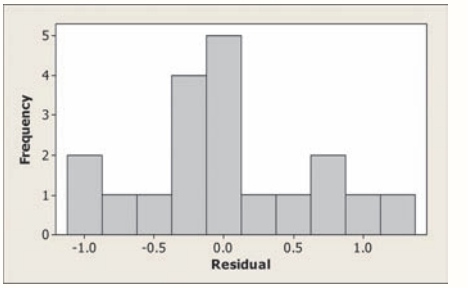
Do Longer Drives Mean Lower Scores on the PGA Tour?

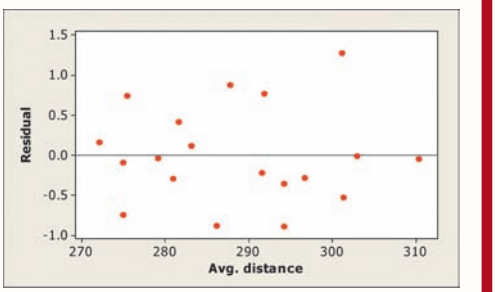
In the chapter-opening Case Study (page 737), we examined data on the mean drive distance (in yards) and mean score per round for an SRS of 19 of the 197 players on the PGA Tour in a recent year. Here is some Minitab output from a least-squares regression analysis on these data:

Predictor	Coef	SE Coef	T	P
Constant	76.904	3.808	20.20	0.000
Avg. distance	−0.02016	0.01319	−1.53	0.145

S = 0.618396 R-Sq = 12.1% R-Sq(adj) = 6.9%







1. Calculate the residual for the player with a mean drive distance of 275.4 yards and a mean score per round of 72.1. Show your work.
2. Interpret the value of s in this setting and explain what parameter s is estimating.
3. Do these data give convincing evidence at the $\alpha = 0.05$ level that the slope of the population regression line is negative?
4. Which kind of mistake—a Type I error or a Type II error—could you have made in Question 3? Justify your answer.

Section 12.2 Summary

- Curved relationships between two quantitative variables can sometimes be changed into linear relationships by **transforming** one or both of the variables. Once we transform the data to achieve linearity, we can fit a least-squares regression line to the transformed data and use this linear model to make predictions.
- When theory or experience suggests that the relationship between two variables follows a **power model** of the form $y = ax^p$, there are two transformations involving powers and roots that can linearize a curved pattern in a scatterplot. Option 1: Raise the values of the explanatory variable x to the power p, then look at a graph of (x^p, y). Option 2: Take the pth root of the values of the response variable y, then look at a graph of $(x, \sqrt[p]{y})$
- Another useful strategy for straightening a curved pattern in a scatterplot is to take the **logarithm** of one or both variables. When a power model describes the relationship between two variables, a plot of log y (ln y) versus log x (ln x) should be linear.
- In a linear model of the form $y = a + bx$, the values of the response variable are predicted to increase by a constant amount b for each increase of 1 unit in the explanatory variable. For an **exponential model** of the form $y = ab^x$, the predicted values of the response variable are multiplied by a factor of b for each increase of 1 unit in the explanatory variable. When an exponential model describes the relationship between two variables, a plot of log y (ln y) versus x should be linear.

12.2 TECHNOLOGY CORNER

TI-Nspire Instructions in Appendix B; HP Prime instructions on the book's Web site.

30. Transforming to achieve linearity on the calculator page 781

Section 12.2 Exercises

31. **The swinging pendulum** Mrs. Hanrahan's precalculus class collected data on the length (in centimeters) of a pendulum and the time (in seconds) the pendulum took to complete one back-and-forth swing (called its period). Here are their data:

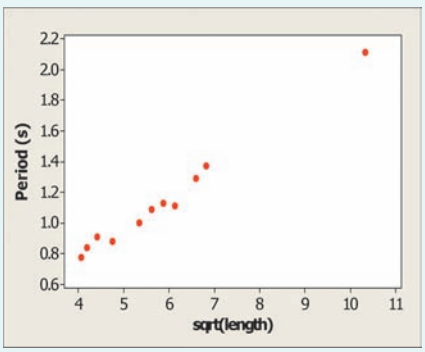
Length (cm)	Period (s)
16.5	0.777
17.5	0.839
19.5	0.912
22.5	0.878
28.5	1.004
31.5	1.087
34.5	1.129
37.5	1.111
43.5	1.290
46.5	1.371
106.5	2.115

(a) Make a reasonably accurate scatterplot of the data by hand, using length as the explanatory variable. Describe what you see.

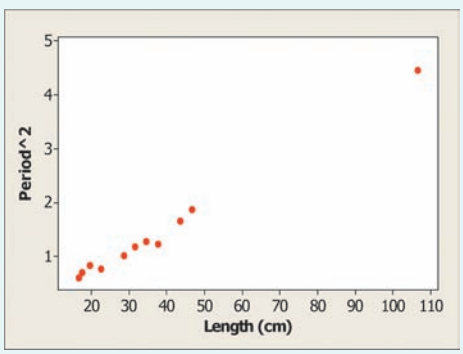
(b) The theoretical relationship between a pendulum's length and its period is

$$\text{period} = \frac{2\pi}{\sqrt{g}}\sqrt{\text{length}}$$

where g is a constant representing the acceleration due to gravity (in this case, $g = 980$ cm/s^2). Use the following graph to identify the transformation that was used to linearize the curved pattern in part (a).



(c) Use the following graph to identify the transformation that was used to linearize the curved pattern in part (a).

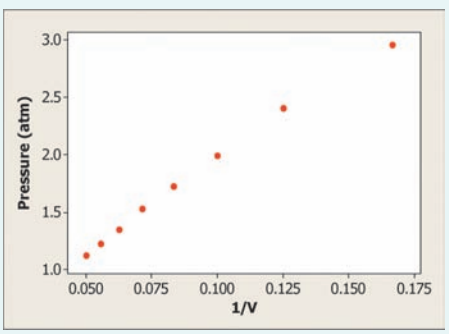


32. **Boyle's law** If you have taken a chemistry or physics class, then you are probably familiar with Boyle's law: for gas in a confined space kept at a constant temperature, pressure times volume is a constant (in symbols, $PV = k$). Students collected the following data on pressure and volume using a syringe and a pressure probe.

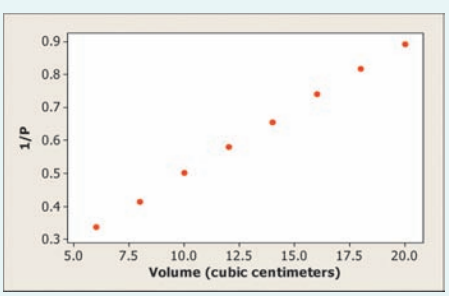
Volume (cubic centimeters)	Pressure (atmospheres)
6	2.9589
8	2.4073
10	1.9905
12	1.7249
14	1.5288
16	1.3490
18	1.2223
20	1.1201

(a) Make a reasonably accurate scatterplot of the data by hand using volume as the explanatory variable. Describe what you see.

(b) If the true relationship between the pressure and volume of the gas is $PV = k$, we can divide both sides of this equation by V to obtain the theoretical model $P = k/V$, or $P = k(1/V)$. Use the graph below to identify the transformation that was used to linearize the curved pattern in part (a).



(c) Use the graph below to identify the transformation that was used to linearize the curved pattern in part (a).

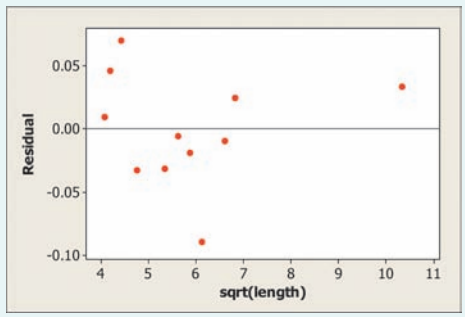


33. The swinging pendulum Refer to Exercise 31. Here is Minitab output from separate regression analyses of the two sets of transformed pendulum data:

pg 770

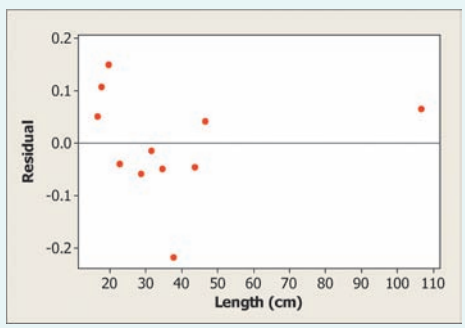
Transformation 1: ($\sqrt{\text{length}}$, period)

```
Predictor  Coef       SE Coef   T      P
Constant   −0.08594   0.05046   −1.70  0.123
sqrt        0.209999  0.008322  25.23  0.000
 (length)
S = 0.0464223  R-Sq = 98.6%  R-Sq(adj) = 98.5%
```



Transformation 2: (length, period^2)

```
Predictor   Coef       SE Coef   T      P
Constant    −0.15465   0.05802   −2.67  0.026
Length (cm)  0.042836  0.001320  32.46  0.000
S = 0.105469   R-Sq = 99.2%   R-Sq(adj) = 99.1%
```



Do each of the following for *both* transformations.

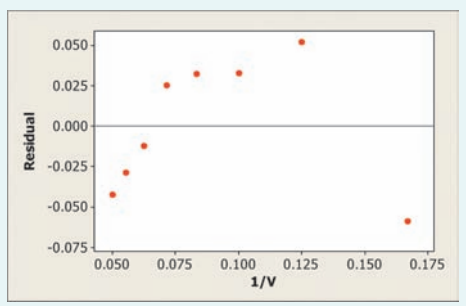
(a) Give the equation of the least-squares regression line. Define any variables you use.

(b) Use the model from part (a) to predict the period of a pendulum with length 80 centimeters. Show your work.

34. Boyle's law Refer to Exercise 32. Here is Minitab output from separate regression analyses of the two sets of transformed pressure data:

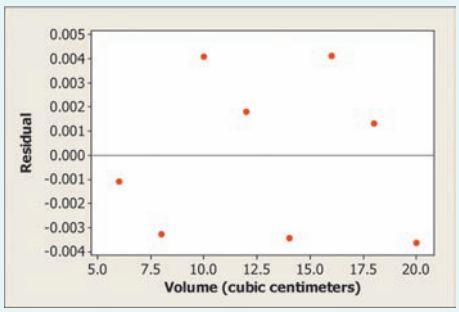
Transformation 1: $\left(\frac{1}{\text{volume}}, \text{pressure}\right)$

```
Predictor  Coef       SE Coef   T      P
Constant    0.36774   0.04055    9.07  0.000
1/V        15.8994    0.4190    37.95  0.000
S = 0.044205   R-Sq = 99.6%   R-Sq(adj) = 99.5%
```



Transformation 2: $\left(\text{volume}, \frac{1}{\text{pressure}}\right)$

```
Predictor Coef        SE Coef    T       P
Constant  0.100170    0.003779    26.51  0.000
Volume    0.0398119   0.0002741  145.23  0.000
S = 0.003553   R-Sq = 100.0%   R-Sq(adj) = 100.0%
```



Do each of the following for *both* transformations.

(a) Give the equation of the least-squares regression line. Define any variables you use.

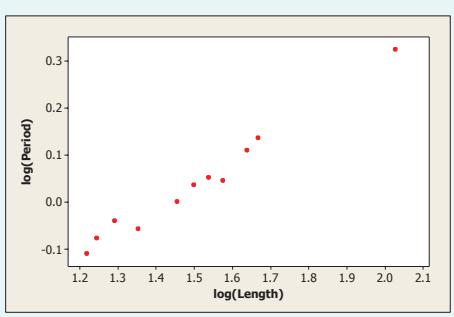
(b) Use the model from part (a) to predict the pressure in the syringe when the volume is 17 cubic centimeters. Show your work.

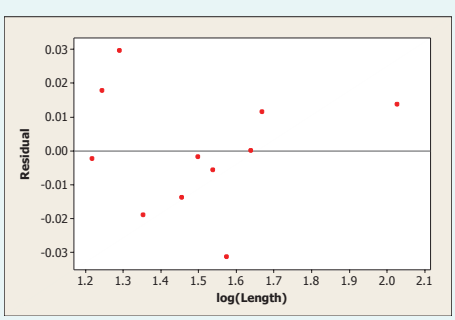
35. The swinging pendulum Refer to Exercise 31. We took the logarithm (base 10) of the values for both variables. Some computer output from a linear regression analysis on the transformed data is shown below.

pg 772

Regression Analysis: log(Period) versus log(Length)

```
Predictor     Coef       SE Coef  T       P
Constant      −0.73675   0.03808  −19.35  0.000
log(Length)   0.51701    0.02511   20.59  0.000
S = 0.0185568   R-Sq = 97.9%   R-Sq(adj) = 97.7%
```





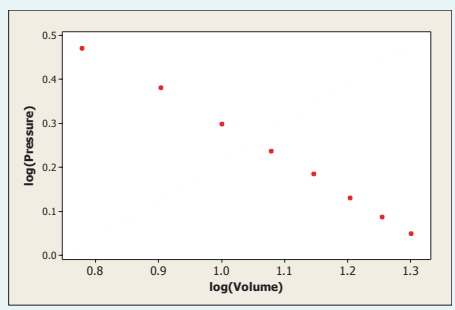
(a) Based on the output, explain why it would be reasonable to use a power model to describe the relationship between the length and period of a pendulum.

