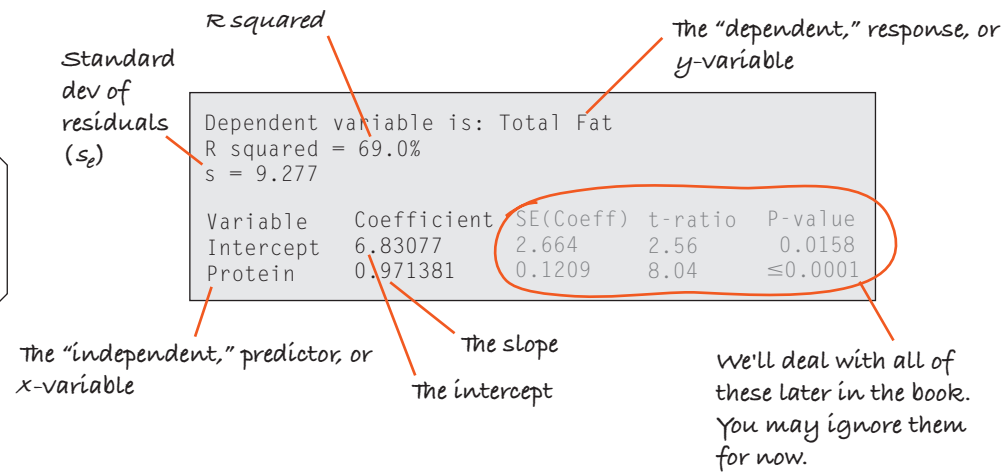
REGRESSION ON THE COMPUTER

All statistics packages make a table of results for a regression. These tables may differ slightly from one package to another, but all are essentially the same—and all include much more than we need to know for now. Every computer regression table includes a section that looks something like this:



Finding Least Squares Lines. We almost always use technology to find regressions. Practice now—just in time for the exercises.

The slope and intercept coefficient are given in a table such as this one. Usually the slope is labeled with the name of the x-variable, and the intercept is labeled "Intercept" or "Constant." So the regression equation shown here is

$$\widehat{Fat} = 6.83077 + 0.97138 Protein.$$

It is not unusual for statistics packages to give many more digits of the estimated slope and intercept than could possibly be estimated from the data. (The original data were reported to the nearest gram.) Ordinarily, you should round most of the reported numbers to one digit more than the precision of the data, and the slope to two. We will learn about the other numbers in the regression table later in the book. For now, all you need to be able to do is find the coefficients, the s_e, and the R^2 value.

EXERCISES

1. **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?

2. **Horsepower.** In Chapter 7's Exercise 33 we examined the relationship between the fuel economy (mpg) and horsepower for 15 models of cars. Further analysis produces the regression model $\widehat{mpg} = 46.87 - 0.084 HP$. If the car you are thinking of buying has a 200-horsepower engine, what does this model suggest your gas mileage would be?

3. **More cereal.** Exercise 1 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?

4. **Horsepower, again.** Exercise 2 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?

5. **Another bowl.** In Exercise 1, 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.

6. **More horsepower.** In Exercise 2, the regression model $\widehat{mpg} = 46.87 - 0.084 HP$ relates cars' horsepower to their fuel economy (in mpg). Explain what the slope means.

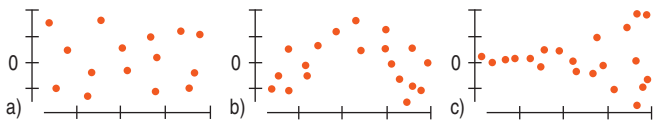
7. **Cereal again.** The correlation between a cereal's fiber and potassium contents is $r = 0.903$. What fraction of the variability in potassium is accounted for by the amount of fiber that servings contain?

8. **Another car.** The correlation between a car's horsepower and its fuel economy (in mpg) is $r = -0.869$. What fraction of the variability in fuel economy is accounted for by the horsepower?

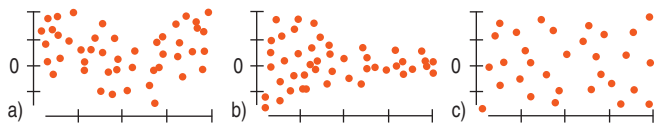
9. **Last bowl!** For Exercise 1's regression model predicting potassium content (in milligrams) from the amount of fiber (in grams) in breakfast cereals, $s_e = 30.77$. Explain in this context what that means.

10. **Last tank!** For Exercise 2's regression model predicting fuel economy (in mpg) from the car's horsepower, $s_e = 3.287$. Explain in this context what that means.

