STATCRUNCH

- Click on **Stat**.
- Choose **Regression » Simple Linear**.
- Choose X and Y variable names from the list of columns.
- Click on **Next** (twice) to **Plot the fitted line** on the scatterplot.
- Click on **Calculate** to see the regression analysis.
- Click on **Next** to see the scatterplot.

COMMENTS

Remember to check the scatterplot to be sure a linear model is appropriate.

Note that before you **Calculate**, clicking on **Next** also allows you to:

- enter an X-value for which you want to find the predicted Y-value;
- save all the fitted values;
- save the residuals;
- ask for a residuals plot.

TI-83/84 PLUS

To find the equation of the regression line (add the line to a scatterplot), choose **LinReg(a+bx)**, tell it the list names, and then add a comma to specify a function name (from **VARS Y-Vars 1:Function**). The final command looks like

LinReg(a+bx)L1, L2, Y1.

- To make a residuals plot, set up a **STATPLOT** as a scatterplot.
- Specify your explanatory data list as Xlist.
- For Ylist, import the name RESID from the **LIST NAMES** menu. **ZoomStat** will now create the residuals plot.

COMMENTS

Each time you execute a **LinReg** command, the calculator automatically computes the residuals and stores them in a data list named RESID. If you want to see them, go to **STAT EDIT**. Space through the names of the lists until you find a blank. Import RESID from the LIST NAMES menu. Now every time you have the calculator compute a regression analysis, it will show you the residuals.

Exercises

Section 7.1

1. **True or false** If false, explain briefly.
   a) We choose the linear model that passes through the most data points on the scatterplot.
   b) The residuals are the observed y-values minus the y-values predicted by the linear model.
   c) Least squares means that the square of the largest residual is as small as it could possibly be.

2. **True or false II** If false, explain briefly.
   a) Some of the residuals from a least squares linear model will be positive and some will be negative.
   b) Least Squares means that some of the squares of the residuals are minimized.
   c) We write $\hat{y}$ to denote the predicted values and y to denote the observed values.

Section 7.2

3. **Least squares interpretations** A least squares regression line was calculated to relate the length (cm) of newborn boys to their weight in kg. The line is $\widehat{weight} = -5.94 + 0.1875\ length$. Explain in words what this model means. Should new parents (who tend to worry) be concerned if their newborn's length and weight don't fit this equation?

4. **Residual interpretations** The newborn grandson of one of the authors was 48 cm long and weighed 3 kg. According to the regression model of Exercise 3, what was his residual? What does that say about him?

Section 7.3

5. **Bookstore sales revisited** Recall the data we saw in Chapter 6, Exercise 3 for a bookstore. The manager wants to predict *Sales* from *Number of Sales People Working*.

Number of Sales People Working	Sales (in $1000)
2	10
3	11
7	13
9	14
10	18
10	20
12	20
15	22
16	22
20	26
$\bar{x} = 10.4$	$\bar{y} = 17.6$
$SD(x) = 5.64$	$SD(y) = 5.34$
$r = 0.965$	

a) Find the slope estimate, b_1.
b) What does it mean, in this context?
c) Find the intercept, b_0.
d) What does it mean, in this context? Is it meaningful?
e) Write down the equation that predicts *Sales* from *Number of Sales People Working*.
f) If 18 people are working, what *Sales* do you predict?
g) If sales are actually $25,000, what is the value of the residual?
h) Have we overestimated or underestimated the sales?

6. **Disk drives again** In Chapter 6, Exercise 4, we saw some data on hard drives. After correcting for an outlier, these data look like this: we want to predict *Price* from *Capacity*.

Capacity (in TB)	Price (in $)
0.080	29.95
0.120	35.00
0.250	49.95
0.320	69.95
1.0	99.00
2.0	205.00
4.0	449.00
$\bar{x} = 1.110$	$\bar{y} = 133.98$
$SD(x) = 1.4469$	$SD(y) = 151.26$
$r = 0.994$	

a) Find the slope estimate, b_1.
b) What does it mean, in this context?
c) Find the intercept, b_0.
d) What does it mean, in this context? Is it meaningful?
e) Write down the equation that predicts *Price* from *Capacity*.
f) What would you predict for the price of a 3.0 TB drive?
g) You have found a 3.0 TB drive for $300. Is this a good buy? How much would you save compared to what you expected to pay?
h) Does the model overestimate or underestimate the price?

Section 7.4

7. **Sophomore slump?** A CEO complains that the winners of his "rookie junior executive of the year" award often turn out to have less impressive performance the following year. He wonders whether the award actually encourages them to slack off. Can you offer a better explanation?

8. **Sophomore slump again?** An online investment blogger advises investing in mutual funds that have performed badly the past year because "regression to the mean tells us that they will do well next year." Is he correct?

Section 7.5

9. **Bookstore sales once more** Here are the residuals for a regression of *Sales* on *Number of Sales People Working* for the bookstore Exercise 5:

Number of Sales People Working	Residual
2	0.07
3	0.16
7	−1.49
9	−2.32
10	0.77
10	2.77
12	0.94
15	0.20
16	−0.72
20	−0.37

a) What are the units of the residuals?
b) Which residual contributes the most to the sum that was minimized according to the Least Squares Criterion to find this regression?
c) Which residual contributes least to that sum?

10. **Disk drives once more** Here are the residuals for a regression of *Price* on *Capacity* for the hard drives of Exercise 6.

Capacity	Residual
0.080	3.02
0.120	3.91
0.250	5.35
0.320	18.075
1.0	−23.55
2.0	−21.475
4.0	14.666

a) Which residual contributes the most to the sum that is minimized by the Least Squares criterion?

b) Two of the residuals are negative. What does that mean about those drives? Be specific and use the correct units.

Section 7.6

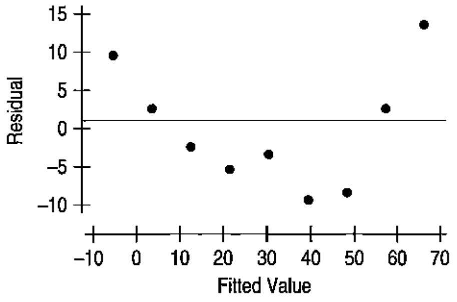
11. **Bookstore sales last time** For the regression model for the bookstore of Exercise 5, what is the value of R^2 and what does it mean?

12. **Disk drives encore** For the hard drive data of Exercise 6, find and interpret the value of R^2.

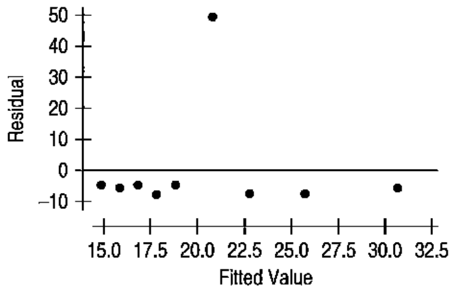
Section 7.7

13. **Residual plots** Here are residual plots (residuals plotted against predicted values) for three linear regression models. Indicate which condition appears to be violated (linearity, outlier or equal spread) in each case.

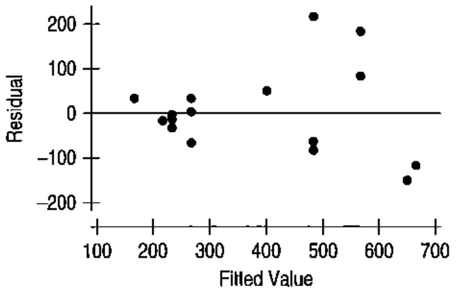
a)



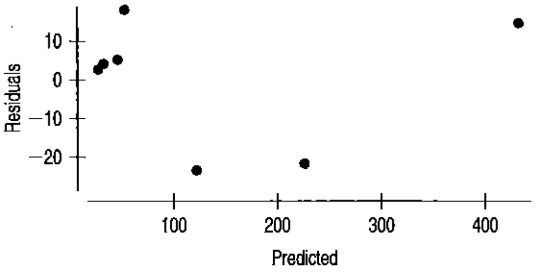
b)



c)



14. **Disk drives last time** Here is a scatterplot of the residuals from the regression of the hard drive prices on their sizes from Exercise 6.



a) Are any assumptions or conditions violated? If so, which ones?

b) What would you recommend about this regression?

Chapter Exercises

15. **Cereals** For many people, breakfast cereal is an important source of fiber in their diets. Cereals also contain potassium, a mineral shown to be associated with maintaining a healthy blood pressure. An analysis of the amount of fiber (in grams) and the potassium content (in milligrams) in servings of 77 breakfast cereals produced the regression model $\widehat{Potassium} = 38 + 27\,Fiber$. If your cereal provides 9 grams of fiber per serving, how much potassium does the model estimate you will get?

16. **Horsepower** In Chapter 6, Exercise 41, we examined the relationship between the fuel economy (mpg) and horsepower for 15 models of cars. Further analysis produces the regression model $\widehat{mpg} = 43.45 - 0.070\,HP$. If the car you are thinking of buying has a 200-horsepower engine, what does this model suggest your gas mileage would be?

17. **More cereal** Exercise 15 describes a regression model that estimates a cereal's potassium content from the amount of fiber it contains. In this context, what does it mean to say that a cereal has a negative residual?

18. **Horsepower again** Exercise 16 describes a regression model that uses a car's horsepower to estimate its fuel economy. In this context, what does it mean to say that a certain car has a positive residual?

19. **Another bowl** In Exercise 15, the regression model $\widehat{Potassium} = 38 + 27\,Fiber$ relates fiber (in grams) and potassium content (in milligrams) in servings of breakfast cereals. Explain what the slope means.

20. **More horsepower** In Exercise 16, the regression model $\widehat{mpg} = 43.45 - 0.070\,HP$ relates cars' horsepower to their fuel economy (in mpg). Explain what the slope means.

21. **Cereal again** The correlation between a cereal's fiber and potassium contents is $r = 0.903$. What fraction of the variability in potassium is accounted for by the amount of fiber that servings contain?

22. **Another car** The correlation between a car's horsepower and its fuel economy (in mpg) is $r = -0.909$. What fraction of the variability in fuel economy is accounted for by the horsepower?

23. **Last bowl!** For Exercise 15's regression model predicting potassium content (in milligrams) from the amount of fiber (in grams) in breakfast cereals, $s_e = 30.77$. Explain in this context what that means.

24. **Last tank!** For Exercise 16's regression model predicting fuel economy (in mpg) from the car's horsepower, $s_e = 2.435$. Explain in this context what that means.

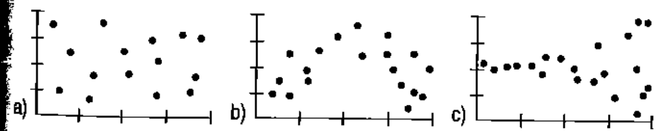
25. **Regression equations** Fill in the missing information in the following table.

	$\bar{x}$	s_x	$\bar{y}$	s_y	r	$\hat{y} = b_0 + b_1x$
a)	10	2	20	3	0.5	
b)	2	0.06	7.2	1.2	−0.4	
c)	12	6			−0.8	$\hat{y} = 200 - 4x$
d)	2.5	1.2		100		$\hat{y} = -100 + 50x$

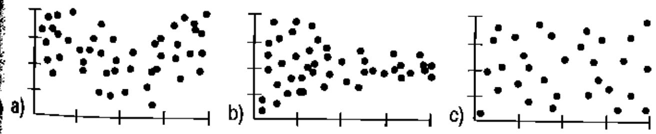
26. **More regression equations** Fill in the missing information in the following table.