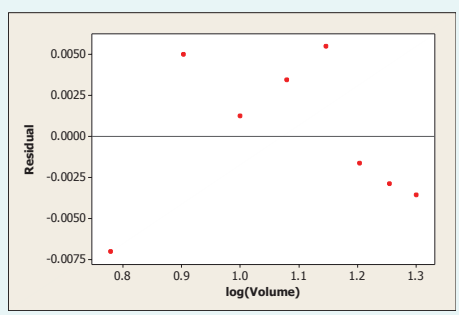
(b) Give the equation of the least-squares regression line. Be sure to define any variables you use.

36. Boyle's law Refer to Exercise 32. We took the logarithm (base 10) of the values for both variables. Some computer output from a linear regression analysis on the transformed data is shown below.

Regression Analysis: log(Pressure) versus log(Volume)

```
Predictor     Coef       SE Coef  T       P
Constant      1.11116    0.01118   99.39  0.000
log(Volume)  −0.81344    0.01020  −79.78  0.000
S = 0.00486926   R-Sq = 99.9%   R-Sq(adj) = 99.9%
```





(a) Based on the output, explain why it would be reasonable to use a power model to describe the relationship between pressure and volume.

(b) Give the equation of the least-squares regression line. Be sure to define any variables you use.

37. The swinging pendulum Use your model from Exercise 35 to predict the period of a pendulum with length 80 centimeters. Show your work.

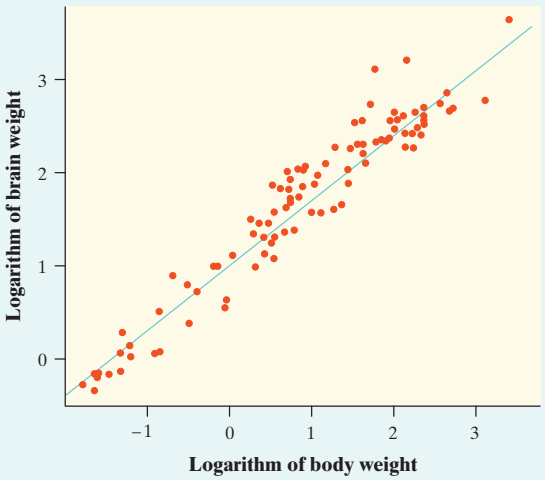
pg 773

38. Boyle's law Use your model from Exercise 36 to predict the pressure in the syringe when the volume is 17 cubic centimeters. Show your work.

39. Brawn versus brain How is the weight of an animal's brain related to the weight of its body? Researchers collected data on the brain weight (in grams) and body weight (in kilograms) for 96 species of mammals.[19] The following figure is a scatterplot of

the logarithm of brain weight against the logarithm of body weight for all 96 species. The least-squares regression line for the transformed data is

$$\widehat{\log y} = 1.01 + 0.72 \log x$$



Based on footprints and some other sketchy evidence, some people believe that a large apelike animal, called Sasquatch or Bigfoot, lives in the Pacific Northwest. His weight is estimated to be about 280 pounds, or 127 kilograms. How big is Bigfoot's brain? Show your method clearly.

40. **Determining tree biomass** It is easy to measure the "diameter at breast height" of a tree. It's hard to measure the total "aboveground biomass" of a tree, because to do this you must cut and weigh the tree. The biomass is important for studies of ecology, so ecologists commonly estimate it using a power model. Combining data on 378 trees in tropical rain forests gives this relationship between biomass y measured in kilograms and diameter x measured in centimeters:[20]

$$\widehat{\ln y} = -2.00 + 2.42 \ln x$$

Use this model to estimate the biomass of a tropical tree 30 centimeters in diameter. Show your work.

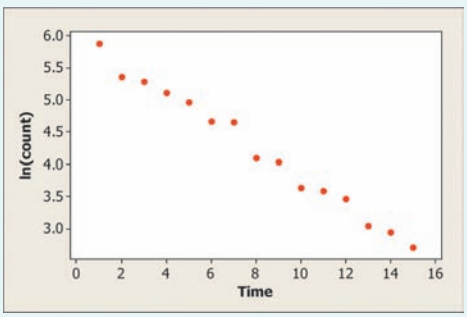
41. **Killing bacteria** Expose marine bacteria to X-rays for time periods from 1 to 15 minutes. Here are the number of surviving bacteria (in hundreds) on a culture plate after each exposure time:[21]

pg 776

Time t	Count y	Time t	Count y
1	355	9	56
2	211	10	38
3	197	11	36
4	166	12	32
5	142	13	21
6	106	14	19
7	104	15	15
8	60		

(a) Make a reasonably accurate scatterplot of the data by hand, using time as the explanatory variable. Describe what you see.

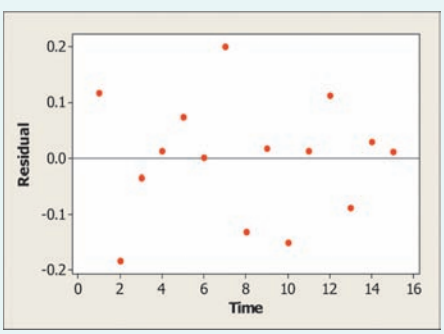
(b) A scatterplot of the natural logarithm of the number of surviving bacteria versus time is shown below. Based on this graph, explain why it would be reasonable to use an exponential model to describe the relationship between count of bacteria and time.



(c) Minitab output from a linear regression analysis on the transformed data is shown below.

```
Predictor  Coef       SE Coef   T       P
Constant    5.97316   0.05978    99.92  0.000
Time       −0.218425  0.006575 −33.22  0.000

S = 0.110016   R-Sq = 98.8%   R-Sq(adj) = 98.7%
```



Give the equation of the least-squares regression line. Be sure to define any variables you use.

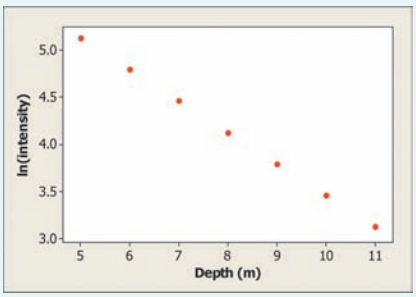
(d) Use your model to predict the number of surviving bacteria after 17 minutes. Show your work.

42. **Light through the water** Some college students collected data on the intensity of light at various depths in a lake. Here are their data:

Depth (m)	Light intensity (lumens)
5	168.00
6	120.42
7	86.31
8	61.87
9	44.34
10	31.78
11	22.78

(a) Make a reasonably accurate scatterplot of the data by hand, using depth as the explanatory variable. Describe what you see.

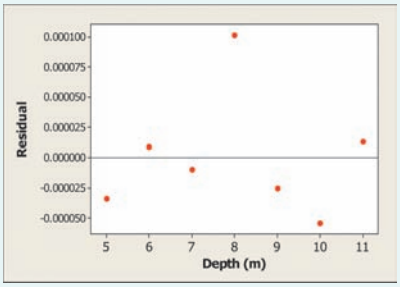
(b) A scatterplot of the natural logarithm of light intensity versus depth is shown below. Based on this graph, explain why it would be reasonable to use an exponential model to describe the relationship between light intensity and depth.



(c) Minitab output from a linear regression analysis on the transformed data is shown below.

```
Predictor  Coef       SE Coef  T          P
Constant    6.78910   0.00009   78575.46  0.000
Depth (m)  −0.333021  0.000010 −31783.44  0.000

S = 0.000055   R-Sq = 100.0%   R-Sq(adj) = 100.0%
```



Give the equation of the least-squares regression line. Be sure to define any variables you use.

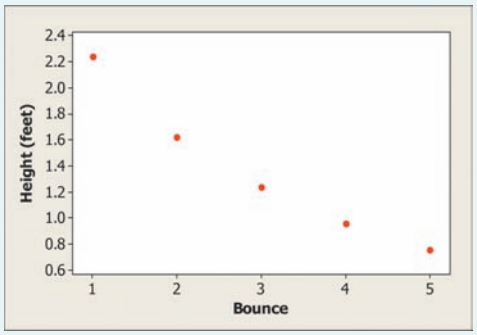
(d) Use your model to predict the light intensity at a depth of 12 meters. Show your work.

43. Follow the bouncing ball Students in Mr. Handford's class dropped a kickball beneath a motion detector. The detector recorded the height of the ball as it bounced up and down several times. Here are the heights of the ball at the highest point on the first five bounces:

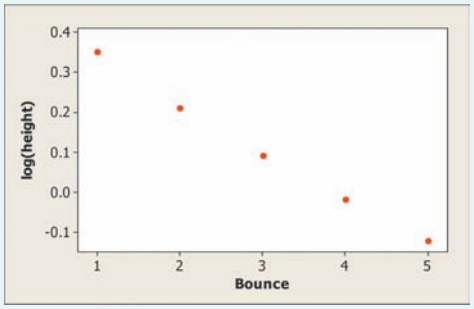
pg 779

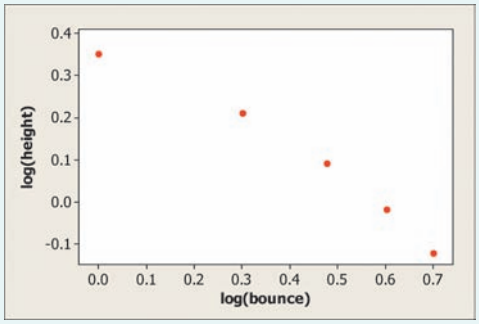
Bounce number	Height (ft)
1	2.240
2	1.620
3	1.235
4	0.958
5	0.756

Here is a scatterplot of the data:



(a) The following graphs show the results of two different transformations of the data. Would an exponential model or a power model provide a better description of the relationship between bounce number and height? Justify your answer.





(b) Minitab output from a linear regression analysis on the transformed data of log(height) versus bounce number is shown below. Give the equation of the least-squares regression line. Be sure to define any variables you use.

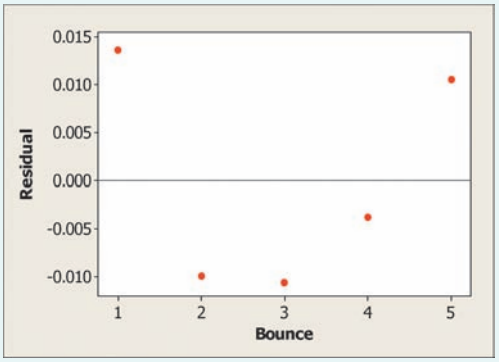
```
Predictor  Coef       SE Coef  T        P
Constant    0.45374   0.01385   32.76  0.000
Bounce     −0.117160  0.004176 −28.06  0.000

S = 0.0132043   R-Sq = 99.6%   R-Sq(adj) = 99.5%
```

(c) Use your model from part (b) to predict the highest point the ball reaches on its seventh bounce. Show your work.

(d) A residual plot for the linear regression in part (b) is shown on the next page. Do you expect your prediction

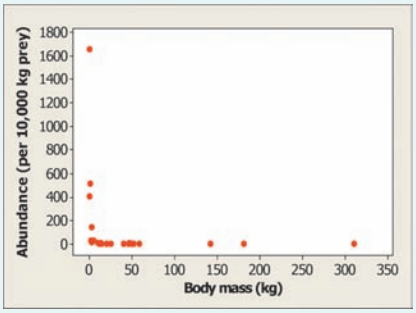
in part (c) to be too high, too low, or about right? Justify your answer.



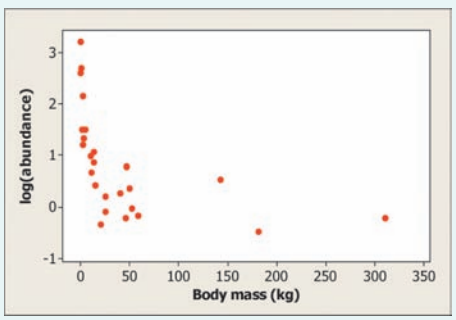
44. Counting carnivores Ecologists look at data to learn about nature's patterns. One pattern they have found relates the size of a carnivore (body mass in kilograms) to how many of those carnivores there are in an area. A good measure of "how many" is to count carnivores per 10,000 kilograms (kg) of their prey in the area. The table below gives data for 25 carnivore species.[22]

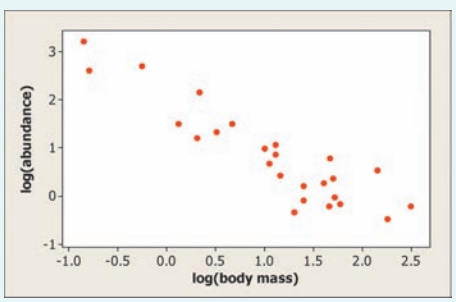
Carnivore species	Body mass (kg)	Abundance (per 10,000 kg of prey)
Least weasel	0.14	1656.49
Ermine	0.16	406.66
Small Indian mongoose	0.55	514.84
Pine marten	1.3	31.84
Kit fox	2.02	15.96
Channel Islands fox	2.16	145.94
Arctic fox	3.19	21.63
Red fox	4.6	32.21
Bobcat	10.0	9.75
Canadian lynx	11.2	4.79
European badger	13.0	7.35
Coyote	13.0	11.65
Ethiopian wolf	14.5	2.70
Eurasian lynx	20.0	0.46
Wild dog	25.0	1.61
Dhole	25.0	0.81
Snow leopard	40.0	1.89
Wolf	46.0	0.62
Leopard	46.5	6.17
Cheetah	50.0	2.29
Puma	51.9	0.94
Spotted hyena	58.6	0.68
Lion	142.0	3.40
Tiger	181.0	0.33
Polar bear	310.0	0.60

Here is a scatterplot of the data.