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



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



13. **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.

14. **What slope?** 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.

15. **Real estate.** A random sample of records of sales of homes 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*-squared 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.

T 16. **Roller coaster.** People who responded to a July 2004 Discovery Channel poll named the 10 best roller coasters in the United States. A table in the last 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 $R^2 = 12.4\%$.
   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.

17. **Real estate again.** The regression of *Price* on *Size* of homes in Albuquerque had $R^2 = 71.4\%$, as described in Exercise 15. Write a sentence (in context, of course) summarizing what the R^2 says about this regression.

T 18. **Coasters again.** Exercise 16 examined the association between the *Duration* of a roller coaster ride and the height of its initial *Drop*, reporting that $R^2 = 12.4\%$. Write a sentence (in context, of course) summarizing what the R^2 says about this regression.

19. **Real estate redux.** The regression of *Price* on *Size* of homes in Albuquerque had $R^2 = 71.4\%$, as described in Exercise 15.
   a) What is the correlation between *Size* and *Price*? Explain why you chose the sign (+ or −) you did.
   b) What would you predict about the *Price* of a home 1 standard deviation above average in *Size*?
   c) What would you predict about the *Price* of a home 2 standard deviations below average in *Size*?

T 20. **Another ride.** The regression of *Duration* of a roller coaster ride on the height of its initial *Drop*, described in Exercise 16, had $R^2 = 12.4\%$.
   a) What is the correlation between *Drop* and *Duration*? Explain why you chose the sign (+ or −) you did.
   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*?

21. **More real estate.** Consider the Albuquerque home sales from Exercise 15 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?

T 22. **Last ride.** Consider the roller coasters described in Exercise 16 again. The regression analysis gives the model $\widehat{Duration} = 91.033 + 0.242\ Drop$.
   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?

23. **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.

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

a) 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*.

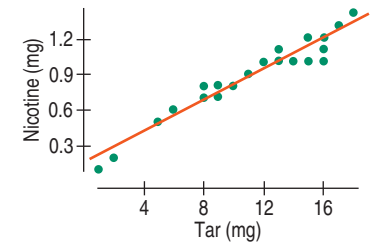
25. ESP. People who claim to "have ESP" participate in a screening test in which they have to guess which of several images someone is thinking of. You and a friend both took the test. You scored 2 standard deviations above the mean, and your friend scored 1 standard deviation below the mean. The researchers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

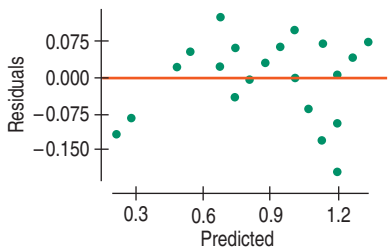
a) Should you choose to take this retest? Explain.

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

26. *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.

T **27. Cigarettes.** Is the nicotine content of a cigarette related to the "tars"? A collection of data (in milligrams) on 29 cigarettes produced the scatterplot, residuals plot, and regression analysis shown:





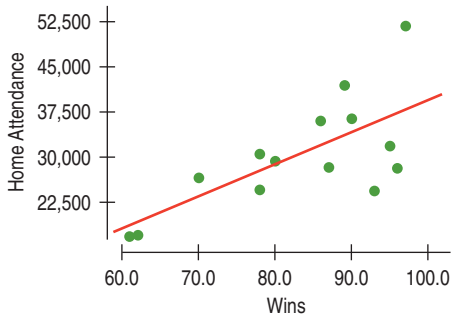
Dependent variable is: nicotine
R squared = 92.4%

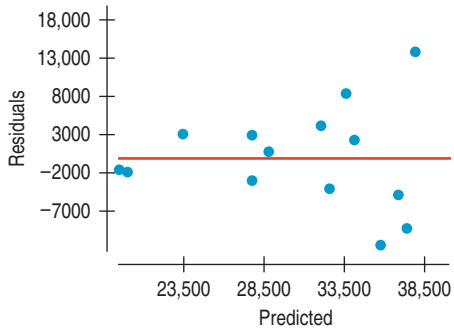
Variable	Coefficient
Constant	0.154030
Tar	0.065052

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

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

T **28. Attendance 2006.** 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 2006 season. Here are the scatterplot, the residuals plot, and part of the regression analysis:





Dependent variable is: Home Attendance
R squared = 48.5%

Variable	Coefficient
Constant	−14364.5
Wins	538.915

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?

T **29. Another cigarette.** Consider again the regression of *Nicotine* content on *Tar* (both in milligrams) for the cigarettes examined in Exercise 27.

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?