	$\bar{x}$	s_x	$\bar{y}$	s_y	r	$\hat{y} = b_0 + b_1x$
a)	30	4	18	6	−0.2	
b)	100	18	60	10	0.9	
c)		0.8	50	15		$\hat{y} = -10 + 15x$
d)			18	4	−0.6	$\hat{y} = 30 - 2x$

27. **Residuals** Tell what each of the residual plots below indicates about the appropriateness of the linear model that was fit to the data.



28. **Residuals** Tell what each of the residual plots below indicates about the appropriateness of the linear model that was fit to the data.



29. **Real estate** A random sample of records of home sales from Feb. 15 to Apr. 30, 1993, from the files maintained by the Albuquerque Board of Realtors gives the *Price* and *Size* (in square feet) of 117 homes. A regression to predict *Price* (in thousands of dollars) from *Size* has an R^2 of 71.4%. The residuals plot indicated that a linear model is appropriate.

a) What are the variables and units in this regression?
b) What units does the slope have?
c) Do you think the slope is positive or negative? Explain.

30. **Roller coaster** The Mitch Hawker poll ranked the Top 10 steel roller coasters in 2011. A table in the previous chapter's exercises shows the length of the initial drop (in feet) and the duration of the ride (in seconds). A regression to predict *Duration* from *Drop* has $R^2 = 15.2\%$.

a) What are the variables and units in this regression?
b) What units does the slope have?
c) Do you think the slope is positive or negative? Explain.

31. **What slope?** If you create a regression model for predicting the *Weight* of a car (in pounds) from its *Length* (in feet), is the slope most likely to be 3, 30, 300, or 3000? Explain.

32. **What slope again?** If you create a regression model for estimating the *Height* of a pine tree (in feet) based on the *Circumference* of its trunk (in inches), is the slope most likely to be 0.1, 1, 10, or 100? Explain.

33. **Real estate again** The regression of *Price* on *Size* of homes in Albuquerque had $R^2 = 71.4\%$, as described in Exercise 29. Write a sentence (in context, of course) summarizing what the R^2 says about this regression.

34. **Coasters again** Exercise 30 examined the association between the *Duration* of a roller coaster ride and the height of its initial *Drop*, reporting that $R^2 = 15.2\%$. Write a sentence (in context, of course) summarizing what the R^2 says about this regression.

35. **Misinterpretations** A Biology student who created a regression model to use a bird's *Height* when perched for predicting its *Wingspan* made these two statements. Assuming the calculations were done correctly, explain what is wrong with each interpretation.

a) My R^2 of 93% shows that this linear model is appropriate.
b) A bird 10 inches tall will have a wingspan of 17 inches.

36. **More misinterpretations** A Sociology student investigated the association between a country's *Literacy Rate* and *Life Expectancy,* and then drew the conclusions listed below. Explain why each statement is incorrect. (Assume that all the calculations were done properly.)

a) The R^2 of 64% means that the *Literacy Rate* determines 64% of the *Life Expectancy* for a country.
b) The slope of the line shows that an increase of 5% in *Literacy Rate* will produce a 2-year improvement in *Life Expectancy.*

37. Real estate redux The regression of *Price* on *Size* of homes in Albuquerque had $R^2 = 71.4\%$, as described in Exercise 29.

a) What is the correlation between *Size* and *Price*?

b) What would you predict about the *Price* of a home 1 SD above average in *Size*?

c) What would you predict about the *Price* of a home 2 SDs below average in *Size*?

38. Another ride The regression of *Duration* of a roller coaster ride on the height of its initial *Drop*, described in Exercise 30, had $R^2 = 15.2\%$.

a) What is the correlation between *Drop* and *Duration*?

b) What would you predict about the *Duration* of the ride on a coaster whose initial *Drop* was 1 standard deviation below the mean *Drop*?

c) What would you predict about the *Duration* of the ride on a coaster whose initial *Drop* was 3 standard deviations above the mean *Drop*?

39. ESP People who claim to "have ESP" participate in a screening test in which they have to guess which of several images someone is thinking of. You and a friend both took the test. You scored 2 standard deviations above the mean, and your friend scored 1 standard deviation below the mean. The researchers offer everyone the opportunity to take a retest.

a) Should you choose to take this retest? Explain.

b) Now explain to your friend what his decision should be and why.

40. *SI* jinx Players in any sport who are having great seasons, turning in performances that are much better than anyone might have anticipated, often are pictured on the cover of *Sports Illustrated.* Frequently, their performances then falter somewhat, leading some athletes to believe in a "*Sports Illustrated* jinx." Similarly, it is common for phenomenal rookies to have less stellar second seasons—the so-called "sophomore slump." While fans, athletes, and analysts have proposed many theories about what leads to such declines, a statistician might offer a simpler (statistical) explanation. Explain.

41. More real estate Consider the Albuquerque home sales from Exercise 29 again. The regression analysis gives the model $\widehat{Price} = 47.82 + 0.061\,Size$.

a) Explain what the slope of the line says about housing prices and house size.

b) What price would you predict for a 3000-square-foot house in this market?

c) A real estate agent shows a potential buyer a 1200-square-foot home, saying that the asking price is $6000 less than what one would expect to pay for a house of this size. What is the asking price, and what is the $6000 called?

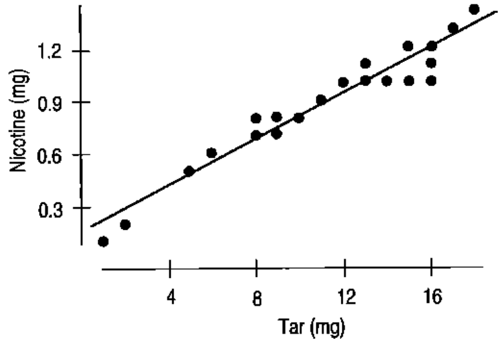
42. Last ride Consider the roller coasters described in Exercise 30 again. The regression analysis gives the model $\widehat{Duration} = 64.232 + 0.180\,Drop$.

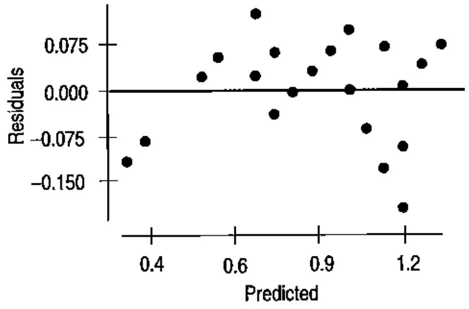
a) Explain what the slope of the line says about how long a roller coaster ride may last and the height of the coaster.

b) A new roller coaster advertises an initial drop of 200 feet. How long would you predict the rides last?

c) Another coaster with a 150-foot initial drop advertises a 2-minute ride. Is this longer or shorter than you'd expect? By how much? What's that called?

43. Cigarettes Is the nicotine content of a cigarette related to the "tar"? A collection of data (in milligrams) on 29 cigarettes produced the scatterplot, residuals plot, and regression analysis shown:





Dependent variable is Nicotine
R-squared = 92.4%

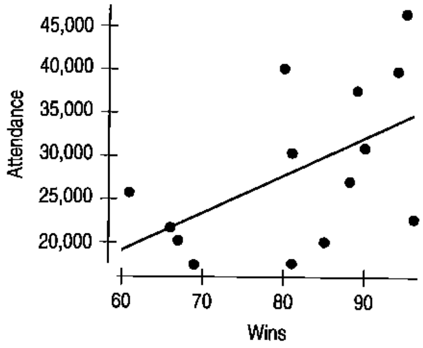
Variable	Coefficient
Constant	0.154030
Tar	0.065052

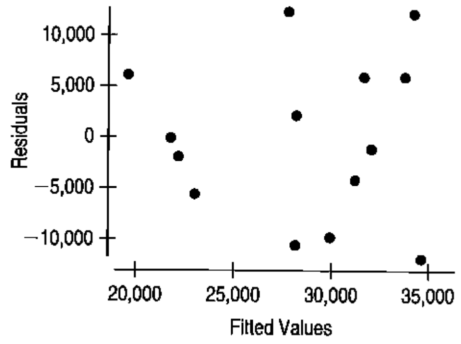
a) Do you think a linear model is appropriate here? Explain.

b) Explain the meaning of R^2 in this context.

44. Attendance 2010 In the previous chapter, you looked at the relationship between the number of wins by American League baseball teams and the average attendance at their home games for the 2010 season.

Here are the scatterplot, the residuals plot, and part of the regression analysis:





Dependent variable is Home Attendance
R-squared = 28.4%

Variable	Coefficient
Constant	−6760.5
Wins	431.22

a) Do you think a linear model is appropriate here? Explain.

b) Interpret the meaning of R^2 in this context.

c) Do the residuals show any pattern worth remarking on?

d) The point in the upper right of the plots is the New York Yankees. What can you say about the residual for the Yankees?

45. **Another cigarette** Consider again the regression of *Nicotine* content on *Tar* (both in milligrams) for the cigarettes examined in Exercise 43.

a) What is the correlation between *Tar* and *Nicotine*?

b) What would you predict about the average *Nicotine* content of cigarettes that are 2 standard deviations below average in *Tar* content?

c) If a cigarette is 1 standard deviation above average in *Nicotine* content, what do you suspect is true about its *Tar* content?

46. **Second inning 2010** Consider again the regression of *Average Attendance* on *Wins* for the baseball teams examined in Exercise 44.

a) What is the correlation between *Wins* and *Average Attendance*?

b) What would you predict about the *Average Attendance* for a team that is 2 standard deviations above average in *Wins*?

c) If a team is 1 standard deviation below average in attendance, what would you predict about the number of games the team has won?

47. **Last cigarette** Take another look at the regression analysis of tar and nicotine content of the cigarettes in Exercise 43.

a) Write the equation of the regression line.

b) Estimate the *Nicotine* content of cigarettes with 4 milligrams of *Tar*.

c) Interpret the meaning of the slope of the regression line in this context.

d) What does the y-intercept mean?

e) If a new brand of cigarette contains 7 milligrams of tar and a nicotine level whose residual is −0.5 mg, what is the nicotine content?

48. **Last inning 2010** Refer again to the regression analysis for average attendance and games won by American League baseball teams, seen in Exercise 44.

a) Write the equation of the regression line.

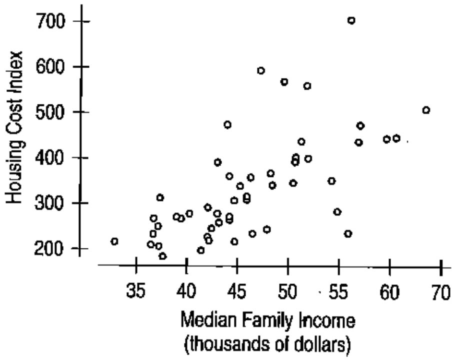
b) Estimate the *Average Attendance* for a team with 50 *Wins*.

c) Interpret the meaning of the slope of the regression line in this context.

d) In general, what would a negative residual mean in this context?

e) The San Francisco Giants, the 2010 World Champions, are not included in these data because they are a National League team. During the 2010 regular season, the Giants won 92 games and averaged 41,736 fans at their home games. Calculate the residual for this team, and explain what it means.