(a) The following graphs show the results of two different transformations of the data. Would an exponential model or a power model provide a better description of the relationship between body mass and abundance? Justify your answer.





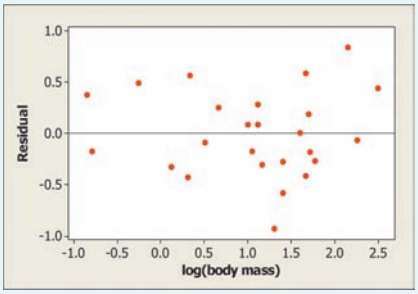
(b) Minitab output from a linear regression analysis on the transformed data of log(abundance) versus log(body mass) is shown below. Give the equation of the least-squares regression line. Be sure to define any variables you use.

```
Predictor      Coef      SE Coef     T        P
Constant     1.9503      0.1342     14.53    0.000
log(body    −1.04811     0.09802   −10.69    0.000
 mass)

S = 0.423352   R-Sq = 83.3%   R-Sq(adj) = 82.5%
```

(c) Use your model from part (b) to predict the abundance of black bears, which have a body mass of 92.5 kilograms. Show your work.

(d) A residual plot for the linear regression in part (b) is shown at top right. Explain what this graph tells you about the appropriateness of the model.



45. **Heart weights of mammals** Here are some data on the hearts of various mammals.[23]

Mammal	Length of cavity of left ventricle (cm)	Heart weight (g)
Mouse	0.55	0.13
Rat	1.0	0.64
Rabbit	2.2	5.8
Dog	4.0	102
Sheep	6.5	210
Ox	12.0	2030
Horse	16.0	3900

(a) Make an appropriate scatterplot for predicting heart weight from length. Describe what you see.

(b) Use transformations to linearize the relationship. Does the relationship between heart weight and length seem to follow an exponential model or a power model? Justify your answer.

(c) Perform least-squares regression on the transformed data. Give the equation of your regression line. Define any variables you use.

(d) Use your model from part (c) to predict the heart weight of a human who has a left ventricle 6.8 centimeters long. Show your work.

46. **Galileo's experiment** Galileo studied motion by rolling balls down ramps. He rolled a ball down a ramp with a horizontal shelf at the end of it so that the ball was moving horizontally when it started to fall off the shelf. The top of the ramp was placed at different heights above the floor (that is, the length of the ramp varied), and Galileo measured the horizontal distance the ball traveled before it hit the floor. Here are Galileo's data. (We won't try to describe the obsolete seventeenth-century units Galileo used to measure distance and height.)[24]

Height	Distance
1000	1500
828	1340
800	1328
600	1172
300	800

(a) Make an appropriate scatterplot for predicting horizontal distance traveled from ramp height. Describe what you see.

(b) Use transformations to linearize the relationship. Does the relationship between distance and height seem to follow an exponential model or a power model? Justify your answer.

(c) Perform least-squares regression on the transformed data. Give the equation of your regression line. Define any variables you use.

(d) Use your model from part (c) to predict the horizontal distance a ball would travel if the ramp height was 700. Show your work.

Multiple Choice: Select the best answer for Exercises 47 to 50.

47. Suppose that the relationship between a response variable y and an explanatory variable x is modeled by $y = 2.7(0.316)^x$. Which of the following scatterplots would approximately follow a straight line?

(a) A plot of y against x

(b) A plot of y against $\log x$

(c) A plot of $\log y$ against x

(d) A plot of $\log y$ against $\log x$

(e) A plot of $\sqrt{y}$ against x.

48. Some high school physics students dropped a ball and measured the distance fallen (in centimeters) at various times (in seconds) after its release. If you have studied physics, then you probably know that the theoretical relationship between the variables is distance $= 490(\text{time})^2$. A scatterplot of the students' data showed a clear curved pattern. At 0.68 seconds after release, the ball had fallen 220.4 centimeters. How much more or less did the ball fall than the theoretical model predicts?

(a) More by 226.576 centimeters

(b) More by 6.176 centimeters

(c) No more and no less

(d) Less by 226.576 centimeters

(e) Less by 6.176 centimeters

49. A scatterplot of $x =$ Super Bowl number and $y =$ cost of a 30-second advertisement on the Super Bowl broadcast (in dollars) shows a strong, positive, nonlinear association. A scatterplot of ln(cost) versus Super Bowl number is roughly linear. The least-squares regression line for this association is $\widehat{\ln(\text{cost})} = 10.97 + 0.0971\,(\text{Super Bowl number})$. Predict the cost of a 30-second advertisement for Super Bowl 40.

(a) $3
(b) $15
(c) $58,153
(d) $83,132
(e) $2,824,947

50. A scatterplot of y versus x shows a positive, nonlinear association. Two different transformations are attempted to try to linearize the association: using the logarithm of the y values and using the square root of the y values. Two least-squares regression lines are calculated, one that uses x to predict $\log(y)$ and the other that uses x to predict $\sqrt{y}$. Which of the following would be the best reason to prefer the least-squares regression line that uses x to predict $\log(y)$?

(a) The value of r^2 is smaller.
(b) The standard deviation of the residuals is smaller.
(c) The slope is greater.
(d) The residual plot has more random scatter.
(e) The distribution of residuals is more Normal.

51. Shower time (1.3, 2.2, 6.3, 7.3) Marcella takes a shower every morning when she gets up. Her time in the shower varies according to a Normal distribution with mean 4.5 minutes and standard deviation 0.9 minutes.

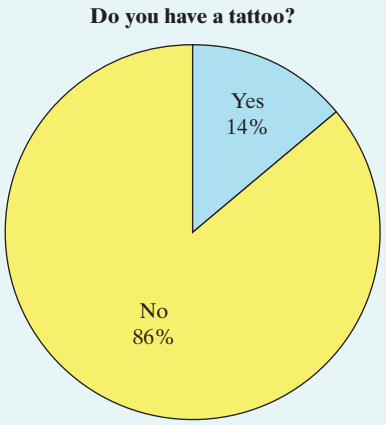
(a) Find the probability that Marcella's shower lasts between 3 and 6 minutes on a randomly selected day. Show your work.

(b) If Marcella took a 7-minute shower, would it be classified as an outlier by the 1.5*IQR* rule? Justify your answer.

(c) Suppose we choose 10 days at random and record the length of Marcella's shower each day. What's the probability that her shower time is 7 minutes or higher on at least 2 of the days? Show your work.

(d) Find the probability that the *mean* length of her shower times on these 10 days exceeds 5 minutes. Show your work.

52. Tattoos (8.2) What percent of U.S. adults have one or more tattoos? The Harris Poll conducted an online survey of 2302 adults during January 2008. According to the published report, "Respondents for this survey were selected from among those who have agreed to participate in Harris Interactive surveys."[25] The pie chart at top right summarizes the responses from those who were surveyed. Explain why it would *not* be appropriate to use these data to construct a 95% confidence interval for the proportion of all U.S. adults who have tattoos.



Exercises 53 and 54 refer to the following setting. About 1100 high school teachers attended a weeklong summer institute for teaching AP® classes. After hearing about the survey in Exercise 52, the teachers in the AP® Statistics class wondered whether the results of the tattoo survey would be similar for teachers. They designed a survey to find out. The class opted to take a random sample of 100 teachers at the institute. One of the questions on the survey was

Do you have any tattoos on your body?

(Circle one) YES NO

53. Tattoos (8.2, 9.2) Of the 98 teachers who responded, 23.5% said that they had one or more tattoos.

(a) Construct and interpret a 95% confidence interval for the actual proportion of teachers at the AP® institute who would say they had tattoos.

(b) Does the interval in part (a) provide convincing evidence that the proportion of teachers at the institute with tattoos is not 0.14 (the value cited in the Harris Poll report)? Justify your answer.

(c) Two of the selected teachers refused to respond to the survey. If both of these teachers had responded, could your answer to part (b) have changed? Justify your answer.

54. Tattoos (4.1) One of the first decisions the class had to make was what kind of sampling method to use.

(a) They knew that a simple random sample was the "preferred" method. With 1100 teachers in 40 different sessions, the class decided not to use an SRS. Give at least two reasons why you think they made this decision.

(b) The AP® Statistics class believed that there might be systematic differences in the proportions of teachers who had tattoos based on the subject areas that they taught. What sampling method would you recommend to account for this possibility? Explain a statistical advantage of this method over an SRS.

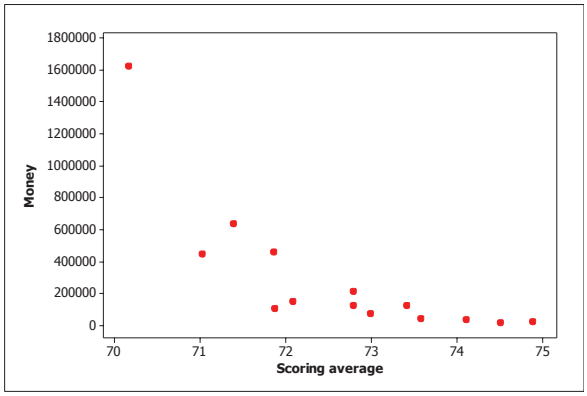
FRAPPY! Free Response AP® Problem, Yay!

The following problem is modeled after actual AP® Statistics exam free response questions. Your task is to generate a complete, concise response in 15 minutes.

Directions: Show all your work. Indicate clearly the methods you use, because you will be scored on the correctness of your methods as well as on the accuracy and completeness of your results and explanations.

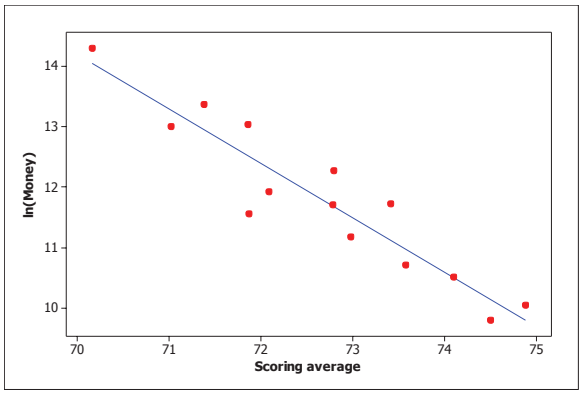
A random sample of 14 golfers was selected from the 147 players on the Ladies Professional Golf Association (LPGA) tour in a recent year. The total amount of money won during the year (in dollars) and the scoring average for each player in the sample was recorded. Lower scoring averages are better in golf.

The scatterplot below displays the relationship between money and scoring average for these 14 players.



(a) Explain why it would *not* be appropriate to construct a confidence interval for the slope of the least-squares regression line relating money to scoring average.

A scatterplot of the natural logarithm of money versus scoring average is shown at top right along with some computer output for a least-squares regression using the transformed data.



Predictor	Coef	SE Coef	T	P
Constant	77.537	7.035	11.02	0.000
Scoring average	−0.90470	0.09679	−9.35	0.000

S = 0.475059 R-Sq = 87.9% R-Sq(adj) = 86.9%

(b) Predict the amount of money won for an LPGA golfer with a scoring average of 70.

(c) Calculate and interpret a 95% confidence interval for the slope of the least-squares regression line relating ln(money) to scoring average. Assume that the conditions for inference have been met.

After you finish, you can view two example solutions on the book's Web site (www.whfreeman.com/tps5e). Determine whether you think each solution is "complete," "substantial," "developing," or "minimal." If the solution is not complete, what improvements would you suggest to the student who wrote it? Finally, your teacher will provide you with a scoring rubric. Score your response and note what, if anything, you would do differently to improve your own score.

Chapter Review

Section 12.1: Inference for Linear Regression

In this section, you learned how to conduct inference about the slope of a population (true) least-squares regression line. The sampling distribution of the sample slope b is the foundation for doing inference about the population (true) slope β. When the conditions are met, the sampling distribution of b has an approximately Normal distribution with mean $\mu_b = \beta$ and standard deviation $\sigma_b = \dfrac{\sigma}{\sigma_x\sqrt{n}}$.

There are five conditions for performing inference about a population (true) slope. Remember them with the acronym LINER.

- The Linear condition says that the mean value of the response variable μ_y falls on the population (true) regression line $\mu_y = \alpha + \beta x$. To check the Linear condition, verify that there are no leftover patterns in the residual plot.
- The Independent condition says that individual observations are independent of each other. To check the Independent condition, verify that the sample size is less than 10% of the population size. Also, convince yourself that knowing the response for one individual won't help you predict the response for another individual.
- The Normal condition says that the distribution of y values is approximately Normal for each value of x. To check the Normal condition, graph a dotplot, histogram, or Normal probability plot of the residuals and verify that there are no outliers or strong skewness.
- The Equal SD condition says that for each value of x, the distribution of y should have the same standard deviation. To check the Equal SD condition, verify that the residuals have roughly the same amount of scatter around the residual = 0 line for each value of x on the residual plot.
- The Random condition says that the data are from a random sample or a randomized experiment. To check the Random condition, verify that randomness was properly used in the data collection process.