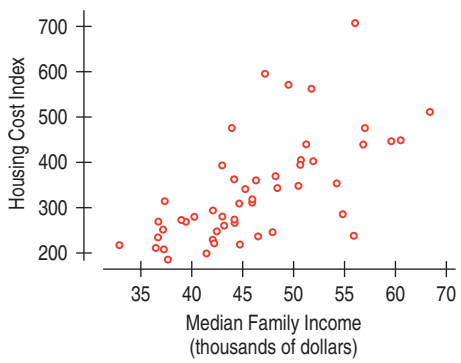
T **30. Second inning 2006.** Consider again the regression of *Average Attendance* on *Wins* for the baseball teams examined in Exercise 28.

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?

31. Last cigarette. Take another look at the regression analysis of tar and nicotine content of the cigarettes in Exercise 27.
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?

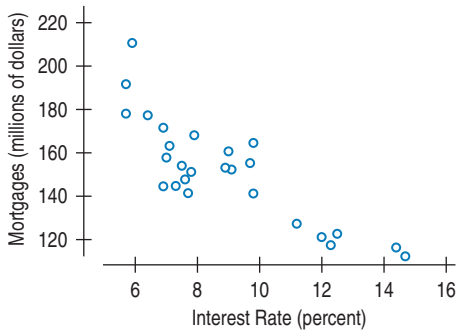
32. Last inning 2006. Refer again to the regression analysis for average attendance and games won by American League baseball teams, seen in Exercise 28.
a) Write the equation of the regression line.
b) Estimate the *Average Attendance* for a team with 50 *Wins*.
c) Interpret the meaning of the slope of the regression line in this context.
d) In general, what would a negative residual mean in this context?
e) The St. Louis Cardinals, the 2006 World Champions, are not included in these data because they are a National League team. During the 2006 regular season, the Cardinals won 83 games and averaged 42,588 fans at their home games. Calculate the residual for this team, and explain what it means.

33. Income and housing revisited. In Chapter 7, Exercise 31, 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.



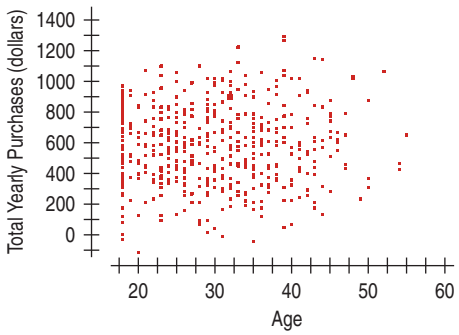
a) Is a regression analysis appropriate? Explain.
b) What is the equation that predicts Housing Cost Index from median family income?
c) For a state with MFI = $44,993, what would be the predicted HCI?
d) Washington, DC, has an MFI of $44,993 and an HCI of 548.02. How far off is the prediction in b) 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?

34. Interest rates and mortgages again. In Chapter 7, Exercise 32, we saw a plot of total mortgages in the United States (in millions of 2005 dollars) versus the interest rate at various times over the past 26 years. The correlation is $r = -0.84$. The mean mortgage amount is $151.9 million and the mean interest rate is 8.88%. The standard deviations are $23.86 million for mortgage amounts and 2.58% 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 20%?
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?

35. 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:*

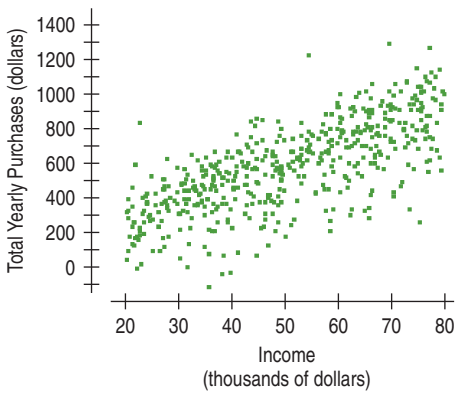


The correlation between *Total Yearly Purchases* and *Age* is $r = 0.037$. Summary statistics for the two variables are:

	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 average *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.

36. Online clothes II. For the online clothing retailer discussed in the previous problem, the scatterplot of *Total Yearly Purchases* by *Income* shows

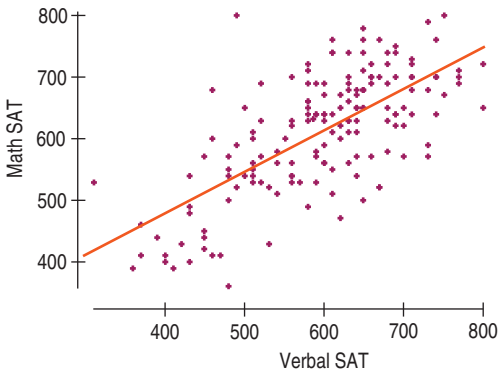


The correlation between *Total Yearly Purchases* and *Income* is 0.722. Summary statistics for the two variables are:

	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 average *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.