49. **Income and housing revisited** In Chapter 6, Exercise 39, we learned that the Office of Federal Housing Enterprise Oversight (OFHEO) collects data on various aspects of housing costs around the United States. Here's a scatterplot (by state) of the *Housing Cost Index* (HCI) versus the *Median Family Income* (MFI) for the 50 states. The correlation is $r = 0.65$. The mean HCI is 338.2, with a standard deviation of 116.55. The mean MFI is \$46,234, with a standard deviation of \$7072.47.



a) Is a regression analysis appropriate? Explain.

b) What is the equation that predicts Housing Cost Index from median family income?

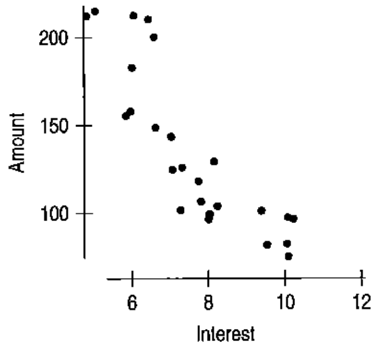
c) For a state with MFI = \$44,993, what would be the predicted HCI?

d) Washington, DC, has an MFI of \$44,993 and an HCI of 548.02. How far off is the prediction in part b from the actual HCI?

e) If we standardized both variables, what would be the regression equation that predicts standardized HCI from standardized MFI?

f) If we standardized both variables, what would be the regression equation that predicts standardized MFI from standardized HCI?

50. Interest rates and mortgages again In Chapter 6, Exercise 40, we saw a plot of mortgages in the United States (in thousands of dollars) versus the interest rate at various times over the past 26 years. The correlation is $r = -0.86$. The mean mortgage amount is \$121.8 thousand and the mean interest rate is 7.74%. The standard deviations are \$47.36 thousand for mortgage amounts and 1.79% for the interest rates.



a) Is a regression model appropriate for predicting mortgage amount from interest rates? Explain.

b) What is the equation that predicts mortgage amount from interest rates?

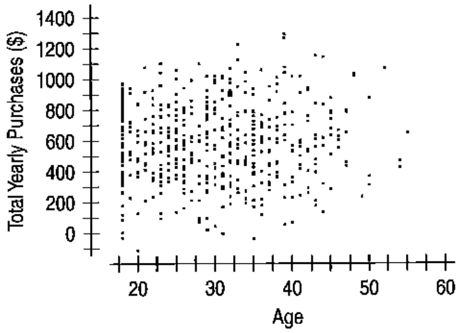
c) What would you predict the mortgage amount would be if the interest rates climbed to 13%?

d) Do you have any reservations about your prediction in part c?

e) If we standardized both variables, what would be the regression equation that predicts standardized mortgage amount from standardized interest rates?

f) If we standardized both variables, what would be the regression equation that predicts standardized interest rates from standardized mortgage amount?

51. Online clothes An online clothing retailer keeps track of its customers' purchases. For those customers who signed up for the company's credit card, the company also has information on the customer's *Age* and *Income.* A random sample of 500 of these customers shows the following scatterplot of *Total Yearly Purchases* by *Age:*



The correlation between *Total Yearly Purchases* and *Age* is $r = 0.037$. Summary statistics for the two variables are:

	Mean	SD
Age	29.67 yrs	8.51 yrs
Total Yearly Purchase	\$572.52	\$253.62

a) What is the linear regression equation for predicting *Total Yearly Purchase* from *Age*?

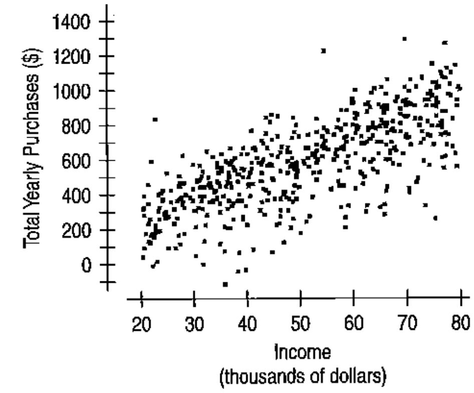
b) Do the assumptions and conditions for regression appear to be met?

c) What is the predicted *Total Yearly Purchase* for an 18-year-old? For a 50-year-old?

d) What percent of the variability in *Total Yearly Purchases* is accounted for by this model?

e) Do you think the regression might be a useful one for the company? Explain.

52. Online clothes II For the online clothing retailer discussed in the previous problem, the scatterplot of *Total Yearly Purchases* by *Income* looks like this:



The correlation between *Total Yearly Purchases* and *Income* is 0.722. Summary statistics for the two variables are:

	Mean	SD
Income	\$50,343.40	\$16,952.50
Total Yearly Purchase	\$572.52	\$253.62

a) What is the linear regression equation for predicting *Total Yearly Purchase* from *Income*?

b) Do the assumptions and conditions for regression appear to be met?

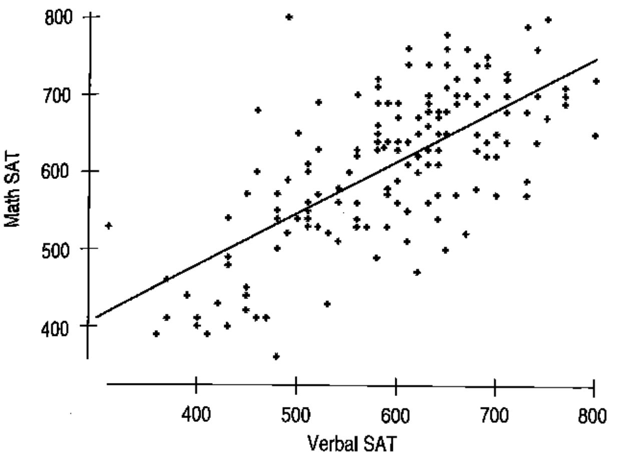
c) What is the predicted *Total Yearly Purchase* for someone with a yearly *Income* of \$20,000? For someone with an annual *Income* of \$80,000?

d) What percent of the variability in *Total Yearly Purchases* is accounted for by this model?

e) Do you think the regression might be a useful one for the company? Comment.

53. SAT scores The SAT is a test often used as part of an application to college. SAT scores are between 200 and 800, but have no units. Tests are given in both Math and

Verbal areas. SAT-Math problems require the ability to read and understand the questions, but can a person's verbal score be used to predict the math score? Verbal and math SAT scores of a high school graduating class are displayed in the scatterplot, with the regression line added.



a) Describe the relationship.

b) Are there any students whose scores do not seem to fit the overall pattern?

c) For these data, $r = 0.685$. Interpret this statistic.

d) These verbal scores averaged 596.3, with a standard deviation of 99.5, and the math scores averaged 612.2, with a standard deviation of 96.1. Write the equation of the regression line.

e) Interpret the slope of this line.

f) Predict the math score of a student with a verbal score of 500.

g) Every year, some students score a perfect 1600. Based on this model, what would such a student's residual be for her math score?

54. **Success in college** Colleges use SAT scores in the admissions process because they believe these scores provide some insight into how a high school student will perform at the college level. Suppose the entering freshmen at a certain college have mean combined *SAT Scores* of 1222, with a standard deviation of 123. In the first semester, these students attained a mean *GPA* of 2.66, with a standard deviation of 0.56. A scatterplot showed the association to be reasonably linear, and the correlation between *SAT* score and *GPA* was 0.47.

a) Write the equation of the regression line.

b) Explain what the y-intercept of the regression line indicates.

c) Interpret the slope of the regression line.

d) Predict the GPA of a freshman who scored a combined 1400.

e) Based upon these statistics, how effective do you think SAT scores would be in predicting academic success during the first semester of the freshman year at this college? Explain.

f) As a student, would you rather have a positive or a negative residual in this context? Explain.

55. **SAT, take 2** Suppose we wanted to use SAT math scores to estimate verbal scores based on the information in Exercise 53.

a) What is the correlation?

b) Write the equation of the line of regression predicting verbal scores from math scores.

c) In general, what would a positive residual mean in this context?

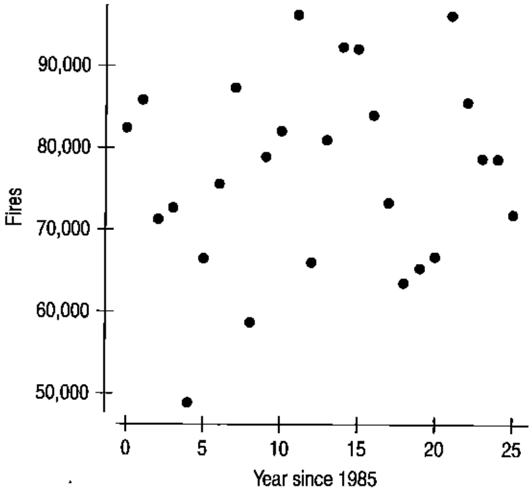
d) A person tells you her math score was 500. Predict her verbal score.

e) Using that predicted verbal score and the equation you created in Exercise 53, predict her math score.

f) Why doesn't the result in part e come out to 500?

56. **Success, part 2** Based on the statistics for college freshmen given in Exercise 54, what SAT score would you predict for a freshmen who attained a first-semester GPA of 3.0?

57. **Wildfires 2010** The National Interagency Fire Center (www.nifc.gov) reports statistics about wildfires. Here's an analysis of the number of wildfires between 1985 and 2010.



```
Dependent variable is Fires
R-squared = 1.9%
s = 11920
Variable             Coefficient
Intercept            74487.1
Years since 1985       209.728
```

a) Is a linear model appropriate for these data? Explain.

b) Interpret the slope in this context.

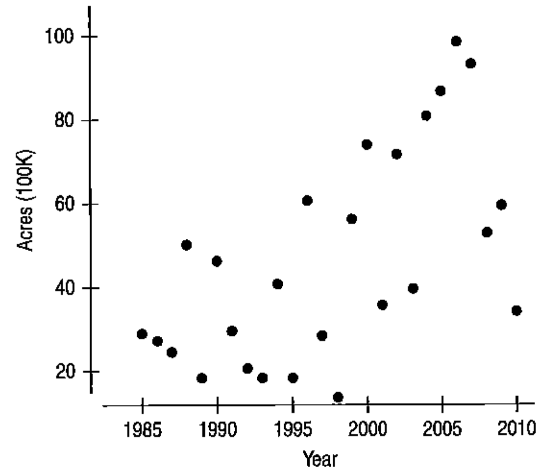
c) Can we interpret the intercept? Why or why not?

d) What does the value of s_e say about the size of the residuals? What does it say about the effectiveness of the model?

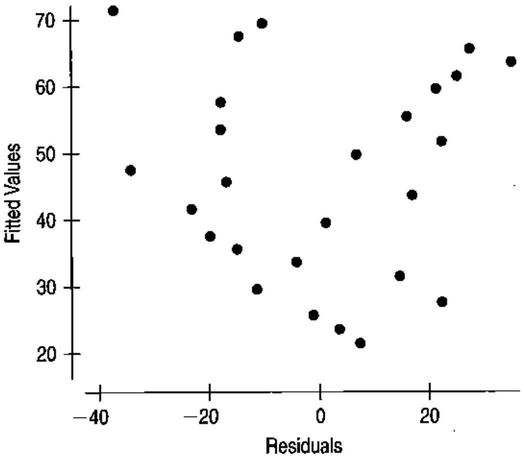
e) What does R^2 mean in this context?

58. **Wildfires 2010—sizes** We saw in Exercise 57 that the number of fires was nearly constant. But has the damage they cause remained constant as well? Here's a regression that examines the trend in *Acres per Fire,*

(in hundreds of thousands of acres) together with some supporting plots:



```
Dependent variable is Acres/fire
R-squared = 36.6%
s = 20.52
Variable              Coefficient
Intercept             −3941
Years since 1985          1.997
```



a) Is the regression model appropriate for these data? Explain.

b) What interpretation (if any) can you give for the R^2 in the regression table?