To construct and interpret a confidence interval for the slope of the population (true) least-squares regression line, follow the familiar four-step process. The formula for the confidence interval is $b \pm t^*\text{SE}_b$, where t^* is the t critical value with df = $n - 2$. The standard error of the slope SE_b describes how far the sample slope typically varies from the population (true) slope in repeated random samples or random assignments. The formula for the standard error of the slope is $\text{SE}_b = \dfrac{s}{s_x\sqrt{n-1}}$. The standard error of the slope is typically provided with standard computer output for least-squares regression.

In most cases, when you conduct a significance test for the slope of the population (true) least-squares regression line, the null hypothesis is H_0:$\beta = 0$. This hypothesis says that a straight-line relationship between x and y is of no value for predicting y. To do the calculations, use the test statistic $t = \dfrac{b - \beta_0}{\text{SE}_b}$ with df = $n - 2$. The value of the test statistic, along with a two-sided P-value, is typically provided with standard computer output for least-squares regression.

Section 12.2: Transforming to Achieve Linearity

When the association between two variables is nonlinear, transforming one or both of the variables can result in a linear association.

If the association between two variables follows a power model in the form $y = ax^p$, there are several transformations that will result in a linear association.

- Raise the values of x to the power of p and plot y versus x^p.
- Calculate the pth root of the y values and plot $\sqrt[p]{y}$ versus x.
- Calculate the logarithms of the x values and the y values and plot $\log(y)$ versus $\log(x)$. You can use base 10 logarithms (log) or base e logarithms (ln).

If the association between two variables follows an exponential model in the form $y = ab^x$, transform the data by computing the logarithms of the y values and plot $\log(y)$ versus x (or $\ln(y)$ versus x).

Once you have achieved linearity, calculate the equation of the least-squares regression line using the transformed data. Remember to include the transformed variables when you are writing the equation of the line. Likewise, when using the line to make predictions, make sure that the prediction is in the original units of y. If you transformed the y variable, you will need to undo the transformation after using the least-squares regression line.

To decide between two or more transformations, look at the residual plots and choose the one with the most random scatter.

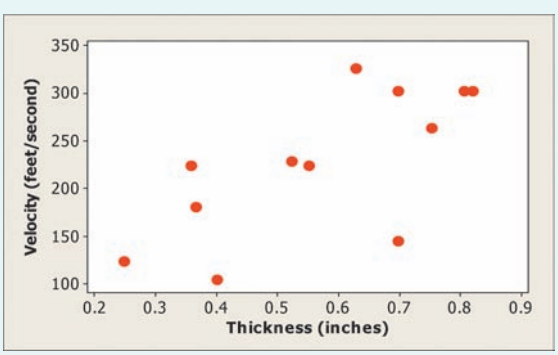
What Did You Learn?

Learning Objective	Section	Related Example on Page(s)	Relevant Chapter Review Exercise(s)
Check the conditions for performing inference about the slope β of the population (true) regression line.	12.1	745	R12.2, R12.3, R12.4
Interpret the values of a, b, s, SE_b, and r^2 in context, and determine these values from computer output.	12.1	748, 754	R12.1
Construct and interpret a confidence interval for the slope β of the population (true) regression line.	12.1	749	R12.3
Perform a significance test about the slope β of the population (true) regression line.	12.1	754	R12.2
Use transformations involving powers and roots to find a power model that describes the relationship between two variables, and use the model to make predictions.	12.2	768, 770	R12.5
Use transformations involving logarithms to find a power model or an exponential model that describes the relationship between two variables, and use the model to make predictions.	12.2	772, 773, 776	R12.6
Determine which of several transformations does a better job of producing a linear relationship.	12.2	779	R12.6

Chapter 12 Chapter Review Exercises

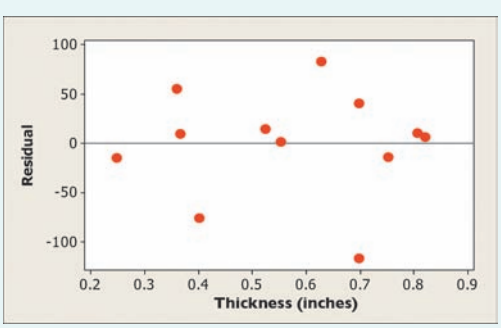
These exercises are designed to help you review the important ideas and methods of the chapter.

Exercises R12.1 to R12.3 refer to the following setting. In the casting of metal parts, molten metal flows through a "gate" into a die that shapes the part. The gate velocity (the speed at which metal is forced through the gate) plays a critical role in die casting. A firm that casts cylindrical aluminum pistons examined a random sample of 12 pistons formed from the same alloy of metal. What is the relationship between the cylinder wall thickness (inches) and the gate velocity (feet per second) chosen by the skilled workers who do the casting? If there is a clear pattern, it can be used to direct new workers or to automate the process. A scatterplot of the data is shown below.[26]



A least-squares regression analysis was performed on the data. Some computer output and a residual plot are shown below. A Normal probability plot of the residuals (not shown) is roughly linear.

```
Predictor    Coef      SE Coef    T       P
Constant     70.44     52.90      1.33    0.213
Thickness    274.78    88.18      ***     ***
S = 56.3641  R-Sq = 49.3%  R-Sq(adj) = 44.2%
```



R12.1 Casting aluminum

(a) Describe what the scatterplot tells you about the relationship between cylinder wall thickness and gate velocity.

(b) What is the equation of the least-squares regression line? Define any variables you use.

(c) One of the cylinders in the sample had a wall thickness of 0.4 inches. The gate velocity chosen for this cylinder was 104.8 feet per second. Does the regression line in part (b) overpredict or underpredict the gate velocity for this cylinder? By how much? Show your work.

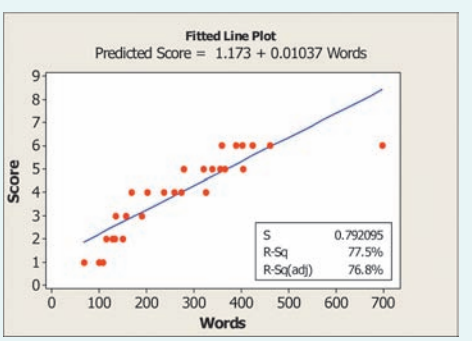
(d) Is a linear model appropriate in this setting? Justify your answer with appropriate evidence.

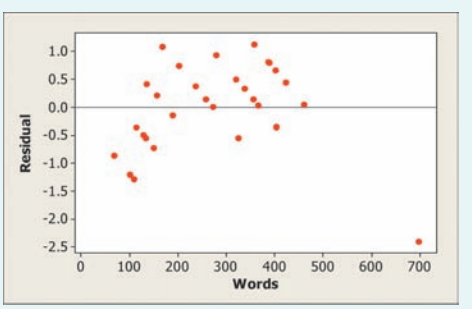
(e) Interpret each of the following in context:
(i) The slope
(ii) s
(iii) r^2
(iv) The standard error of the slope

R12.2 Casting aluminum Do the data provide convincing evidence at the $\alpha = 0.05$ level of a linear relationship between thickness and gate velocity in the population of pistons formed from this alloy of metal?

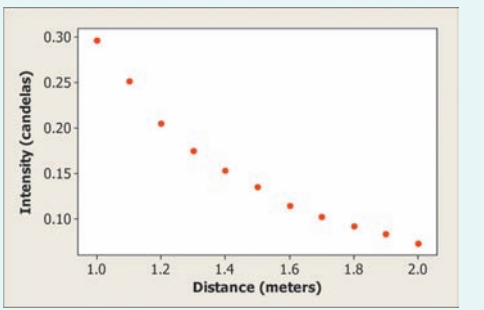
R12.3 Casting aluminum Construct and interpret a 95% confidence interval for the slope of the population regression line. Explain how this interval is consistent with the results of Exercise R12.2.

R12.4 SAT essay—is longer better? Following the debut of the new SAT Writing test in March 2005, Dr. Les Perelman from the Massachusetts Institute of Technology recorded the number of words and score for each essay in a sample provided by the College Board. A least-squares regression analysis was performed on these data. The two graphs at bottom left display the results of that analysis. Explain why the conditions for performing inference are not met in this setting.





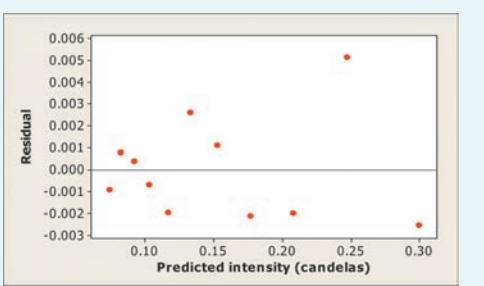
R12.5 Light intensity In a physics class, the intensity of a 100-watt lightbulb was measured by a sensor at various distances from the light source. A scatterplot of the data is shown below. Note that a candela is a unit of luminous intensity in the International System of Units.



Physics textbooks suggest that the relationship between light intensity y and distance x should follow an "inverse square law," that is, a power law model of the form $y = ax^{-2} = a\frac{1}{x^2}$. We transformed the distance measurements by squaring them and then taking their reciprocals. Some computer output and a residual plot from a least-squares regression analysis on the transformed data are shown below. Note that the horizontal axis on the residual plot displays predicted light intensity.

Predictor	Coef	SE Coef	T	P
Constant	−0.000595	0.001821	−0.33	0.751
Distance^(−2)	0.299624	0.003237	92.56	0.000

S = 0.00248369 R-Sq = 99.9% R-Sq(adj) = 99.9%

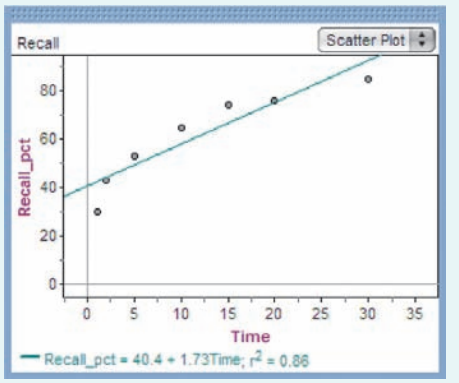


(a) Did this transformation achieve linearity? Give appropriate evidence to justify your answer.

(b) What is the equation of the least-squares regression line? Define any variables you use.

(c) What would you predict for the intensity of a 100-watt bulb at a distance of 2.1 meters? Show your work.

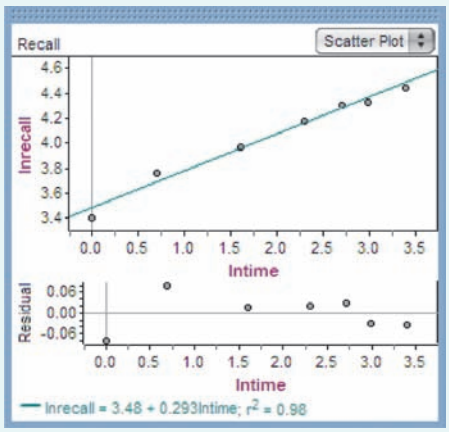
R12.6 An experiment was conducted to determine the effect of practice time (in seconds) on the percent of unfamiliar words recalled. Here is a Fathom scatterplot of the results with a least-squares regression line superimposed.

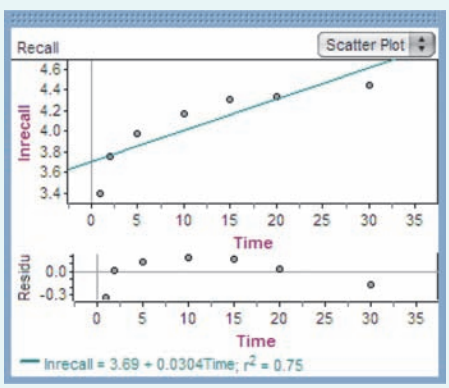


(a) Sketch a residual plot. Be sure to label your axes.

(b) Explain why a linear model is not appropriate for describing the relationship between practice time and percent of words recalled.

(c) We used Fathom to transform the data in hopes of achieving linearity. The screen shots on the right show the results of two different transformations. Would an exponential model or a power model describe the relationship better? Justify your answer.





(d) Use each model to predict the word recall for 25 seconds of practice. Show your work. Which prediction do you think will be better?

Chapter 12 AP® Statistics Practice Test

Section I: Multiple Choice *Select the best answer for each question.*

T12.1 Which of the following is *not* one of the conditions that must be satisfied in order to perform inference about the slope of a least-squares regression line?

(a) For each value of x, the population of y-values is Normally distributed.

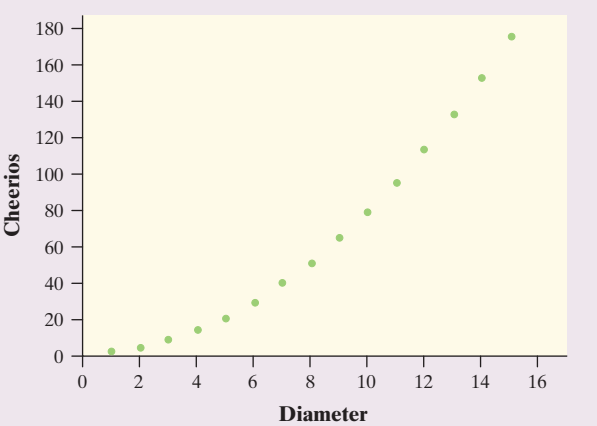
(b) The standard deviation σ of the population of y-values corresponding to a particular value of x is always the same, regardless of the specific value of x.