T **37. 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. Doing the SAT-Math problems also involves 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 student scores a perfect 1600. Based on this model, what would be that student's Math score residual?

38. 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 1833, 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 2100.

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.

39. SAT, take 2. Suppose we wanted to use SAT math scores to estimate verbal scores based on the information in Exercise 37.
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 37, predict her math score.
f) Why doesn't the result in part e) come out to 500?

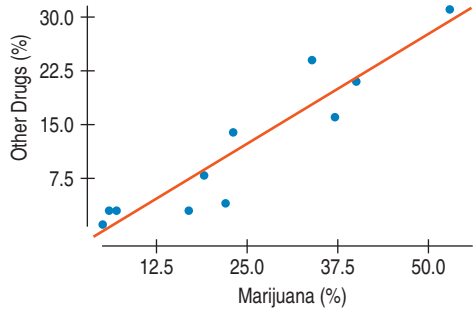
40. Success, part 2. Based on the statistics for college freshmen given in Exercise 38, what SAT score might be expected among freshmen who attained a first-semester GPA of 3.0?

T **41. Used cars 2007.** Classified ads in the *Ithaca Journal* offered several used Toyota Corollas for sale. Listed below are the ages of the cars and the advertised prices.

Age (yr)	Price Advertised ($)
1	13,990
1	13,495
3	12,999
4	9500
4	10,495
5	8995
5	9495
6	6999
7	6950
7	7850
8	6999
8	5995
10	4950
10	4495
13	2850

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 $R^2 = 94.4\%$. 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?

T **42. 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?

T **43. More used cars 2007.** Use the advertised prices for Toyota Corollas given in Exercise 41 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 $3500. What is the residual?
g) Would this regression model be useful in establishing a fair price for a 20-year-old car? Explain.

T **44. Birthrates 2005.** 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
Rate	19.4	18.4	14.8	15.9	15.6	16.4	14.8	14.4	14.0

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.

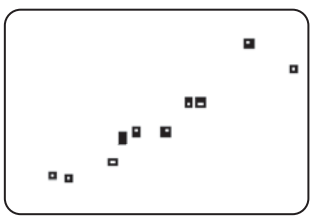
45. Burgers. In the last chapter, you examined the association between the amounts of *Fat* and *Calories* in fast-food hamburgers. Here are the data:

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?

46. 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 to estimate calories from the fat content.
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?

47. A second helping of burgers. In Exercise 45 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.

48. A second helping of chicken. In Exercise 46 you created a model to estimate the number of *Calories* in a chicken sandwich when you know the *Fat*.
a) Explain why you cannot use that model to estimate the fat content of a 400-calorie sandwich.
b) Make that estimate using an appropriate model.

T **49. 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.

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

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?

T **50. 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 49, create and analyze that model.

T **51. Heptathlon 2004.** We discussed the women's 2004 Olympic heptathlon in Chapter 6. The table on the next page shows 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.

Name	Country	High Jump (m)	800-m (sec)	Long Jump (m)
Carolina Klüft	SWE	1.91	134.15	6.51
Austra Skujyté	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.)

T **52. Heptathlon 2004 again.** We saw the data for the women's 2004 Olympic heptathlon in Exercise 51. 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.)

53. 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.

54. 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.

JUST CHECKING Answers

1. You should expect the price to be 0.77 standard deviations above the mean.
2. You should expect the size to be $2(0.77) = 1.54$ standard deviations below the mean.
3. The home is 1.5 standard deviations above the mean in size no matter how size is measured.
4. An increase in home size of 1000 square feet is associated with an increase in price of $94,454, on average.
5. Units are thousands of dollars per thousand square feet.
6. About $188,908, on average
7. No. Even if it were positive, no one wants a house with 0 square feet!
8. Negative; that indicates it's priced lower than a typical home of its size.
9. $280,245
10. $19,755 (positive!)
11. Differences in the size of houses account for about 59.5% of the variation in the house prices.
12. It's positive. The correlation and the slope have the same sign.
13. R^2 would not change, but the slope would. Slope depends on the units used but correlation doesn't.
14. No, the standard deviation of the residuals is 53.79 thousand dollars. We shouldn't be surprised by any residual smaller than 2 standard deviations, and a residual of $100,000 is less than 2(53,790).