59. Used cars 2011 Carmax.com lists numerous Toyota Corollas for sale within a 250 mile radius of Redlands, CA. Listed at the top of the next column are the ages of the cars and the advertised prices.

a) Make a scatterplot for these data.

b) Describe the association between *Age* and *Price* of a used Corolla.

c) Do you think a linear model is appropriate?

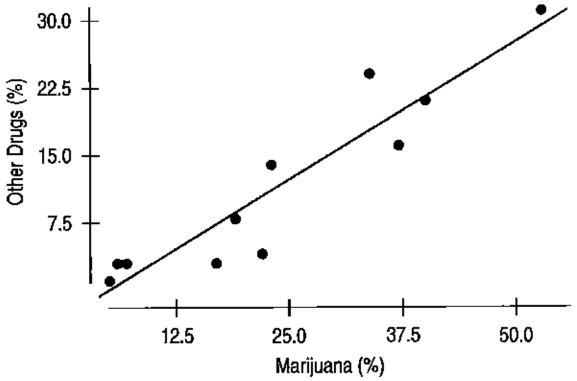
d) Computer software says that $R^2 = 89.1\%$. What is the correlation between *Age* and *Price*?

e) Explain the meaning of R^2 in this context.

f) Why doesn't this model explain 100% of the variability in the price of a used Corolla?

Age (yr)	Price Advertised ($)
1	17,599
2	14,998
2	15,998
4	13,998
4	14,998
5	14,599
5	13,998
6	11,998
7	9,998
7	11,559
8	10,849
8	10,899
10	9,998

60. Drug abuse In the exercises of the last chapter, you examined results of a survey conducted in the United States and 10 countries of Western Europe to determine the percentage of teenagers who had used marijuana and other drugs. Below is the scatterplot. Summary statistics showed that the mean percent that had used marijuana was 23.9%, with a standard deviation of 15.6%. An average of 11.6% of teens had used other drugs, with a standard deviation of 10.2%.



a) Do you think a linear model is appropriate? Explain.

b) For this regression, R^2 is 87.3%. Interpret this statistic in this context.

c) Write the equation you would use to estimate the percentage of teens who use other drugs from the percentage who have used marijuana.

d) Explain in context what the slope of this line means.

e) Do these results confirm that marijuana is a "gateway drug," that is, that marijuana use leads to the use of other drugs?

61. More used cars 2011 Use the advertised prices for Toyota Corollas given in Exercise 59 to create a linear model for the relationship between a car's *Age* and its *Price*.

a) Find the equation of the regression line.

b) Explain the meaning of the slope of the line.

c) Explain the meaning of the *y*-intercept of the line.

d) If you want to sell a 7-year-old Corolla, what price seems appropriate?

e) You have a chance to buy one of two cars. They are about the same age and appear to be in equally good condition. Would you rather buy the one with a positive residual or the one with a negative residual? Explain.

f) You see a "For Sale" sign on a 10-year-old Corolla stating the asking price as $8,500. What is the residual?

g) Would this regression model be useful in establishing a fair price for a 25-year-old car? Explain.

62. **Veggie burgers** Burger King introduced a meat-free burger in 2002. The nutrition label is shown here:

Nutrition Facts	
Calories	330
Fat	10g*
Sodium	760g
Sugars	5g
Protein	14g
Carbohydrates	43g
Dietary Fiber	4g
Cholesterol	0

* (2 grams of saturated fat)

RECOMMENDED DAILY VALUES
(based on a 2,000-calorie/day diet)

Iron	20%
Vitamin A	10%
Vitamin C	10%
Calcium	6%

a) Use the regression model created in this chapter, $\widehat{Fat} = 6.8 + 0.97\ Protein$, to predict the fat content of this burger from its protein content.

b) What is its residual? How would you explain the residual?

c) Write a brief report about the *Fat* and *Protein* content of this menu item. Be sure to talk about the variables by name and in the correct units.

63. **Burgers** In the last chapter, you examined the association between the amounts of *Fat* and *Calories* in fast-food hamburgers. Here are the data:

Fat (g)	19	31	34	35	39	39	43
Calories	410	580	590	570	640	680	660

a) Create a scatterplot of *Calories* vs. *Fat.*

b) Interpret the value of R^2 in this context.

c) Write the equation of the line of regression.

d) Use the residuals plot to explain whether your linear model is appropriate.

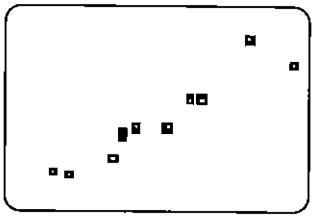
e) Explain the meaning of the y-intercept of the line.

f) Explain the meaning of the slope of the line.

g) A new burger containing 28 grams of fat is introduced. According to this model, its residual for calories is +33. How many calories does the burger have?

64. **Chicken** Chicken sandwiches are often advertised as a healthier alternative to beef because many are lower in fat. Tests on 11 brands of fast-food chicken sandwiches produced the following summary statistics and scatterplot from a graphing calculator:

	Fat (g)	Calories
Mean	20.6	472.7
St. Dev.	9.8	144.2
Correlation	0.947	



a) Do you think a linear model is appropriate in this situation?

b) Describe the strength of this association.

c) Write the equation of the regression line.

d) Explain the meaning of the slope.

e) Explain the meaning of the y–intercept.

f) What does it mean if a certain sandwich has a negative residual?

g) If a chicken sandwich and a burger each advertised 35 grams of fat, which would you expect to have more calories (see Exercise 63)?

h) McDonald's Filet-O-Fish sandwich has 26 grams of fat and 470 calories. Does the fat–calorie relationship in this sandwich appear to be very different from that found in chicken sandwiches or in burgers (see Exercise 63)? Explain.

65. **A second helping of burgers** In Exercise 63, you created a model that can estimate the number of *Calories* in a burger when the *Fat* content is known.

a) Explain why you cannot use that model to estimate the fat content of a burger with 600 calories.

b) Using an appropriate model, estimate the fat content of a burger with 600 calories.

66. **Cost of living 2008** The *Worldwide Cost of Living Survey City Rankings* determine the cost of living in the 25 most expensive cities in the world. (www.finfacts.com/costofliving.htm) These rankings scale New York City as 100, and express the cost of living in other cities as a percentage of the New York cost. For example, the table on the following page indicates that in Tokyo the cost of living was 22.1% higher than New York in 2007, and increased to 27.0% higher in 2008.

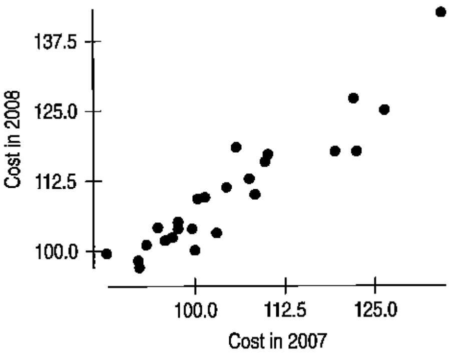
a) Using the scatterplot on the next page, describe the association between costs of living in 2007 and 2008.

b) The correlation is 0.938. Find and interpret the value of R^2.

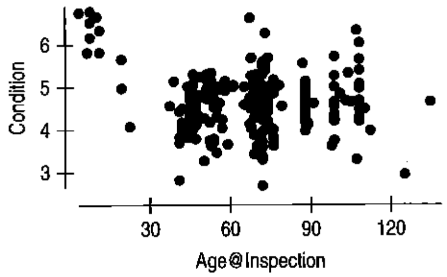
c) The regression equation predicting the 2008 cost of living from the 2007 figure is $\widehat{Cost08} = 21.75 + 0.84\ Cost07$. Use this equation to find the residual for Oslo.

d) Explain what the residual means.

City	2007	2008
Moscow	134.4	142.4
Tokyo	122.1	127.0
London	126.3	125.0
Oslo	105.8	118.3
Seoul	122.4	117.7
Hong Kong	119.4	117.6
Copenhagen	110.2	117.2
Geneva	109.8	115.8
Zurich	107.6	112.7
Milan	104.4	111.3
Osaka	108.4	110.0
Paris	101.4	109.4
Singapore	100.4	109.1
Tel Aviv	97.7	105.0
Sydney	94.9	104.1
Dublin	99.6	103.9
Rome	97.6	103.9
St. Petersburg	103.0	103.1
Vienna	96.9	102.3
Beijing	95.9	101.9
Helsinki	93.3	101.1
New York City	100.0	100.0
Istanbul	87.7	99.4
Shanghai	92.1	98.3
Amsterdam	92.2	97.0



67. New York bridges We saw in this chapter that in Tompkins County, New York, older bridges were in worse condition than newer ones. Tompkins is a rural area. Is this relationship true in New York City as well? Here are data on the *Condition* (as measured by the state Department of Transportation Condition Index) and *Age at Inspection* for bridges in New York City.



```
Dependent variable is Condition
R-squared = 2.6%
s = 0.6708

Variable            Coefficient
Intercept              4.95147
Age@Inspection        −0.00481
```

a) New York State defines any bridge with a condition score less than 5 as *deficient.* What does this model predict for the condition scores of New York City bridges?

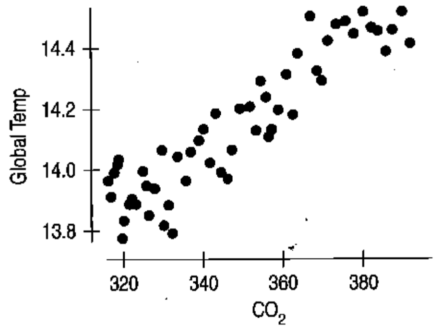
b) Our earlier model found that the condition of bridges in Tompkins County was decreasing at about 0.025 per year. What does this model say about New York City bridges?

c) How much faith would you place in this model? Explain.

68. Candy The table shows the increase in Halloween candy sales over a 6-year period as reported by the National Confectioners Association (www.candyusa.com). Using these data, estimate the amount of candy sold in 2009. Discuss the appropriateness of your model and your faith in the estimate. Then comment on the fact that NCA reported 2009 sales of $2.207 million. (Enter *Year* as 3, 4, ...)

2003	1.993
2004	2.041
2005	2.088
2006	2.146
2007	2.202
2008	2.209

69. Climate change 2011 The earth's climate is getting warmer. The most common theory attributes the increase to an increase in atmospheric levels of carbon dioxide (CO_2), a greenhouse gas. Here is a scatterplot showing the mean annual CO_2 concentration in the atmosphere, measured in parts per million (ppm) at the top of Mauna Loa in Hawaii, and the mean annual air temperature over both land and sea across the globe, in degrees Celsius (°C) for the years 1959 to 2011.

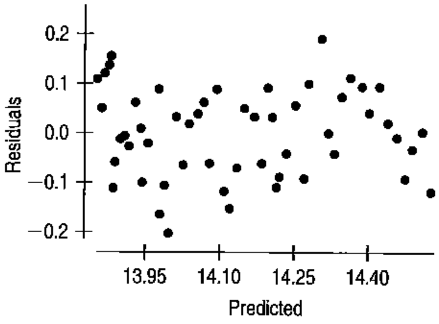


A regression predicting *Temperature* from CO_2 produces the following output table (in part):

```
Dependent variable is Global Temperature (°C)
R-squared = 84.0%

Variable            Coefficient
Intercept             11.0276
CO2 (ppm)              0.0089
```

a) What is the correlation between CO_2 and *Temperature*?
b) Explain the meaning of *R*-squared in this context.
c) Give the regression equation.
d) What is the meaning of the slope in this equation?
e) What is the meaning of the *y*-intercept of this equation?
f) Here is a scatterplot of the residuals vs. CO_2. Does this plot show evidence of the violation of any assumptions behind the regression? If so, which ones?



g) CO_2 levels will probably reach 400 ppm by 2020. What mean *Temperature* does the regression predict for that concentration of CO_2?