(c) The sample size—that is, the number of paired observations (x, y)—exceeds 30.

(d) There exists a straight line $y = \alpha + \beta x$ such that, for each value of x, the mean μ_y of the corresponding population of y-values lies on that straight line.

(e) The data come from a random sample or a randomized experiment.

T12.2 Students in a statistics class drew circles of varying diameters and counted how many Cheerios could be placed in the circle. The scatterplot shows the results.



The students want to determine an appropriate equation for the relationship between diameter and the number of Cheerios. The students decide to transform the data to make it appear more linear before computing a least-squares regression line. Which of the following transformations would be reasonable for them to try?

I. Plot the square root of the number of Cheerios against diameter.

II. Plot the cube of the number of Cheerios against diameter.

III. Plot the log of the number of Cheerios against the log of the diameter.

IV. Plot the number of Cheerios against the log of the diameter.

(a) I and II **(c)** II and III **(e)** I and IV

(b) I and III **(d)** II and IV

T12.3 Inference about the slope β of a least-squares regression line is based on which of the following distributions?

(a) The t distribution with $n - 1$ degrees of freedom

(b) The standard Normal distribution

(c) The chi-square distribution with $n - 1$ degrees of freedom

(d) The t distribution with $n - 2$ degrees of freedom

(e) The Normal distribution with mean μ and standard deviation σ

Exercises T12.4 through T12.8 refer to the following setting. An old saying in golf is "You drive for show and you putt for dough." The point is that good putting is more important than long driving for shooting low scores and hence winning money. To see if this is the case, data from a random sample of 69 of the nearly 1000 players on the PGA Tour's world money list are examined. The average number of putts per hole and the player's total winnings for the previous season are recorded. A least-squares regression line was fitted to the data. The following results were obtained from statistical software.

Predictor	Coef	SE Coef	T	P
Constant	7897179	3023782	6.86	0.000
Avg. Putts	−4139198	1698371	****	****

S = 281777 R-Sq = 8.1% R-Sq(adj) = 7.8%

T12.4 The correlation between total winnings and average number of putts per hole for these players is

(a) −0.285. **(c)** −0.007. **(e)** 0.285.

(b) −0.081. **(d)** 0.081.

T12.5 Suppose that the researchers test the hypotheses $H_0: \beta = 0$, $H_a: \beta < 0$. The value of the t statistic for this test is

(a) 2.61. **(c)** 0.081. **(e)** −20.24.

(b) 2.44. **(d)** −2.44.

T12.6 The P-value for the test in Question T12.5 is 0.0087. A correct interpretation of this result is that

(a) the probability that there is no linear relationship between average number of putts per hole and total winnings for these 69 players is 0.0087.

(b) the probability that there is no linear relationship between average number of putts per hole and total winnings for all players on the PGA Tour's world money list is 0.0087.

(c) if there is no linear relationship between average number of putts per hole and total winnings for the players in the sample, the probability of getting a random sample of 69 players that yields a least-squares regression line with a slope of −4139198 or less is 0.0087.

(d) if there is no linear relationship between average number of putts per hole and total winnings for the players on the PGA Tour's world money list, the probability of getting a random sample of 69 players that yields a least-squares regression line with a slope of −4139198 or less is 0.0087.

(e) the probability of making a Type I error is 0.0087.

T12.7 A 95% confidence interval for the slope β of the population regression line is

(a) 7,897,179 ± 3,023,782.

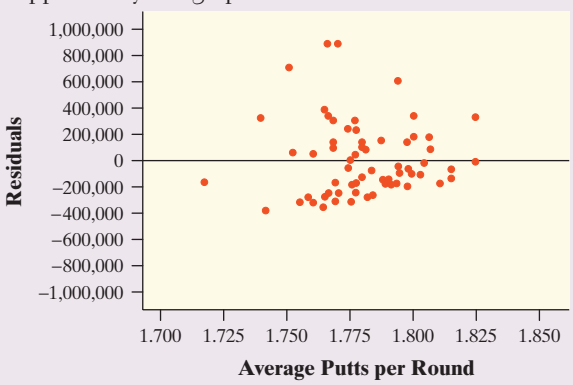
(b) 7,897,179 ± 6,047,564.

(c) −4,139,198 ± 1,698,371.

(d) −4,139,198 ± 3,328,807.

(e) −4,139,198 ± 3,396,742.

T12.8 A residual plot from the least-squares regression is shown below. Which of the following statements is supported by the graph?



(a) The residual plot contains dramatic evidence that the standard deviation of the response about the population regression line increases as the average number of putts per round increases.

(b) The sum of the residuals is not 0. Obviously, there is a major error present.

(c) Using the regression line to predict a player's total winnings from his average number of putts almost always results in errors of less than $200,000.

(d) For two players, the regression line underpredicts their total winnings by more than $800,000.

(e) The residual plot reveals no correlation between average putts per round and prediction errors from the least-squares line for these players.

T12.9 Which of the following would provide evidence that a power law model of the form $y = ax^b$, where $b \neq 0$ and $b \neq 1$, describes the relationship between a response variable y and an explanatory variable x?

(a) A scatterplot of y versus x looks approximately linear.

(b) A scatterplot of ln y versus x looks approximately linear.

(c) A scatterplot of y versus ln x looks approximately linear.

(d) A scatterplot of ln y versus ln x looks approximately linear.

(e) None of these

T12.10 We record data on the population of a particular country from 1960 to 2010. A scatterplot reveals a clear curved relationship between population and year. However, a different scatterplot reveals a strong linear relationship between the logarithm (base 10) of the population and the year. The least-squares regression line for the transformed data is

$$\widehat{\log(\text{population})} = -13.5 + 0.01(\text{year})$$

Based on this equation, the population of the country in the year 2020 should be about

(a) 6.7. **(b)** 812. **(c)** 5,000,000. **(d)** 6,700,000. **(e)** 8,120,000.

Section II: Free Response *Show all your work. Indicate clearly the methods you use, because you will be graded on the correctness of your methods as well as on the accuracy and completeness of your results and explanations.*

T12.11 Growth hormones are often used to increase the weight gain of chickens. In an experiment using 15 chickens, 3 chickens were randomly assigned to each of 5 different doses of growth hormone (0, 0.2, 0.4, 0.8, and 1.0 milligrams). The subsequent weight gain (in ounces) was recorded for each chicken. A researcher plots the data and finds that a linear relationship appears to hold. Computer output from a least-squares regression analysis for these data is shown below. Assume that the conditions for performing inference about the slope β of the true regression line are met.

Predictor	Coef	SE Coef	T	P
Constant	4.5459	0.6166	7.37	<0.0001
Dose	4.8323	1.0164	4.75	0.0004

S = 3.135 R-Sq = 38.4% R-Sq(adj) = 37.7%

(a) What is the equation of the least-squares regression line for these data? Define any variables you use.

(b) Interpret each of the following in context:

(i) The slope

(ii) The y intercept

(iii) s

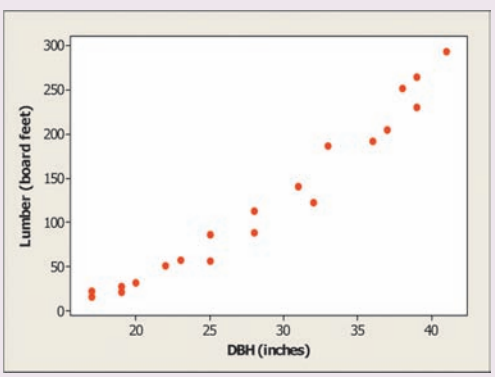
(iv) The standard error of the slope

(v) r^2

(c) Do the data provide convincing evidence of a linear relationship between dose and weight gain? Carry out a significance test at the $\alpha = 0.05$ level.

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

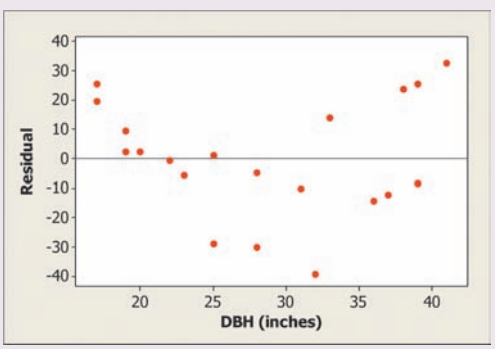
T12.12 Foresters are interested in predicting the amount of usable lumber they can harvest from various tree species. They collect data on the diameter at breast height (DBH) in inches and the yield in board feet of a random sample of 20 Ponderosa pine trees that have been harvested. (Note that a board foot is defined as a piece of lumber 12 inches by 12 inches by 1 inch.) A scatterplot of the data is shown below.



(a) Some computer output and a residual plot from a least-squares regression on these data appear below. Explain why a linear model may not be appropriate in this case.

Predictor	Coef	SE Coef	T	P
Constant	−191.12	16.98	−11.25	0.000
DBH (inches)	11.0413	0.5752	19.19	0.000

S = 20.3290 R-Sq = 95.3% R-Sq(adj) = 95.1%



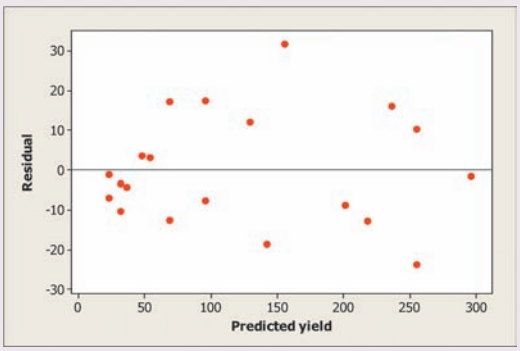
The foresters are considering two possible transformations of the original data: (1) cubing the

diameter values or (2) taking the natural logarithm of the yield measurements. After transforming the data, a least-squares regression analysis is performed. Some computer output and a residual plot for each of the two possible regression models follow.

Option 1: Cubing the diameter values

Predictor	Coef	SE Coef	T	P
Constant	2.078	5.444	0.38	0.707
DBH^3	0.0042597	0.0001549	27.50	0.000

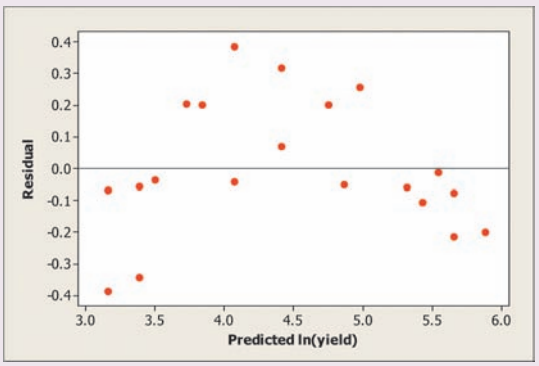
S = 14.3601 R-Sq = 97.7% R-Sq(adj) = 97.5%



Option 2: Taking natural logarithm of yield measurements

Predictor	Coef	SE Coef	T	P
Constant	1.2319	0.1795	6.86	0.000
DBH (inches)	0.113417	0.006081	18.65	0.000

S = 0.214894 R-Sq = 95.1% R-Sq(adj) = 94.8%



(b) Use both models to predict the amount of usable lumber from a Ponderosa pine with diameter 30 inches. Show your work.

(c) Which of the predictions in part (b) seems more reliable? Give appropriate evidence to support your choice.

Cumulative AP® Practice Test 4

Section I: Multiple Choice *Choose the best answer.*

AP4.1 A major agricultural company is testing a new variety of wheat to determine whether it is more resistant to certain insects than is the current wheat variety. The proportion of a current wheat crop lost to insects is 4%. Thus, the company wishes to test the following hypotheses:

$$H_0: p = 0.04$$
$$H_a: p < 0.04$$

Which of the following significance levels and sample sizes would lead to the highest power for this test?

(a) $n = 200$ and $\alpha = 0.01$

(b) $n = 400$ and $\alpha = 0.05$

(c) $n = 400$ and $\alpha = 0.01$

(d) $n = 500$ and $\alpha = 0.01$

(e) $n = 500$ and $\alpha = 0.05$

AP4.2 If $P(A) = 0.24$ and $P(B) = 0.52$ and events A and B are independent, what is $P(A \text{ or } B)$?

(a) 0.1248 **(b)** 0.28

(c) 0.6352 **(d)** 0.76

(e) The answer cannot be determined from the given information.

AP4.3 As part of a bear population study, data were gathered on a sample of black bears in the western United States to examine the relationship between the bear's neck girth (distance around the neck) and the weight of the bear. A scatterplot of the data reveals a straight-line pattern. The r^2-value from a least-squares regression analysis was determined to be $r^2 = 0.963$. Which one of the following is the correct value and corresponding interpretation for the correlation?

(a) The correlation is -0.963, and 96.3% of the variation in a bear's weight can be explained by the least-squares regression line using neck girth as the explanatory variable.

(b) The correlation is 0.963. There is a strong positive linear relationship between a bear's neck girth and its weight.

(c) The correlation is 0.981, and 98.1% of the variation in a bear's weight can be explained by the least-

squares regression line using neck girth as the explanatory variable.

(d) The correlation is −0.981. There is a strong negative linear relationship between a bear's neck girth and its weight.

(e) The correlation is 0.981. There is a strong positive linear relationship between a bear's neck girth and its weight.

AP4.4 The school board in a certain school district obtained a random sample of 200 residents and asked if they were in favor of raising property taxes to fund the hiring of more statistics teachers. The resulting confidence interval for the true proportion of residents in favor of raising taxes was (0.183, 0.257). The margin of error for this confidence interval is

(a) 0.037. (c) 0.220. (e) 0.740.

(b) 0.183. (d) 0.257.

AP4.5 After a name-brand drug has been sold for several years, the Food and Drug Administration (FDA) will allow other companies to produce a generic equivalent. The FDA will permit the generic drug to be sold as long as there isn't convincing evidence that it is less effective than the name brand drug. For a proposed generic drug intended to lower blood pressure, the following hypotheses will be used:

$$H_0: \mu_G = \mu_N \text{ versus } H_a: \mu_G < \mu_N$$

where

μ_G = true mean reduction in blood pressure using the generic drug

μ_N = true mean reduction in blood pressure using the name-brand drug.

In the context of this situation, which of the following describes a Type I error?

(a) The FDA finds convincing evidence that the generic drug is less effective, when in reality it is less effective.

(b) The FDA finds convincing evidence that the generic drug is less effective, when in reality it is equally effective.

(c) The FDA fails to find convincing evidence that the generic drug is less effective, when in reality it is less effective.

(d) The FDA fails to find convincing evidence that the generic drug is less effective, when in reality it is equally effective.

(e) The FDA finds convincing evidence that the generic drug is equally effective, when in reality it is less effective.

AP4.6 Which of the following sampling plans for estimating the proportion of all adults in a medium-sized town who favor a tax increase to support the local school system does *not* suffer from undercoverage bias?

(a) A random sample of 250 names from the local phone book

(b) A random sample of 200 parents whose children attend one of the local schools

(c) A sample consisting of 500 people from the city who take an online survey about the issue

(d) A random sample of 300 homeowners in the town

(e) A random sample of 100 people from an alphabetical list of all adults who live in the town

AP4.7 Which of the following is a categorical variable?

(a) The weight of automobiles

(b) The time required to complete the Olympic marathon

(c) The average gas mileage of a hybrid car

(d) The brand of shampoo purchased by shoppers in a grocery store

(e) The average closing price of a particular stock on the New York Stock Exchange

AP4.8 A large machine is filled with thousands of small pieces of candy, 40% of which are orange. When money is deposited, the machine dispenses 60 randomly selected pieces of candy. The machine will be recalibrated if a group of 60 candies contains fewer than 18 that are orange. What is the approximate probability that this will happen if the machine is working correctly?

(a) $P\left(Z < \dfrac{0.3 - 0.4}{\sqrt{\dfrac{(0.4)(0.6)}{60}}}\right)$ (d) $P\left(Z < \dfrac{0.3 - 0.4}{\dfrac{(0.4)(0.6)}{\sqrt{60}}}\right)$

(b) $P\left(Z < \dfrac{0.3 - 0.4}{\sqrt{\dfrac{(0.3)(0.7)}{60}}}\right)$ (e) $P\left(Z < \dfrac{0.4 - 0.3}{\dfrac{\sqrt{(0.3)(0.7)}}{60}}\right)$

(c) $P\left(Z < \dfrac{0.3 - 0.4}{\dfrac{\sqrt{(0.4)(0.6)}}{60}}\right)$

AP4.9 A random sample of 900 students at a very large university was asked which social-networking site they used most often during a typical week. Their responses are shown in the table below.

Networking site	Male	Female	Total
Facebook	221	283	**504**
Twitter	42	38	**80**
LinkedIn	108	87	**195**
Pinterest	23	26	**49**
MySpace	29	43	**72**
Total	**423**	**477**	**900**

Assuming that gender and preferred networking site are independent, how many females do you expect to choose LinkedIn?

(a) 18.85 (c) 87.00 (e) 103.35
(b) 46.11 (d) 91.65

AP4.10 Insurance adjusters are always vigilant about being overcharged for accident repairs. The adjusters suspect that Repair Shop 1 quotes higher estimates than Repair Shop 2. To check their suspicion, the adjusters randomly select 12 cars that were recently involved in an accident and then take each of the cars to both repair shops to obtain separate estimates of the cost to fix the vehicle. The estimates are given below in hundreds of dollars.

Car:	1	2	3	4	5	6
Shop 1:	21.2	25.2	39.0	11.3	15.0	18.1
Shop 2:	21.3	24.1	36.8	11.5	13.7	17.6

Car:	7	8	9	10	11	12
Shop 1:	25.3	23.2	12.4	42.6	27.6	12.9
Shop 2:	24.8	21.3	12.1	42.0	26.7	12.5

Assuming that the conditions for inference are reasonably met, which of the following significance tests could legitimately be used to determine whether the adjusters' suspicion is correct?

(a) A paired t test
(b) A two-sample t test
(c) A t test to see if the slope of the population regression line is 0
(d) A chi-square test for homogeneity
(e) A two-sample z test for comparing two proportions

AP4.11 A survey firm wants to ask a random sample of adults in Ohio if they support an increase in the state sales tax from 5% to 6%, with the additional revenue going to education. Let $\hat{p}$ denote the proportion in the sample who say that they support the increase. Suppose that 40% of all adults in Ohio support the increase. How large a sample would be needed to guarantee that the standard deviation of $\hat{p}$ is no more than 0.01?

(a) 1500 (c) 2401 (e) 9220
(b) 2400 (d) 2500

AP4.12 A set of 10 cards consists of 5 red cards and 5 black cards. The cards are shuffled thoroughly, and you choose one at random, observe its color, and replace it in the set. The cards are thoroughly reshuffled, and you again choose a card at random, observe its color, and replace it in the set. This is done a total of four times. Let X be the number of red cards observed in these four trials. The random variable X has which of the following probability distributions?

(a) The Normal distribution with mean 2 and standard deviation 1
(b) The binomial distribution with $n = 10$ and $p = 0.5$
(c) The binomial distribution with $n = 5$ and $p = 0.5$
(d) The binomial distribution with $n = 4$ and $p = 0.5$
(e) The geometric distribution with $p = 0.5$

AP4.13 A study of road rage asked random samples of 596 men and 523 women about their behavior while driving. Based on their answers, each respondent was assigned a road rage score on a scale of 0 to 20. The respondents were chosen by random digit dialing of telephone numbers. Are the conditions for two-sample t inference satisfied?

(a) Maybe. The data came from independent random samples, but we need to examine the data to check for Normality.
(b) No. Road rage scores in a range between 0 and 20 can't be Normal.
(c) No. A paired t test should be used in this case.
(d) Yes. The large sample sizes guarantee that the corresponding population distributions will be Normal.
(e) Yes. We have two independent random samples and large sample sizes, and the 10% condition is met.

AP4.14 Do hummingbirds prefer store-bought food made from concentrate or a simple mixture of sugar and water? To find out, a researcher obtains 10 identical hummingbird feeders and fills 5, chosen at random, with store-bought food from concentrate and the other 5 with a mixture of sugar and water. The feeders are then randomly assigned to 10 possible hanging locations in the researcher's yard. Which inference procedure should you use to test whether hummingbirds show a preference for store-bought food based on amount consumed?

(a) A one-sample z test for a proportion
(b) A two-sample z test for a difference in proportions
(c) A chi-square test for independence
(d) A two-sample t test
(e) A paired t test

AP4.15 A Harris Poll found that 54% of American adults don't think that human beings developed from earlier species. The poll's margin of error for 95% confidence was 3%. This means that

(a) there is a 95% chance that the interval (51%, 57%) contains the true percent of American adults who do not think that human beings developed from earlier species.

(b) the poll used a method that provides an estimate within 3% of the truth about the population 95% of the time.

(c) if Harris takes another poll using the same method, the results of the second poll will lie between 51% and 57%.

(d) there is a 3% chance that the interval is correct.

(e) the poll used a method that would result in an interval that contains 54% in 95% of all possible samples of the same size from this population.

AP4.16 Two six-sided dice are rolled and the sum of the faces showing is recorded after each roll. Let X = the number of rolls required until a sum greater than 7 is obtained. If 100 trials are conducted, which of the following is most likely to be part of the probability distribution of X?

(a)

Number of rolls *X*	Number of trials
1	34
2	20
3	16
4	10
5	6
6	6
7	3
8	2
9	1
10	0
11	1
12	0
13	1

(b)

Number of rolls *X*	Number of trials
0	34
1	20
2	16
3	10
4	6
5	6
6	3
7	2
8	1
9	0
10	1
11	0
12	1

(c)

Number of rolls *X*	Number of trials
1	18
2	23
3	26
4	15
5	9
6	6
7	1
8	0
9	1
10	0
11	0
12	0
13	1

(d)

Number of rolls *X*	Number of trials
1	10
2	9
3	10
4	12
5	7
6	13
7	10
8	7
9	9
10	10
11	2
12	1

(e)

Number of rolls *X*	Number of trials
1	2
2	2
3	5
4	10
5	11
6	15
7	22
8	17
9	9
10	4
11	2
12	0
13	1

AP4.17 Women who are severely overweight suffer economic consequences, a study has shown. They have household incomes that are an average of $6710 lower. The findings are from an eight-year observational study of 10,039 randomly selected women who were 16 to 24 years old when the research began. Does this study give strong evidence that being severely overweight causes a woman to have a lower income?

(a) Yes. The study included both women who were severely overweight and women who were not.

(b) Yes. The subjects in the study were selected at random.

(c) No. The study showed that there is no connection between income and being severely overweight.

(d) No. The study suggests an association between income and being severely overweight, but we can't draw a cause-and-effect conclusion.

(e) There is not enough information to answer this question.

Questions AP 4.18 and 4.19 refer to the following situation. Could mud wrestling be the cause of a rash contracted by University of Washington students? Two physicians at the University of Washington student health center wondered about this when one male and six female students complained of rashes after participating in a mud-wrestling event. Questionnaires were sent to a random sample of students who participated in the event. The results, by gender, are summarized in the following table.

	Men	Women
Developed rash	12	12
No rash	38	12

Some Minitab output for the previous table is given below. The output includes the observed counts, the expected counts, and the chi-square statistic.

```
Expected counts are printed below observed
counts

                    MEN      WOMEN     Total
Developed rash       12         12        24
                  16.22       7.78

No rash              38         12        50
                  33.78      16.22

Total                50         24        74

ChiSq = 5.002
```

AP4.18 The cell that contributes most to the chi-square statistic is

(a) men who developed a rash.

(b) men who did not develop a rash.

(c) women who developed a rash.

(d) women who did not develop a rash.

(e) both (a) and (d).

AP4.19 From the chi-square test performed in this study, we may conclude that

(a) there is convincing evidence of an association between the gender of an individual participating in the event and development of a rash.

(b) mud wrestling causes a rash, especially for women.

(c) there is absolutely no evidence of any relation between the gender of an individual participating in the event and the subsequent development of a rash.

(d) development of a rash is a real possibility if you participate in mud wrestling, especially if you do so regularly.

(e) the gender of the individual participating in the event and the development of a rash are independent.

AP4.20 Random assignment is part of a well-designed comparative experiment because

(a) it is more fair to the subjects.

(b) it helps create roughly equivalent groups before treatments are imposed on the subjects.

(c) it allows researchers to generalize the results of their experiment to a larger population.

(d) it helps eliminate any possibility of bias in the experiment.

(e) it prevents the placebo effect from occurring.

AP4.21 The following back-to-back stemplots compare the ages of players from two minor-league hockey teams (1|7 = 17 years).

Team A		Team B
98777	1	788889
44333221	2	00123444
7766555	2	556679
521	3	023
86	3	55

Which of the following *cannot* be justified from the plots?

(a) Team A has the same number of players in their 30s as does Team B.

(b) The median age of both teams is the same.

(c) Both age distributions are skewed to the right.

(d) The age ranges of both teams are similar.

(e) There are no outliers by the 1.5*IQR* rule in either distribution.

AP4.22 A distribution that represents the number of cars X parked in a randomly selected residential driveway on any night is given by

x_i:	0	1	2	3	4
p_i:	0.1	0.2	0.35	0.25	0.15

Which of the following statements is correct?

(a) This is a legitimate probability distribution because each of the p_i-values is between 0 and 1.

(b) This is a legitimate probability distribution because $\sum x_i$ is exactly 10.

(c) This is a legitimate probability distribution because each of the p_i-values is between 0 and 1 and the $\sum x_i$ is exactly 10.

(d) This is not a legitimate probability distribution because $\sum x_i$ is not exactly 10.

(e) This is not a legitimate probability distribution because $\sum p_i$ is not exactly 1.

AP4.23 Which sampling method was used in each of the following settings, in order from I to IV?

I. A student chooses for a survey the first 20 students to arrive at school.

II. The name of each student in a school is written on a card, the cards are well mixed, and 10 names are drawn.

III. A state agency randomly selects 50 people from each of the state's senatorial districts.

IV. A city council randomly selects eight city blocks and then surveys all the voting-age residents of those blocks.

(a) Voluntary response, SRS, stratified, cluster
(b) Convenience, SRS, stratified, cluster
(c) Convenience, cluster, SRS, stratified
(d) Convenience, SRS, cluster, stratified
(e) Cluster, SRS, stratified, convenience

AP4.24 Western lowland gorillas, whose main habitat is the central African continent, have a mean weight of 275 pounds with a standard deviation of 40 pounds. Capuchin monkeys, whose main habitat is Brazil and a few other parts of Latin America, have a mean weight of 6 pounds with a standard deviation of 1.1 pounds. Both weight distributions are approximately Normally distributed. If a particular western lowland gorilla is known to weigh 345 pounds, approximately how much would a capuchin monkey have to weigh, in pounds, to have the same standardized weight as the lowland gorilla?