70. Birthrates 2009 The table shows the number of live births per 1000 women aged 15–44 years in the United States, starting in 1965. (National Center for Health Statistics, www.cdc.gov/nchs/)

Year	1965	1970	1975	1980	1985	1990	1995	2000	2005	2009
Rate	19.4	18.4	14.8	15.9	15.6	16.4	14.8	14.4	14.0	13.5

a) Make a scatterplot and describe the general trend in *Birthrates*. (Enter *Year* as years since 1900: 65, 70, 75, etc.)
b) Find the equation of the regression line.
c) Check to see if the line is an appropriate model. Explain.
d) Interpret the slope of the line.
e) The table gives rates only at 5-year intervals. Estimate what the rate was in 1978.
f) In 1978, the birthrate was actually 15.0. How close did your model come?
g) Predict what the *Birthrate* will be in 2010. Comment on your faith in this prediction.
h) Predict the *Birthrate* for 2025. Comment on your faith in this prediction.

71. Body fat It is difficult to determine a person's body fat percentage accurately without immersing him or her in water. Researchers hoping to find ways to make a good estimate immersed 20 male subjects, then measured their waists and recorded their weights shown in the table at the top of the next column.

a) Create a model to predict *%Body Fat* from *Weight*.
b) Do you think a linear model is appropriate? Explain.
c) Interpret the slope of your model.
d) Is your model likely to make reliable estimates? Explain.
e) What is the residual for a person who weighs 190 pounds and has 21% body fat?

Waist (in.)	Weight (lb)	Body Fat (%)	Waist (in.)	Weight (lb)	Body Fat (%)
32	175	6	33	188	10
36	181	21	40	240	20
38	200	15	36	175	22
33	159	6	32	168	9
39	196	22	44	246	38
40	192	31	33	160	10
41	205	32	41	215	27
35	173	21	34	159	12
38	187	25	34	146	10
38	188	30	44	219	28

72. Body fat again Would a model that uses the person's *Waist* size be able to predict the *%Body Fat* more accurately than one that uses *Weight*? Using the data in Exercise 71, create and analyze that model.

73. Heptathlon 2004 We discussed the women's 2008 Olympic heptathlon in Chapter 6. Here are the results from the high jump, 800-meter run, and long jump for the 26 women who successfully completed all three events in the 2004 Olympics (www.espn.com):

Name	Country	High Jump (m)	800-m (sec)	Long Jump (m)
Carolina Klüft	SWE	1.91	134.15	6.51
Austra Skujyte	LIT	1.76	135.92	6.30
Kelly Sotherton	GBR	1.85	132.27	6.51
Shelia Burrell	USA	1.70	135.32	6.25
Yelena Prokhorova	RUS	1.79	131.31	6.21
Sonja Kesselschlaeger	GER	1.76	135.21	6.42
Marie Collonville	FRA	1.85	133.62	6.19
Natalya Dobrynska	UKR	1.82	137.01	6.23
Margaret Simpson	GHA	1.79	137.72	6.02
Svetlana Sokolova	RUS	1.70	133.23	5.84
J. J. Shobha	IND	1.67	137.28	6.36
Claudia Tonn	GER	1.82	130.77	6.35
Naide Gomes	POR	1.85	140.05	6.10
Michelle Perry	USA	1.70	133.69	6.02
Aryiro Strataki	GRE	1.79	137.90	5.97
Karin Ruckstuhl	NED	1.85	133.95	5.90
Karin Ertl	GER	1.73	138.68	6.03
Kylie Wheeler	AUS	1.79	137.65	6.36
Janice Josephs	RSA	1.70	138.47	6.21
Tiffany Lott Hogan	USA	1.67	145.10	6.15
Magdalena Szczepanska	POL	1.76	133.08	5.98
Irina Naumenko	KAZ	1.79	134.57	6.16
Yuliya Akulenko	UKR	1.73	142.58	6.02
Soma Biswas	IND	1.70	132.27	5.92
Marsha Mark-Baird	TRI	1.70	141.21	6.22
Michaela Hejnova	CZE	1.70	145.68	5.70

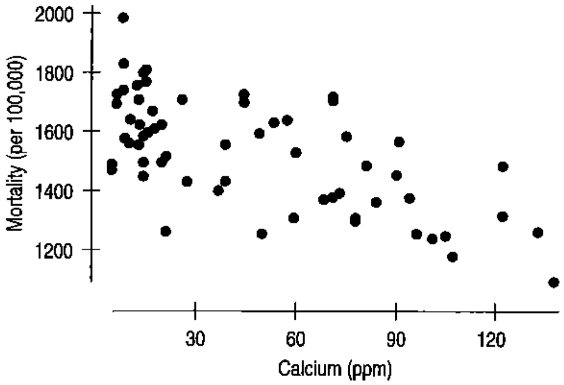
Let's examine the association among these events. Perform a regression to predict high-jump performance from the 800-meter results.

a) What is the regression equation? What does the slope mean?
b) What percent of the variability in high jumps can be accounted for by differences in 800-m times?
c) Do good high jumpers tend to be fast runners? (Be careful—low times are good for running events and high distances are good for jumps.)
d) What does the residuals plot reveal about the model?
e) Do you think this is a useful model? Would you use it to predict high-jump performance? (Compare the residual standard deviation to the standard deviation of the high jumps.)

74. Heptathlon 2004 again We saw the data for the women's 2004 Olympic heptathlon in Exercise 73. Are the two jumping events associated? Perform a regression of the long-jump results on the high-jump results.

a) What is the regression equation? What does the slope mean?
b) What percentage of the variability in long jumps can be accounted for by high-jump performances?
c) Do good high jumpers tend to be good long jumpers?
d) What does the residuals plot reveal about the model?
e) Do you think this is a useful model? Would you use it to predict long-jump performance? (Compare the residual standard deviation to the standard deviation of the long jumps.)

75. Hard water In an investigation of environmental causes of disease, data were collected on the annual mortality rate (deaths per 100,000) for males in 61 large towns in England and Wales. In addition, the water hardness was recorded as the calcium concentration (parts per million, ppm) in the drinking water. The following display shows the relationship between *Mortality* and *Calcium* concentration for these towns:



a) Describe what you see in this scatterplot, in context.
b) Here is the regression analysis of *Mortality* and *Calcium* concentration. What is the regression equation?

Dependent variable is Mortality
R-squared = 43%
s = 143.0

Variable	Coefficient	SE(Coeff)	t-Ratio	P-Value
Intercept	1676	29.30	57.2	<0.0001
Calcium	−3.23	0.48	26.66	<0.0001

c) Interpret the slope and y-intercept of the line, in context.
d) The largest residual, with a value of −348.6, is for the town of Exeter. Explain what this value means.
e) The hardness of Derby's municipal water is about 100 ppm of calcium. Use this equation to predict the mortality rate in Derby.
f) Explain the meaning of R-squared in this situation.

76. Gators Wildlife researchers monitor many wildlife populations by taking aerial photographs. Can they estimate the weights of alligators accurately from the air? Here is a regression analysis of the *Weight* of alligators (in pounds) and their *Length* (in inches) based on data collected about captured alligators.

Dependent variable is Weight
R-squared = 83.6%
s = 54.01

Variable	Coefficient	SE(Coeff)	t-Ratio	P-Value
Intercept	−393	47.53	−8.27	<0.0001
Length	5.9	0.5448	10.8	<0.0001

a) Did they choose the correct variable to use as the dependent variable and the predictor? Explain.
b) What is the correlation between an alligator's length and weight?
c) Write the regression equation.
d) Interpret the slope of the equation in this context.
e) Do you think this equation will allow the scientists to make accurate predictions about alligators? What part of the regression analysis indicates this? What additional concerns do you have?

77. Least squares Consider the four points (10,10), (20,50), (40,20), and (50,80). The least squares line is $\hat{y} = 7.0 + 1.1x$. Explain what "least squares" means, using these data as a specific example.

78. Least squares Consider the four points (200,1950), (400,1650), (600,1800), and (800,1600). The least squares line is $\hat{y} = 1975 - 0.45x$. Explain what "least squares" means, using these data as a specific example.

Review of Part II

Exploring Relationships Between Variables

Quick Review

You have now survived your second major unit of Statistics. Here's a brief summary of the key concepts and skills:

- We treat data two ways: as categorical and as quantitative.
- To explore relationships in categorical data, check out Chapter 2.
- To explore relationships in quantitative data:
  - Make a picture. Use a scatterplot. Put the explanatory variable on the *x*-axis and the response variable on the *y*-axis.
  - Describe the association between two quantitative variables in terms of direction, form, and strength.
  - The amount of scatter determines the strength of the association.
  - If, as one variable increases so does the other, the association is positive. If one increases as the other decreases, it's negative.
  - If the form of the association is linear, calculate a correlation to measure its strength numerically, and do a regression analysis to model it.
  - Correlations closer to -1 or $+1$ indicate stronger linear associations. Correlations near 0 indicate weak linear relationships, but other forms of association may still be present.
  - The line of best fit is also called the least squares regression line because it minimizes the sum of the squared residuals.
  - The regression line predicts values of the response variable from values of the explanatory variable.
  - A residual is the difference between the true value of the response variable and the value predicted by the regression model.
  - The slope of the line is a rate of change, best described in "*y*-units" per "*x*-unit."
  - R^2 gives the fraction of the variation in the response variable that is accounted for by the model.
  - The standard deviation of the residuals measures the amount of scatter around the line.
  - Outliers and influential points can distort any of our models.
  - If you see a pattern (a curve) in the residuals plot, your chosen model is not appropriate; use a different model. You may, for example, straighten the relationship by re-expressing one of the variables.
  - To straighten bent relationships, re-express the data using logarithms or a power (squares, square roots, reciprocals, etc.).
  - Always remember that an association is not necessarily an indication that one of the variables causes the other.

Need more help with some of this? Try rereading some sections of Chapters 6 through 8. And see below for more opportunities to review these concepts and skills.

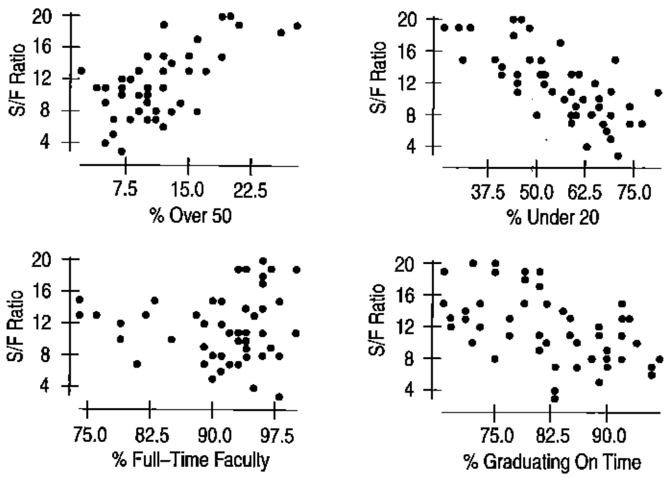
"One must learn by doing the thing; though you think you know it, you have no certainty until you try."