(a) 4.08 (c) 7.93
(b) 7.27 (d) 8.20
(e) There is not enough information to determine the weight of a capuchin monkey.

AP4.25 Suppose that the mean weight of a certain type of pig is 280 pounds with a standard deviation of 80 pounds. The weight distribution of pigs tends to be somewhat skewed to the right. A random sample of 100 pigs is taken. Which of the following statements about the sampling distribution of the sample mean weight $\bar{x}$ is true?

(a) It will be Normally distributed with a mean of 280 pounds and a standard deviation of 80 pounds.
(b) It will be Normally distributed with a mean of 280 pounds and a standard deviation of 8 pounds.
(c) It will be approximately Normally distributed with a mean of 280 pounds and a standard deviation of 80 pounds.
(d) It will be approximately Normally distributed with a mean of 280 pounds and a standard deviation of 8 pounds.
(e) There is not enough information to determine the mean and standard deviation of the sampling distribution.

AP4.26 Which of the following statements about the t distribution with degrees of freedom df is (are) true?

I. It is symmetric.

II. It has more variability than the t distribution with df + 1 degrees of freedom.

III. As df increases, the t distribution approaches the standard Normal distribution.

(a) I only (c) III only (e) I, II, and III
(b) II only (d) I and III

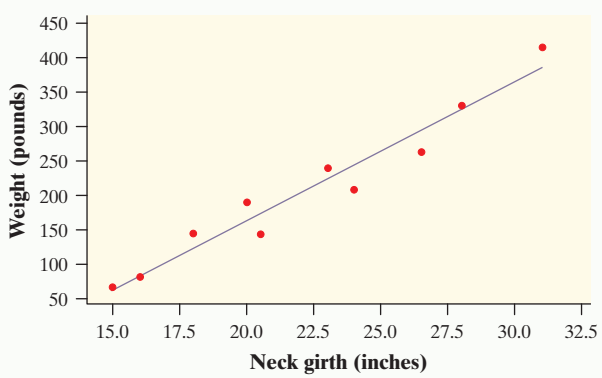
AP4.27 A company has been running television commercials for a new children's product on five different family programs during the evening hours in a large city over a one-month period. A random sample of families is taken, and they are asked to indicate which of the five programs they viewed most often and their rating of the advertised product. The results are summarized in the following table.

	Family program				
Product rating	**A**	**B**	**C**	**D**	**E**
Excellent	23	29	42	48	51
Good	25	33	44	53	49
Fair	31	29	25	16	10
Poor	38	32	25	18	12

The advertiser decided to use a chi-square test to see if there is a relationship between the family program viewed and the product's rating. What would be the degrees of freedom for this test?

(a) 3 (c) 12 (e) 19
(b) 4 (d) 18

Questions AP4.28 and AP4.29 refer to the following situation. Park rangers are interested in estimating the weight of the bears that inhabit their state. The rangers have data on weight (in pounds) and neck girth (distance around the neck in inches) for 10 randomly selected bears. Some regression output for these data is shown below.



Predictor	Coef	SE Coef	T	P
Constant	−241.70	38.57	−6.27	0.000
Neck Girth	20.230	1.695	11.93	0.000
S = 26.7565	R-Sq = 94.7%			

AP4.28 Which of the following represents a 95% confidence interval for the true slope of the least-squares regression line relating the weight of a bear and its neck girth?

(a) 20.230 ± 1.695
(b) 20.230 ± 3.83
(c) 20.230 ± 3.91
(d) 20.230 ± 20.22
(e) 26.7565 ± 3.83

AP4.29 A bear was recently captured whose neck girth was 35 inches and whose weight was 466.35 pounds. If this bear were added to the data set given above, what would be the effect on the value of s?

(a) It would decrease the value of s because the added point is an outlier.
(b) It would decrease the value of s because the added point lies on the least-squares regression line.
(c) It would increase the value of s because the added point is an outlier.
(d) It would increase the value of s because the added point lies on the least-squares regression line.
(e) It would have no effect on the value of s because the added point lies on the least-squares regression line.

AP4.30 An experimenter wishes to test whether or not two types of fish food (a standard fish food and a new product) work equally well at producing fish of equal weight after a two-month feeding program. The experimenter has two identical fish tanks (1 and 2) to put fish in and is considering how to assign 40 fish, each of which has a numbered tag, to the tanks. The best way to do this would be to

(a) put all the odd-numbered fish in one tank, the even in the other, and give the standard food type to the odd-numbered ones.
(b) obtain pairs of fish whose weights are roughly equal at the start of the experiment and randomly assign one to Tank 1 and the other to Tank 2, with the feed assigned at random to the tanks.
(c) proceed as in part (b), but put the heavier of the pair into Tank 2.
(d) assign the fish completely at random to the two tanks and give the standard feed to Tank 1.
(e) assign the fish to the tanks using any method that the researcher wants. The placebo effect doesn't apply to fish.

AP4.31 A city wants to conduct a poll of taxpayers to determine the level of support for constructing a new city-owned baseball stadium. Which of the following is the primary reason for using a large sample size in constructing a confidence interval to estimate the proportion of city taxpayers who would support such a project?

(a) To increase the confidence level
(b) To eliminate any confounding variables
(c) To reduce nonresponse bias
(d) To increase the precision of the estimate
(e) To reduce undercoverage

AP4.32 A standard deck of playing cards contains 52 cards, of which 4 are aces and 13 are hearts. You are offered a choice of the following two wagers:

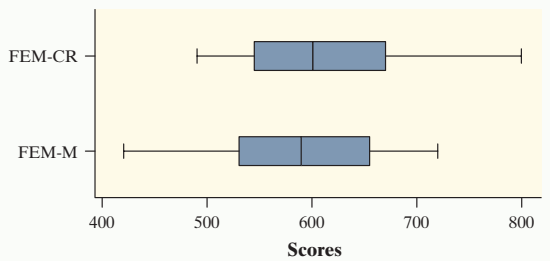
I. Draw one card at random from the deck. You win \$10 if the card drawn is an ace. Otherwise, you lose \$1.

II. Draw one card at random from the deck. If the card drawn is a heart, you win \$2. Otherwise, you lose \$1.

Which of the two wagers should you prefer?

(a) Wager 1, because it has a higher expected value
(b) Wager 2, because it has a higher expected value
(c) Wager 1, because it has a higher probability of winning
(d) Wager 2, because it has a higher probability of winning
(e) Both wagers are equally favorable.

AP4.33 Below are boxplots of SAT Critical Reading and Math scores for a randomly selected group of female juniors at a highly competitive suburban school.



Which of the following *cannot* be justified by the plots shown above?

(a) The maximum Critical Reading score is higher than the maximum Math score.
(b) Critical Reading scores are skewed to the right, whereas Math scores are somewhat skewed to the left.
(c) The median Critical Reading score for females is slightly higher than the median Math score.
(d) There appear to be no outliers in the SAT score distributions.
(e) The mean Critical Reading score and the mean Math score for females are about the same.

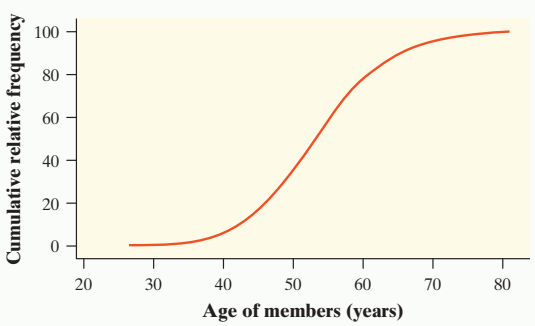
AP4.34 A distribution of exam scores has mean 60 and standard deviation 18. If each score is doubled, and then 5 is subtracted from that result, what will be the mean and standard deviation, respectively, of the new scores?

(a) mean = 115 and standard deviation = 31
(b) mean = 115 and standard deviation = 36
(c) mean = 120 and standard deviation = 6
(d) mean = 120 and standard deviation = 31
(e) mean = 120 and standard deviation = 36

AP4.35 In a clinical trial, 30 patients with a certain blood disease are randomly assigned to two groups. One group is then randomly assigned the currently marketed medicine, and the other group receives the experimental medicine. Each week, patients report to the clinic where blood tests are conducted. The lab technician is unaware of the kind of medicine the patient is taking, and the patient is also unaware of which medicine he or she has been given. This design can be described as

(a) a double-blind, completely randomized experiment, with the currently marketed medicine and the experimental medicine as the two treatments.
(b) a single-blind, completely randomized experiment, with the currently marketed medicine and the experimental medicine as the two treatments.
(c) a double-blind, matched pairs design, with the currently marketed medicine and the experimental medicine forming a pair.
(d) a double-blind, block design that is not a matched pairs design, with the currently marketed medicine and the experimental medicine as the two blocks.
(e) a double-blind, randomized observational study.

AP4.36 A local investment club that meets monthly has 200 members ranging in age from 27 to 81. A cumulative relative frequency graph is shown below. Approximately how many members of the club are more than 60 years of age?



(a) 20 (c) 78 (e) 110
(b) 44 (d) 90

AP4.37 A manufacturer of electronic components is testing the durability of a newly designed integrated circuit to determine whether its life span is longer than that of the earlier model, which has a mean life span of 58 months. The company takes a simple random sample of 120 integrated circuits and simulates normal use until they stop working. The null and alternative hypotheses used for the significance test are given by $H_0: \mu = 58$ and $H_a: \mu > 58$. The P-value for the resulting one-sample t test is 0.035. Which of the following best describes what the P-value measures?

(a) The probability that the new integrated circuit has the same life span as the current model is 0.035.
(b) The probability that the test correctly rejects the null hypothesis in favor of the alternative hypothesis is 0.035.
(c) The probability that a single new integrated circuit will not last as long as one of the earlier circuits is 0.035.
(d) The probability of getting a sample statistic as far or farther from 58 if there really is no difference between the new and the old circuits is 0.035.
(e) The probability of getting a sample mean for the new integrated circuit that is lower than the mean for the earlier model is 0.035.

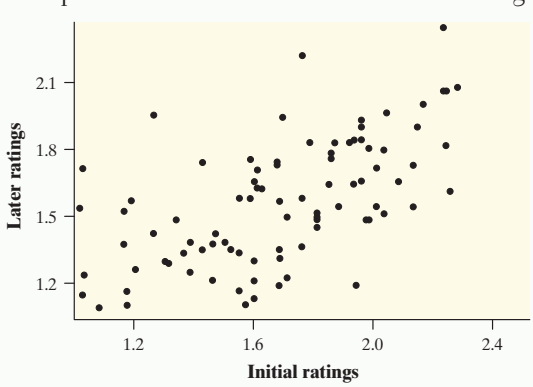
Questions AP4.38 and AP4.39 refer to the following situation. Do children's fear levels change over time and, if so, in what ways? Little research has been done on the prevalence and persistence of fears in children. Several years ago, two researchers surveyed a randomly selected group of 94 third- and fourth-grade children, asking them to rate their level of fearfulness about a variety of situations. Two years later, the children again completed the same survey. The researchers computed the overall fear rating for each child in both years and were interested in the relationship between these ratings. They then assumed that the true regression line was

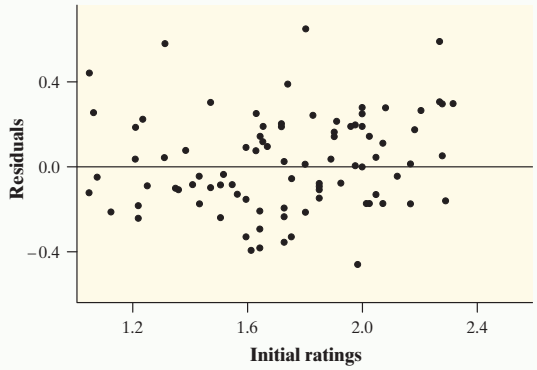
$$\mu_{\text{later rating}} = \alpha + \beta\,(\text{initial rating})$$

and that the assumptions for regression inference were satisfied. This model was fitted to the data using least-squares regression. The following results were obtained from statistical software.

Predictor	Coefficient	St. Dev.
Constant	0.877917	0.1184
Initial Rating	0.397911	0.0676
S = 0.2374	R-Sq = 0.274	

Here is a scatterplot of the later ratings versus the initial ratings and a plot of the residuals versus the initial ratings.





AP4.38 Which of the following statements is supported by these plots?

(a) There is no striking evidence that the assumptions for regression inference are violated.

(b) The abundance of outliers and influential observations in the plots means that the assumptions for regression are clearly violated.

(c) These plots contain dramatic evidence that the standard deviation of the response about the true regression line is not approximately the same for each x-value.

(d) These plots call into question the validity of the assumption that the later ratings vary Normally about the least-squares line for each value of the initial ratings.

(e) A linear model isn't appropriate here because the residual plot shows no association.

AP4.39 George's initial fear rating was 0.2 higher than Jonny's. What does the model predict about their final fear ratings?

(a) George's will be about 0.96 higher than Jonny's.

(b) George's will be about 0.40 higher than Jonny's.

(c) George's will be about 0.20 higher than Jonny's

(d) George's will be about 0.08 higher than Jonny's.

(e) George's will be about the same as Jonny's.

AP4.40 The table below provides data on the political affiliation and opinion about the death penalty of 850 randomly selected voters from a congressional district.

	Favor	Oppose	Total
Republican	299	98	**397**
Democrat	77	171	**248**
Other	118	87	**205**
Total	**494**	**356**	**850**

Which of the following does *not* support the conclusion that being a Republican and favoring the death penalty are not independent?

(a) $\frac{299}{494} \neq \frac{98}{356}$

(b) $\frac{299}{494} \neq \frac{397}{850}$

(c) $\frac{494}{850} \neq \frac{299}{397}$

(d) $\frac{494}{850} \neq \frac{397}{850}$

(e) $\frac{(397)(494)}{850} \neq 299$

Section II: Free Response *Show all your work. Indicate clearly the methods you use, because you will be graded on the correctness of your methods as well as on the accuracy and completeness of your results and explanations.*

AP4.41 The body's natural electrical field helps wounds heal. If diabetes changes this field, it might explain why people with diabetes heal more slowly. A study of this idea compared randomly selected normal mice and randomly selected mice bred to spontaneously develop diabetes. The investigators attached sensors to the right hip and front feet of the mice and measured the difference in electrical potential (in millivolts) between these locations. Graphs of the data for each group reveal no outliers or strong skewness. The following computer output provides numerical summaries of the data.[27]

Variable	N	Mean	StDev	Minimum
Diabetic mice	24	13.090	4.839	1.050
Normal mice	18	10.022	2.915	4.950

Q1	Median	Q3	Maximum
10.038	12.650	17.038	22.600
8.238	9.250	12.375	16.100

The researchers want to know whether the difference in mean electrical potentials between normal mice and mice with diabetes is statistically significant at the $\alpha = 0.05$ level. Carry out a test and report your conclusion.

AP4.42 Can physical activity in youth lead to mental sharpness in old age? A 2010 study investigating this question involved 9344 randomly selected, mostly white women over age 65 from four U.S. states. These women were asked about their levels of physical activity during their teenage years, thirties, fifties, and later years. Those who reported being physically active as teens enjoyed the lowest level of cognitive decline—only 8.5% had cognitive impairment—compared with 16.7% of women who reported not being physically active at that time.

(a) State an appropriate pair of hypotheses that the researchers could use to test whether the proportion of women who suffered a cognitive decline was

significantly lower for women who were physically active in their youth than for women who were not physically active at that time. Be sure to define any parameters you use.

(b) Assuming the conditions for performing inference are met, what inference method would you use to test the hypotheses you identified in part (b)? Do *not* carry out the test.

(c) Suppose the test in part (b) shows that the proportion of women who suffered a cognitive decline was significantly lower for women who were physically active in their youth than for women who were not physically active at that time. Can we generalize the results of this study to all women aged 65 and older? Justify your answer.

(d) We cannot conclude that being physically active as a teen *causes* a lower level of cognitive decline for women over 65, due to possible confounding with other variables. Explain the concept of confounding and give an example of a potential confounding variable in this study.

AP4.43 In a recent poll, randomly selected New York State residents at various fast-food restaurants were asked if they supported or opposed a "fat tax" on nondiet sugared soda. Thirty-one percent said that they were in favor of such a tax and 66% were opposed. But when asked if they would support such a tax if the money raised were used to fund health care given the high incidence of obesity in the United States, 48% said that they were in favor and 49% were opposed.

(a) In this situation, explain how bias may have been introduced based on the way the questions were worded *and* suggest a way that they could have been worded differently in order to avoid this bias.

(b) In this situation, explain how bias may have been introduced based on the way the sample was taken *and* suggest a way that the sample could have been obtained in order to avoid this bias.

(c) This poll was conducted only in New York State. Suppose the pollsters wanted to ensure that estimates for the proportion of people who would support a tax on nondiet sugared soda were available for each state as well as an overall estimate for the nation as a whole. Identify a sampling method that would achieve this goal *and* briefly describe how the sample would be taken.

AP4.44 Each morning, coffee is brewed in the school workroom by one of three faculty members, depending on who arrives first at work. Mr. Worcester arrives first 10% of the time, Dr. Currier arrives first 50% of the time, and Mr. Legacy arrives first on the remaining mornings. The probability that the coffee is strong when brewed by Dr. Currier is 0.1, while the corresponding probabilities when it is brewed by Mr. Legacy and Mr. Worcester are 0.2 and 0.3, respectively. Mr. Worcester likes strong coffee!

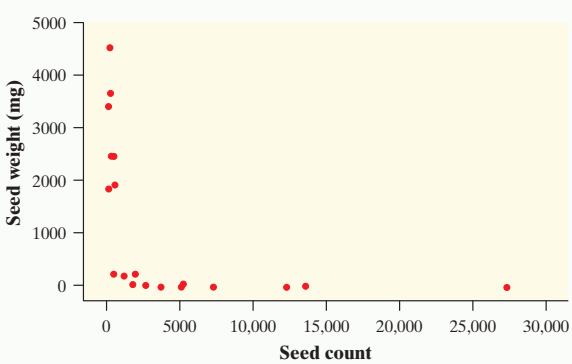
(a) What is the probability that on a randomly selected morning the coffee will be strong? Show your work.

(b) If the coffee is strong on a randomly selected morning, what is the probability that it was brewed by Dr. Currier? Show your work.

AP4.45 The following table gives data on the mean number of seeds produced in a year by several common tree species and the mean weight (in milligrams) of the seeds produced. Two species appear twice because their seeds were counted in two locations. We might expect that trees with heavy seeds produce fewer of them, but what mathematical model best describes the relationship?[28]

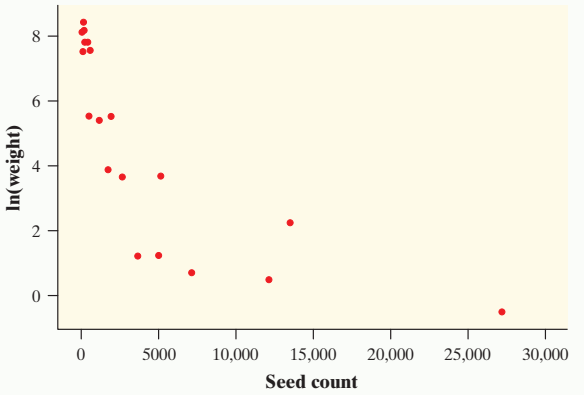
Tree species	Seed count	Seed weight (mg)
Paper birch	27,239	0.6
Yellow birch	12,158	1.6
White spruce	7202	2.0
Engelmann spruce	3671	3.3
Red spruce	5051	3.4
Tulip tree	13,509	9.1
Ponderosa pine	2667	37.7
White fir	5196	40.0
Sugar maple	1751	48.0
Sugar pine	1159	216
American beech	463	247
American beech	1892	247
Black oak	93	1851
Scarlet oak	525	1930
Red oak	411	2475
Red oak	253	2475
Pignut hickory	40	3423
White oak	184	3669
Chestnut oak	107	4535

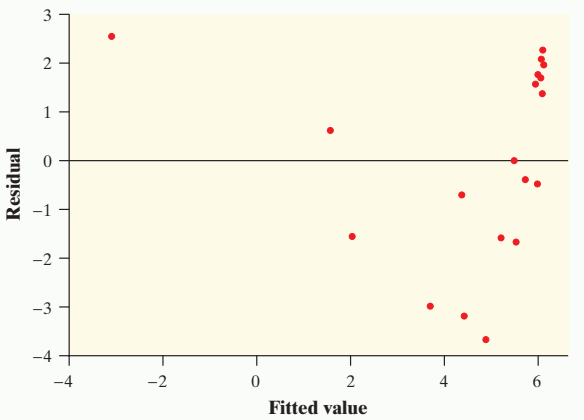
(a) Based on the scatterplot below, is a linear model appropriate to describe the relationship between seed count and seed weight? Explain.



(b) Two alternative models based on transforming the original data are proposed to predict the seed weight from the seed count. Graphs and computer output from a least-squares regression analysis on the transformed data are shown below.

Model A:

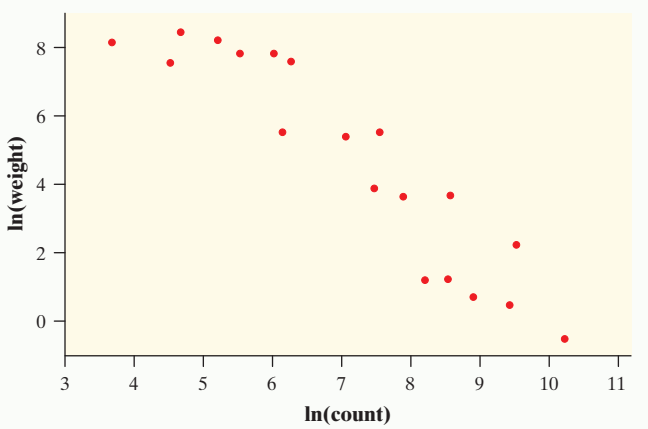


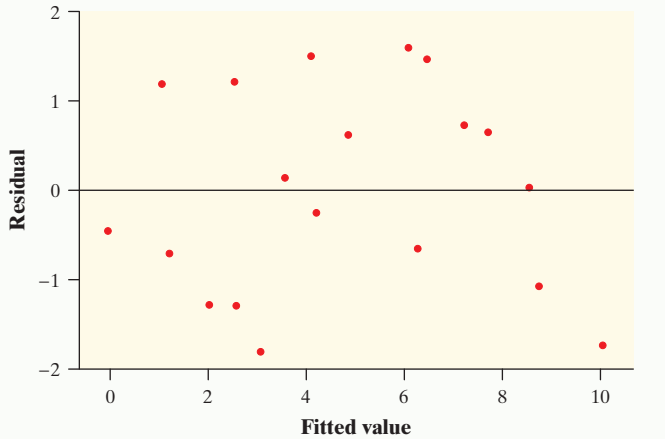


Predictor	Coef	SE Coef	T	P
Constant	6.1394	0.5726	10.72	0.000
Seed Count	−0.00033869	0.00007187	−4.71	0.000

S = 2.08100 R-Sq = 56.6% R-Sq(adj) = 54.1%

Model B:





Predictor	Coef	SE Coef	T	P
Constant	15.491	1.081	14.33	0.000
ln(count)	−1.5222	0.1470	−10.35	0.000

S = 1.16932 R-Sq = 86.3% R-Sq(adj) = 85.5%

Which model, A or B, is more appropriate for predicting seed weight from seed count? Justify your answer.

(c) Using the model you chose in part (b), predict the seed weight if the seed count is 3700.

(d) Interpret the value of r^2 for your model.

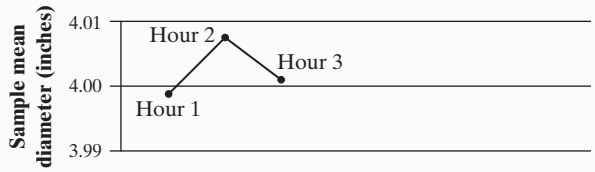
AP4.46 Suppose a company manufactures plastic lids for disposable coffee cups. When the manufacturing process is working correctly, the diameters of the lids are approximately Normally distributed with a mean diameter of 4 inches and a standard deviation of 0.02 inches. To make sure the machine is not producing lids that are too big or too small, each hour a random sample of 25 lids is selected and the sample mean is calculated.

(a) Describe the shape, center, and spread of the sampling distribution of the sample mean diameter, assuming the machine is working properly.

The company decides that it will shut down the machine if the sample mean diameter is less than 3.99 inches or greater than 4.01 inches, because this indicates that some lids will be too small or too large for the cups. If the sample mean is less than 3.99 or greater than 4.01, all the lids from that hour are thrown away because the company does not want to sell bad products.

(b) Assuming that the machine is working properly, what is the probability that a random sample of 25 lids will have a mean diameter less than 3.99 inches or greater than 4.01 inches? Show your work.

Also, to look for any trends, each hour the company records the value of the sample mean on a chart, like the one at top right.



One benefit of using this type of chart is that out-of-control production trends can be noticed before it is too late and lids have to be thrown away. For example, if the sample mean increased in 3 consecutive samples, this would suggest that something might be wrong with the machine. If this trend can be noticed before the sample mean gets larger than 4.01, then the machine can be fixed without having to throw away any lids.

(c) Assuming that the manufacturing process is working correctly, what is the probability that the sample mean diameter will be above the desired mean of 4.00 but below the upper boundary of 4.01? Show your work.

(d) Assuming that the manufacturing process is working correctly, what is the probability that in 5 consecutive samples, 4 or 5 of the sample means will be above the desired mean of 4.00 but below the upper boundary of 4.01? Show your work.

(e) Which of the following results gives more convincing evidence that the machine needs to be shut down? Explain.

1. Getting a single sample mean below 3.99 or above 4.01

or

2. Taking 5 consecutive samples and having at least 4 of the sample means be between 4.00 and 4.01.

(f) Suggest a different rule (other than 1 and 2 stated in part (e)) for stopping the machine before it starts producing lids that have to be thrown away. Assuming that the machine is working properly, calculate the probability that the machine will be shut down when using your rule.