—*Sophocles (495–406 B.C.E.)*

Review Exercises

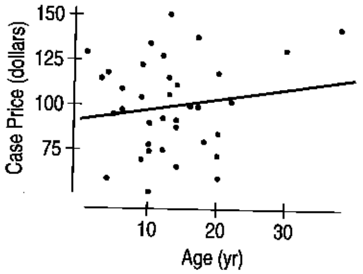
1. College Every year, *US News and World Report* publishes a special issue on many U.S. colleges and universities. The scatterplots below have *Student/Faculty Ratio* (number of students per faculty member) for the colleges and universities on the *y*-axes plotted against 4 other variables. The correct correlations for these scatterplots appear in this list. Match them.

-0.98 -0.71 -0.51 0.09 0.23 0.69



2. **Togetherness** Are good grades in high school associated with family togetherness? A random sample of 142 high school students was asked how many meals per week their families ate together. Their responses produced a mean of 3.78 meals per week, with a standard deviation of 2.2. Researchers then matched these responses against the students' grade point averages (GPAs). The scatterplot appeared to be reasonably linear, so they created a line of regression. No apparent pattern emerged in the residuals plot. The equation of the line was $\widehat{GPA} = 2.73 + 0.11\,Meals$.
   a) Interpret the y-intercept in this context.
   b) Interpret the slope in this context.
   c) What was the mean GPA for these students?
   d) If a student in this study had a negative residual, what did that mean?
   e) Upon hearing of this study, a counselor recommended that parents who want to improve the grades their children get should get the family to eat together more often. Do you agree with this interpretation? Explain.

3. **Vineyards** Here are the scatterplot and regression analysis for *Case Prices* of 36 wines from vineyards in the Finger Lakes region of New York State and the *Ages* of the vineyards.

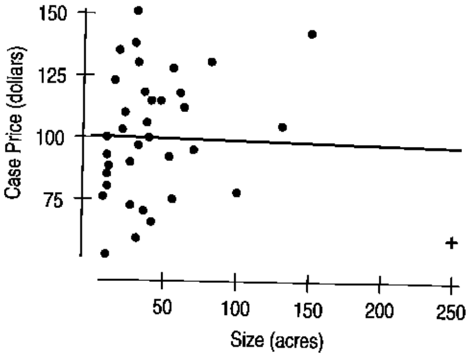


Dependent variable is Case Price
R-squared = 2.7%

Variable	Coefficient
Constant	92.7650
Age	0.567284

   a) Does it appear that vineyards in business longer get higher prices for their wines? Explain.
   b) What does this analysis tell us about vineyards in the rest of the world?
   c) Write the regression equation.
   d) Explain why that equation is essentially useless.

4. **Vineyards again** Instead of *Age*, perhaps the *Size* of the vineyard (in acres) is associated with the price of the wines. Look at the scatterplot:



   a) Do you see any evidence of an association?
   b) What concern do you have about this scatterplot?
   c) If the red "+" data point is removed, would the correlation become stronger or weaker? Explain.
   d) If the red "+" data point is removed, would the slope of the line increase or decrease? Explain.

5. **More twins 2009?** As the table shows, the number of twins born in the United States has been increasing. (www.cdc.gov/nchs/births.htm)

Year	Twin Births	Year	Twin Births
1980	68,339	1995	96,736
1981	70,049	1996	100,750
1982	71,631	1997	104,137
1983	72,287	1998	110,670
1984	72,949	1999	114,307
1985	77,102	2000	118,916
1986	79,485	2001	121,246
1987	81,778	2002	125,134
1988	85,315	2003	128,665
1989	90,118	2004	132,219
1990	93,865	2005	133,122
1991	94,779	2006	137,085
1992	95,372	2007	138,961
1993	96,445	2008	138,660
1994	97,064	2009	137,217

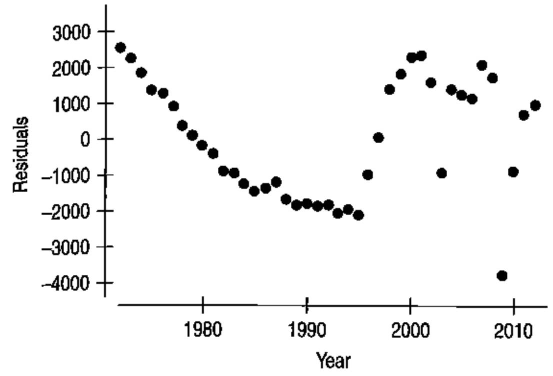
   a) Find the equation of the regression line for predicting the number of twin births.
   b) Explain in this context what the slope means.
   c) Predict the number of twin births in the United States for the year 2014. Comment on your faith in that prediction.
   d) Comment on the residuals plot.

6. **Dow Jones 2012** The Dow Jones stock index measures the performance of the stocks of America's largest companies (finance.yahoo.com). A regression of the Dow prices on years 1972–2012 looks like this:

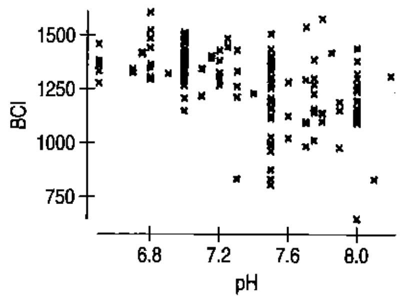
Dependent variable is Dow Index
R-squared = 83.9% s = 1659

Variable	Coefficient
Intercept	−667396
Year	337.605

   a) What is the correlation between *Dow Index* and *Year*?
   b) Write the regression equation.
   c) Explain in this context what the equation says.
   d) Here's a scatterplot of the residuals. Which assumption(s) of the regression analysis appear to be violated?



7. **Acid rain** Biologists studying the effects of acid rain on wildlife collected data from 163 streams in the Adirondack Mountains. They recorded the *pH* (acidity) of the water and the *BCI*, a measure of biological diversity, and they calculated $R^2 = 27\%$. Here's a scatterplot of *BCI* against *pH*:



a) What is the correlation between *pH* and *BCI*?
b) Describe the association between these two variables.
c) If a stream has average *pH*, what would you predict about the *BCI*?
d) In a stream where the *pH* is 3 standard deviations above average, what would you predict about the *BCI*?

8. **Manatees 2010** Marine biologists warn that the growing number of powerboats registered in Florida threatens the existence of manatees. The data below come from the Florida Fish and Wildlife Conservation Commission (myfwc.com/research/manatee/) and the National Marine Manufacturers Association (www.nmma.org/).

Year	Manatees Killed	Powerboat Registrations (in 1000s)
1982	13	447
1983	21	460
1984	24	481
1985	16	498
1986	24	513
1987	20	512
1988	15	527
1989	34	559
1990	33	585
1992	33	614
1993	39	646
1994	43	675
1995	50	711
1996	47	719
1997	53	716
1998	38	716
1999	35	716
2000	49	735
2001	81	860
2002	95	923
2003	73	940
2004	69	946
2005	79	974
2006	92	988
2007	73	992
2008	90	932
2009	97	949
2010	83	914

a) In this context, which is the explanatory variable?
b) Make a scatterplot of these data and describe the association you see.
c) Find the correlation between *Boat Registrations* and *Manatee Deaths*.
d) Interpret the value of R^2.
e) Does your analysis prove that powerboats are killing manatees?

9. **A manatee model 2010** Continue your analysis of the manatee situation from the previous exercise.

a) Create a linear model of the association between *Manatee Deaths* and *Powerboat Registrations*.
b) Interpret the slope of your model.
c) Interpret the y-intercept of your model.
d) How accurately did your model predict the high number of manatee deaths in 2010?
e) Which is better for the manatees, positive residuals or negative residuals? Explain.
f) What does your model suggest about the future for the manatee?

10. **Grades** A Statistics instructor created a linear regression equation to predict students' final exam scores from their midterm exam scores. The regression equation was $\widehat{Fin} = 10 + 0.9\,Mid$.

a) If Susan scored a 70 on the midterm, what did the instructor predict for her score on the final?
b) Susan got an 80 on the final. How big is her residual?
c) If the standard deviation of the final was 12 points and the standard deviation of the midterm was 10 points, what is the correlation between the two tests?
d) How many points would someone need to score on the midterm to have a predicted final score of 100?

e) Suppose someone scored 100 on the final. Explain why you can't estimate this student's midterm score from the information given.

f) One of the students in the class scored 100 on the midterm but got overconfident, slacked off, and scored only 15 on the final exam. What is the residual for this student?

g) No other student in the class "achieved" such a dramatic turnaround. If the instructor decides not to include this student's scores when constructing a new regression model, will the R^2 value of the regression increase, decrease, or remain the same? Explain.

h) Will the slope of the new line increase or decrease?

11. Traffic Highway planners investigated the relationship between traffic *Density* (number of automobiles per mile) and the average *Speed* of the traffic on a moderately large city thoroughfare. The data were collected at the same location at 10 different times over a span of 3 months. They found a mean traffic *Density* of 68.6 cars per mile (cpm) with standard deviation of 27.07 cpm. Overall, the cars' average *Speed* was 26.38 mph, with standard deviation of 9.68 mph. These researchers found the regression line for these data to be $\widehat{Speed} = 50.55 - 0.352\,Density$.

a) What is the value of the correlation coefficient between *Speed* and *Density*?

b) What percent of the variation in average *Speed* is explained by traffic *Density*?

c) Predict the average *Speed* of traffic on the thoroughfare when the traffic *Density* is 50 cpm.

d) What is the value of the residual for a traffic *Density* of 56 cpm with an observed *Speed* of 32.5 mph?

e) The data set initially included the point *Density* $=$ 125 cpm, *Speed* $=$ 55 mph. This point was considered an outlier and was not included in the analysis. Will the slope increase, decrease, or remain the same if we redo the analysis and include this point?

f) Will the correlation become stronger, weaker, or remain the same if we redo the analysis and include this point $(125, 55)$?

g) A European member of the research team measured the *Speed* of the cars in kilometers per hour $(1 \text{ km} \approx 0.62 \text{ miles})$ and the traffic *Density* in cars per kilometer. Find the value of his calculated correlation between speed and density.

12. Cramming One Thursday, researchers gave students enrolled in a section of basic Spanish a set of 50 new vocabulary words to memorize. On Friday, the students took a vocabulary test. When they returned to class the following Monday, they were retested—without advance warning. Here are the test scores for the 25 students.

Fri.	Mon.	Fri.	Mon.	Fri.	Mon.
42	36	48	37	39	41
44	44	43	41	46	32
45	46	45	32	37	36
48	38	47	44	40	31
44	40	50	47	41	32
43	38	34	34	48	39
41	37	38	31	37	31
35	31	43	40	36	41
43	32				

a) What is the correlation between *Friday* and *Monday* scores?

b) What does a scatterplot show about the association between the scores?

c) What does it mean for a student to have a positive residual?

d) What would you predict about a student whose *Friday* score was one standard deviation below average?

e) Write the equation of the regression line.

f) Predict the *Monday* score of a student who earned a 40 on Friday.

13. Car correlations What factor most explains differences in *Fuel Efficiency* among cars? Below is a correlation matrix exploring that relationship for the car's *Weight, Horsepower*, engine *size* (*Displacement*), and number of *Cylinders*.

	MPG	Weight	Horse- power	Displacement	Cylinders
MPG	1.000				
Weight	−0.903	1.000			
Horsepower	−0.871	0.917	1.000		
Displacement	−0.786	0.951	0.872	1.000	
Cylinders	−0.806	0.917	0.864	0.940	1.000

a) Which factor seems most strongly associated with *Fuel Efficiency*?

b) What does the negative correlation indicate?

c) Explain the meaning of R^2 for that relationship.

14. Autos, revisited Look again at the correlation table for cars in the previous exercise.

a) Which two variables in the table exhibit the strongest association?

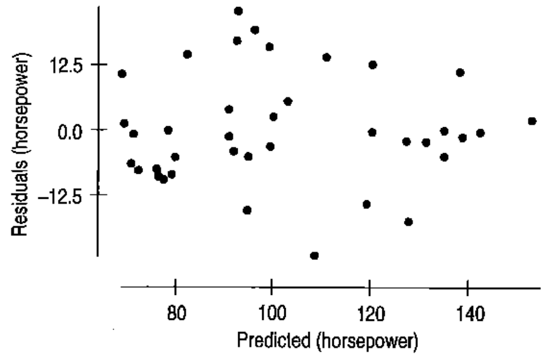
b) Is that strong association necessarily cause-and-effect? Offer at least two explanations why that association might be so strong.

c) Engine displacements for U.S.-made cars are often measured in cubic inches. For many foreign cars, the units are either cubic centimeters or liters. How would changing from cubic inches to liters affect the calculated correlations involving *Displacement*?

d) What would you predict about the *Fuel Efficiency* of a car whose engine *Displacement* is one standard deviation above the mean?

Cars, one more time! Can we predict the *Horsepower* of the engine that manufacturers will put in a car by knowing the *Weight* of the car? Here are the regression analysis and residuals plot:

```
Dependent variable is Horsepower
R-squared = 84.1%
Variable      Coefficient
Intercept       3.49834
Weight         34.3144
```

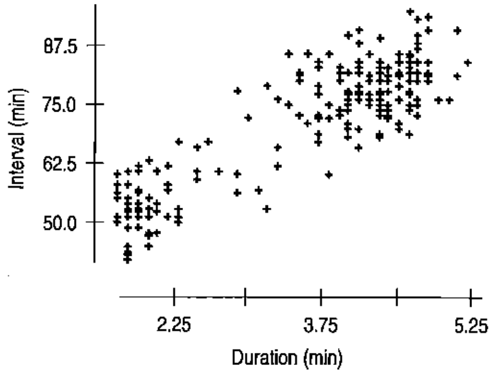


a) Write the equation of the regression line.
b) Do you think the car's *Weight* is measured in pounds or thousands of pounds? Explain.
c) Do you think this linear model is appropriate? Explain.
d) The highest point in the residuals plot, representing a residual of 22.5 horsepower, is for a Chevy weighing 2595 pounds. How much horsepower does this car have?

16. **Colorblind** Although some women are colorblind, this condition is found primarily in men. Why is it wrong to say there's a strong correlation between *Sex* and *Colorblindness*?

17. **Old Faithful** There is evidence that eruptions of Old Faithful can best be predicted by knowing the duration of the previous eruption.

a) Describe what you see in the scatterplot of *Intervals* between eruptions vs. *Duration* of the previous eruption.



b) Write the equation of the line of best fit. Here's the regression analysis:

```
Dependent variable is Interval
R-squared = 77.0%    s = 6.16 min
Variable     Coefficient
Intercept     33.9668
Duration      10.3582
```

c) Carefully explain what the slope of the line means in this context.
d) How accurate do you expect predictions based on this model to be? Cite statistical evidence.
e) If you just witnessed an eruption that lasted 4 minutes, how long do you predict you'll have to wait to see the next eruption?
f) So you waited, and the next eruption came in 79 minutes. Use this as an example to define a residual.

18. **Crocodile lengths** The ranges inhabited by the Indian gharial crocodile and the Australian saltwater crocodile overlap in Bangladesh. Suppose a very large crocodile skeleton is found there, and we wish to determine the species of the animal. Wildlife scientists have measured the lengths of the heads and the complete bodies of several crocs (in centimeters) of each species, creating the regression analyses below:

```
Indian Crocodile                      Australian Crocodile
Dependent variable is IBody           Dependent variable is ABody
R-squared = 97.2%                     R-squared = 98.0%
Variable      Coefficient             Variable      Coefficient
Intercept     −69.3693                Intercept     −20.2245
IHead           7.40004               AHead           7.71726
```

a) Do the associations between the sizes of the heads and bodies of the two species appear to be strong? Explain.
b) In what ways are the two relationships similar? Explain.
c) What is different about the two models? What does that mean?
d) The crocodile skeleton found had a head length of 62 cm and a body length of 380 cm. Which species do you think it was? Explain why.

19. **How old is that tree?** One can determine how old a tree is by counting its rings, but that requires either cutting the tree down or extracting a sample from the tree's core. Can we estimate the tree's age simply from its diameter? A forester measured 27 trees of the same species that had been cut down, and counted the rings to determine the ages of the trees.

Diameter (in.)	Age (yr)	Diameter (in.)	Age (yr)
1.8	4	10.3	23
1.8	5	14.3	25
2.2	8	13.2	28
4.4	8	9.9	29
6.6	8	13.2	30
4.4	10	15.4	30
7.7	10	17.6	33
10.8	12	14.3	34
7.7	13	15.4	35
5.5	14	11.0	38
9.9	16	15.4	38
10.1	18	16.5	40
12.1	20	16.5	42
12.8	22		

a) Find the correlation between *Diameter* and *Age*. Does this suggest that a linear model may be appropriate? Explain.
b) Create a scatterplot and describe the association.
c) Create the linear model.
d) Check the residuals. Explain why a linear model is probably not appropriate.
e) If you used this model, would it generally overestimate or underestimate the ages of very large trees? Explain.

20. Improving trees In the last exercise, you saw that the linear model had some deficiencies. Let's create a better model.

a) Perhaps the cross-sectional area of a tree would be a better predictor of its age. Since area is measured in square units, try re-expressing the data by squaring the diameters. Does the scatterplot look better?
b) Create a model that predicts *Age* from the square of the *Diameter*.
c) Check the residuals plot for this new model. Is this model more appropriate? Why?
d) Estimate the age of a tree 18 inches in diameter.

21. Big screen An electronics website collects data on the size of new HD flat panel televisions (measuring the diagonal of the screen in inches) to predict the cost (in hundreds of dollars). Which of these is most likely to be the slope of the regression line: 0.03, 0.3, 3, 30? Explain.

22. Smoking and pregnancy 2006 The Child Trends Data Bank monitors issues related to children. The table shows a 50-state average of the percent of expectant mothers who smoked cigarettes during their pregnancies.

Year	% Smoking While Pregnant	Year	% Smoking While Pregnant
1990	19.2	1999	14.1
1991	18.7	2000	14.0
1992	17.9	2001	13.8
1993	16.8	2002	13.3
1994	16.0	2003	12.7
1995	15.4	2004	10.9
1996	15.3	2005	10.1
1997	14.9	2006	10.0
1998	14.8		

a) Create a scatterplot and describe the trend you see.
b) Find the correlation.
c) How is the value of the correlation affected by the fact that the data are averages rather than percentages for each of the 50 states?
d) Write a linear model and interpret the slope in context.

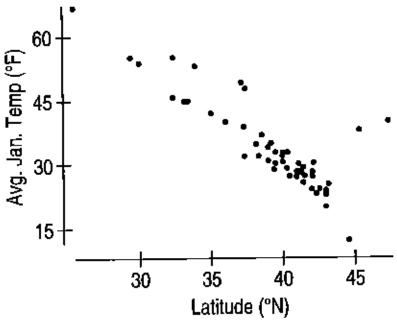
23. No smoking? The downward trend in smoking you saw in the last exercise is good news for the health of babies, but will it ever stop?

a) Explain why you can't use the linear model you created in Exercise 22 to see when smoking during pregnancy will cease altogether.
b) Create a model that could estimate the year in which the level of smoking would be 0%.
c) Comment on the reliability of such a prediction.

24. Tips It's commonly believed that people use tips to reward good service. A researcher for the hospitality industry examined tips and ratings of service quality from 2645 dining parties at 21 different restaurants. The correlation between ratings of service and tip percentages was 0.11. (Source: M. Lynn and M. McCall, "Gratitude and Gratuity." *Journal of Socio-Economics* 29: 203–214)

a) Describe the relationship between *Quality of Service* and *Tip Size*.
b) Find and interpret the value of R^2 in this context.

25. U.S. cities Data from 50 large U.S. cities show the mean *January Temperature* and the *Latitude*. Describe what you see in the scatterplot.



26. Correlations The study of U.S. cities in Exercise 25 found the mean *January Temperature* (degrees Fahrenheit), *Altitude* (feet above sea level), and *Latitude* (degrees north of the equator) for 55 cities. Here's the correlation matrix:

	Jan. Temp	Latitude	Altitude
Jan. Temp	1.000		
Latitude	−0.848	1.000	
Altitude	−0.369	0.184	1.000

a) Which seems to be more useful in predicting *January Temperature: Altitude* or *Latitude*? Explain.
b) If the *Temperature* were measured in degrees Celsius, what would be the correlation between Temperature and Latitude?
c) If the *Temperature* were measured in degrees Celsius and the *Altitude* in meters, what would be the correlation? Explain.
d) What would you predict about the January *Temperature* in a city whose *Altitude* is two standard deviations higher than the average *Altitude*?

27. **Winter in the city** Summary statistics for the data relating the latitude and average January temperature for 55 large U.S. cities are given below.

Variable	Mean	StdDev
Latitude	39.02	5.42
JanTemp	26.44	13.49

Correlation = −0.848

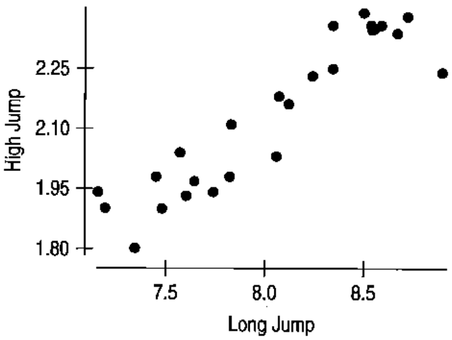
a) What percent of the variation in January *Temperature* can be explained by variation in *Latitude*?
b) What is indicated by the fact that the correlation is negative?
c) Write the equation of the line of regression for predicting average January *Temperature* from *Latitude*.
d) Explain what the slope of the line means.
e) Do you think the *y*-intercept is meaningful? Explain.
f) The latitude of Denver is 40°N. Predict the mean January temperature there.
g) What does it mean if the residual for a city is positive?

28. **Depression** The September 1998 issue of the *American Psychologist* published an article by Kraut et al. that reported on an experiment examining "the social and psychological impact of the Internet on 169 people in 73 households during their first 1 to 2 years online." In the experiment, 73 households were offered free Internet access for 1 or 2 years in return for allowing their time and activity online to be tracked. The members of the households who participated in the study were also given a battery of tests at the beginning and again at the end of the study. The conclusion of the study made news headlines: Those who spent more time online tended to be more depressed at the end of the experiment. Although the paper reports a more complex model, the basic result can be summarized in the following regression of *Depression* (at the end of the study, in "depression scale units") vs. *Internet Use* (in mean hours per week):

```
Dependent variable is Depression
R-squared = 4.6%   s = 0.4563
Variable        Coefficient
Intercept         0.5655
Internet use      0.0199
```

The news reports about this study clearly concluded that using the Internet causes depression. Discuss whether such a conclusion can be drawn from this regression. If so, discuss the supporting evidence. If not, say why not.

29. **Jumps 2008** How are Olympic performances in various events related? The plot shows winning long-jump and high-jump distances, in meters, for the Summer Olympics from 1912 through 2008.



a) Describe the association.
b) Do long-jump performances somehow influence the high-jumpers? How do you account for the relationship you see?
c) The correlation for the given scatterplot is 0.920. If we converted the jump lengths to centimeters by multiplying by 100, would that make the actual correlation higher or lower?
d) What would you predict about the long jump in a year when the high-jumper jumped one standard deviation better than the average high jump?

30. **Modeling jumps 2008** Here are the summary statistics for the Olympic long jumps and high jumps displayed in the previous exercise.

Event	Mean	StdDev
High Jump	2.13880	0.191884
Long Jump	8.03960	0.521380

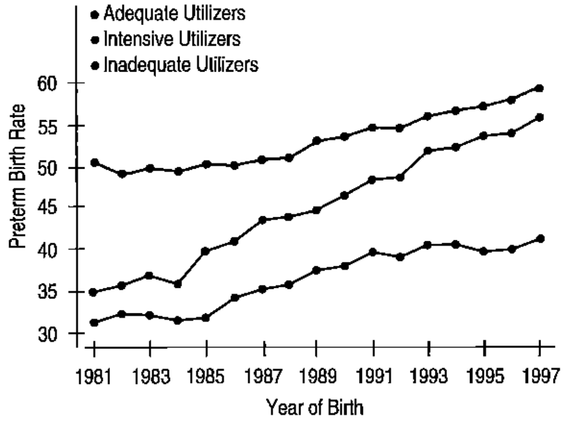
Correlation = 0.920

a) Write the equation of the line of regression for estimating *High Jump* from *Long Jump*.
b) Interpret the slope of the line.
c) In a year when the long jump is 8.9 m, what high jump would you predict?
d) Why can't you use this line to estimate the long jump for a year when you know the high jump was 2.25 m?
e) Write the equation of the line you need to make that prediction.

31. **French** Consider the association between a student's score on a French vocabulary test and the weight of the student. What direction and strength of correlation would you expect in each of the following situations? Explain.
a) The students are all in third grade.
b) The students are in third through twelfth grades in the same school district.
c) The students are in tenth grade in France.
d) The students are in third through twelfth grades in France.

32. **Twins** Twins are often born at less than 9 months gestation. The graph from the *Journal of the American Medical Association (JAMA)* shows the rate of preterm twin births in the United States over the past 20 years. In this study, *JAMA* categorized mothers by the level of prenatal medical care they received: inadequate, adequate, or intensive.

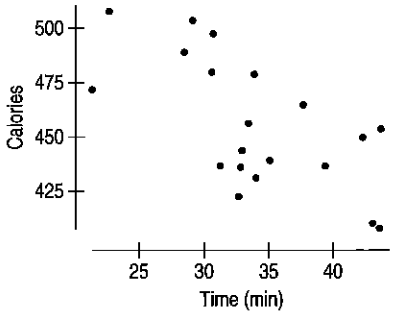
a) Describe the overall trend in preterm twin births.
b) Describe any differences you see in this trend, depending on the level of prenatal medical care the mother received.
c) Should expectant mothers be advised to cut back on the level of medical care they seek in the hope of avoiding preterm births? Explain.



Preterm Birth Rate per 100 live twin births among U.S. twins by intensive, adequate, and less than adequate prenatal care utilization, 1981–1997. (Source: *JAMA* 284[2000]: 335–341)

33. Lunchtime Does how long toddlers sit at the lunch table help predict how much they eat? The table and graph show the number of minutes the kids stayed at the table and the number of calories they consumed. Create and interpret a model for these data.

Calories	Time	Calories	Time
472	21.4	450	42.4
498	30.8	410	43.1
465	37.7	504	29.2
456	33.5	437	31.3
423	32.8	489	28.6
437	39.5	436	32.9
508	22.8	480	30.6
431	34.1	439	35.1
479	33.9	444	33.0
454	43.8	408	43.7



34. Gasoline Since clean-air regulations have dictated the use of unleaded gasoline, the supply of leaded gas in New York state has diminished. The following table was given on the August 2001 New York State Math B exam, a statewide achievement test for high school students.

Year	1984	1988	1992	1996	2000
Gallons (1000's)	150	124	104	76	50

a) Create a linear model and predict the number of gallons that will be available in 2015. Comment.
b) The exam then asked students to estimate the year when leaded gasoline will first become unavailable, expecting them to use the model from part a to answer the question. Explain why that method is incorrect.
c) Create a model that *would* be appropriate for that task, and make the estimate.
d) The "wrong" answer from the other model is fairly accurate in this case. *Why?*

35. Tobacco and alcohol Are people who use tobacco products more likely to consume alcohol? Here are data on household spending (in pounds) taken by the British government on 11 regions in Great Britain. Do tobacco and alcohol spending appear to be related? What questions do you have about these data? What conclusions can you draw?

Region	Alcohol	Tobacco
North	6.47	4.03
Yorkshire	6.13	3.76
Northeast	6.19	3.77
East Midlands	4.89	3.34
West Midlands	5.63	3.47
East Anglia	4.52	2.92
Southeast	5.89	3.20
Southwest	4.79	2.71
Wales	5.27	3.53
Scotland	6.08	4.51
Northern Ireland	4.02	4.56

36. Football weights The Sears Cup was established in 1993 to honor institutions that maintain a broad-based athletic program, achieving success in many sports, both men's and women's. Since its Division III inception in 1995, the cup has been won by Williams College 15 of 17 years. Their football team has an 85.3% winning record under their current coach. Why does the football team win so much? Is it because they're heavier than their opponents? The table shows the average team weights for selected years from 1973 to 1993.

Year	Weight (lb)	Year	Weight (lb)
1973	185.5	1983	192.0
1975	182.4	1987	196.9
1977	182.1	1989	202.9
1979	191.1	1991	206.0
1981	189.4	1993	198.7

a) Fit a straight line to the relationship between *Weight* and *Year*.

b) Does a straight line seem reasonable?

c) Predict the average weight of the team for the year 2015. Does this seem reasonable?

d) What about the prediction for the year 2103? Explain.

e) What about the prediction for the year 3003? Explain.

37. Models Find the predicted value of y, using each model for $x = 10$.

a) $\hat{y} = 2 + 0.8 \ln x$ b) $\log \hat{y} = 5 - 0.23x$

c) $\dfrac{1}{\sqrt{\hat{y}}} = 17.1 - 1.66x$

38. Williams vs. Texas Here are the average weights of the football team for the University of Texas for various years in the 20th century.

Year	1905	1919	1932	1945	1955	1965
Weight (lb)	164	163	181	192	195	199

a) Fit a straight line to the relationship of *Weight* by *Year* for Texas football players.

b) According to these models, in what year will the predicted weight of the Williams College team from Exercise 36 first be more than the weight of the University of Texas team?

c) Do you believe this? Explain.

39. Vehicle weights The Minnesota Department of Transportation hoped that they could measure the weights of big trucks without actually stopping the vehicles by using a newly developed "weigh-in-motion" scale. After installation of the scale, a study was conducted to find out whether the scale's readings correspond to the true weights of the trucks being monitored. In Chapter 6, Exercise 50, you examined the scatterplot for the data they collected, finding the association to be approximately linear with $R^2 = 93\%$. Their regression equation is $\widehat{Wt} = 10.85 + 0.64\,Scale$, where both the scale reading and the predicted weight of the truck are measured in thousands of pounds.

a) Estimate the weight of a truck if this scale read 31,200 pounds.

b) If that truck actually weighed 32,120 pounds, what was the residual?

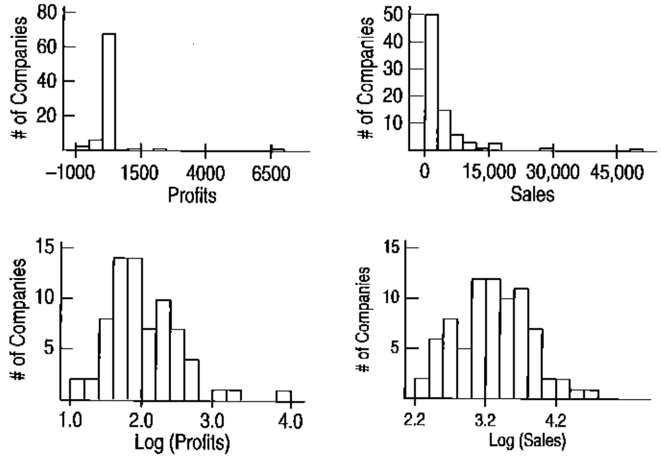
c) If the scale reads 35,590 pounds, and the truck has a residual of −2440 pounds, how much does it actually weigh?

d) In general, do you expect estimates made using this equation to be reasonably accurate? Explain.

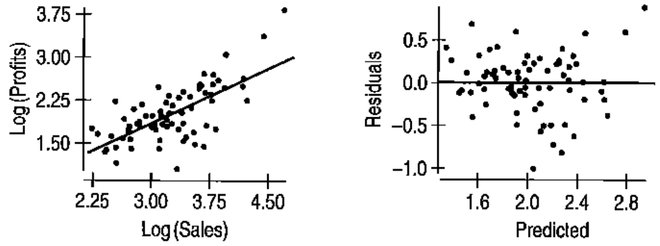
e) If the police plan to use this scale to issue tickets to trucks that appear to be overloaded, will negative or positive residuals be a greater problem? Explain.

40. Profit How are a company's profits related to its sales? Let's examine data from 71 large U.S. corporations. All amounts are in millions of dollars.

a) Histograms of *Profits* and *Sales* and histograms of the logarithms of *Profits* and *Sales* are seen below. Why are the re-expressed data better for regression?



b) Here are the scatterplot and residuals plot for the regression of logarithm of *Profits* vs. log of *Sales*. Do you think this model is appropriate? Explain.



c) Here's the regression analysis. Write the equation.

```
Dependent variable is Log Profit
R-squared = 48.1%

Variable      Coefficient
Intercept     −0.106259
LogSales       0.647798
```

d) Use your equation to estimate profits earned by a company with sales of 2.5 billion dollars. (That's 2500 million.)

41. Down the drain Most water tanks have a drain plug so that the tank may be emptied when it's to be moved or repaired. How long it takes a certain size of tank to drain depends on the size of the plug, as shown in the table. Create a model.

Plug Dia (in.)	$\frac{3}{8}$	$\frac{1}{2}$	$\frac{3}{4}$	1	$1\frac{1}{4}$	$1\frac{1}{2}$	2
Drain Time (min.)	140	80	35	20	13	10	5

42. Chips A start-up company has developed an improved electronic chip for use in laboratory equipment. The company needs to project the manufacturing cost, so it develops a spreadsheet model that takes into account the purchase of production equipment, overhead, raw materials, depreciation, maintenance, and other business costs. The spreadsheet estimates the cost of producing 10,000 to 200,000 chips per year, as seen in the table. Develop a regression model to predict *Costs* based on the *Level* of production.

Chips Produced (1000s)	Cost per Chip ($)	Chips Produced (1000s)	Cost per Chip ($)
10	146.10	90	47.22
20	105.80	100	44.31
30	85.75	120	42.88
40	77.02	140	39.05
50	66.10	160	37.47
60	63.92	180	35.09
70	58.80	200	34.04
80	50.91		