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

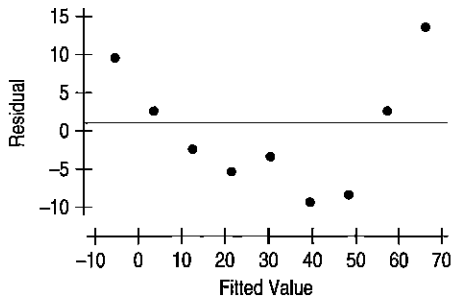
11. Bookstore sales last time For the regression model for the bookstore of Exercise 5, what is the value of R^2 and what does it mean?

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

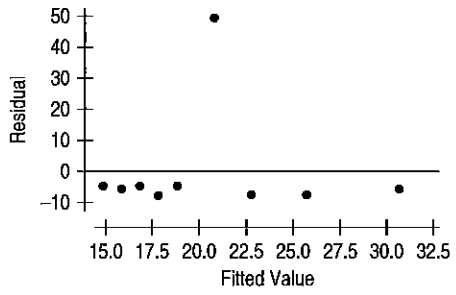
Section 7.7

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.

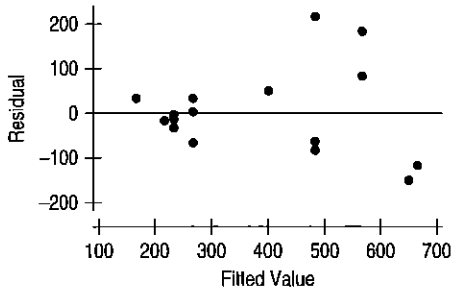
a)



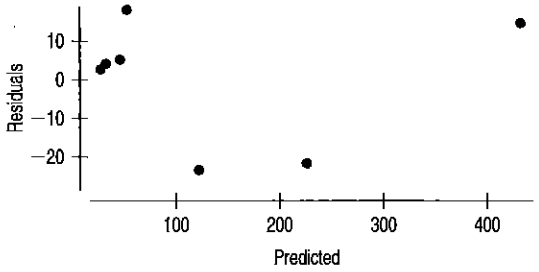
b)



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.

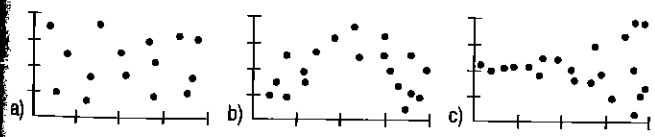
25. Regression equations Fill in the missing information in the following table.

	$\bar{x}$	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		$\hat{y} = -100 + 50x$

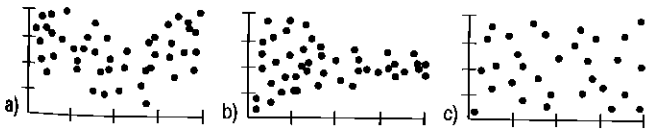
26. More regression equations Fill in the missing information in the following table.

	$\bar{x}$	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.

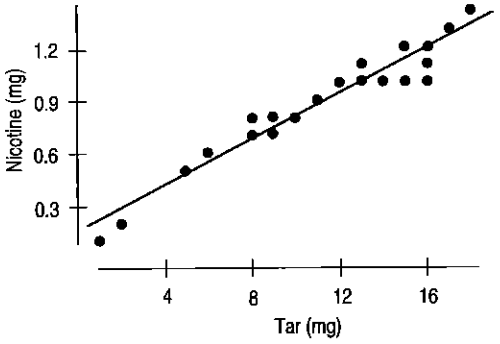
41. More real estate Consider the Albuquerque home sales from Exercise 29 again. The regression analysis gives the model $\widehat{Price} = 47.82 + 0.061\ Size$.

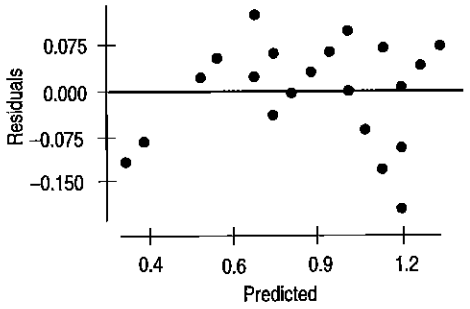
a) Explain what the slope of the line says about housing prices and house size.
b) What price would you predict for a 3000-square-foot house in this market?
c) A real estate agent shows a potential buyer a 1200-square-foot home, saying that the asking price is \$6000 less than what one would expect to 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:





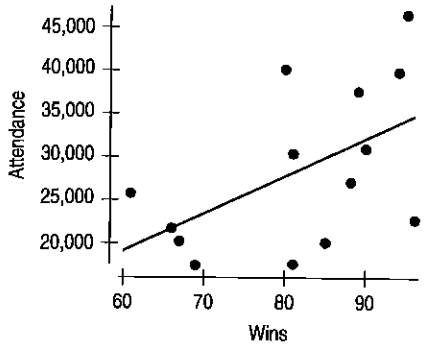
Dependent variable is Nicotine
R-squared = 92.4%

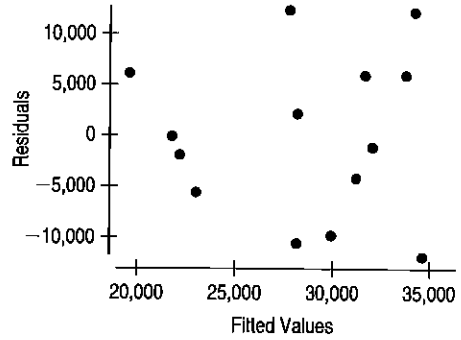
Variable	Coefficient
Constant	0.154030
Tar	0.065052

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?

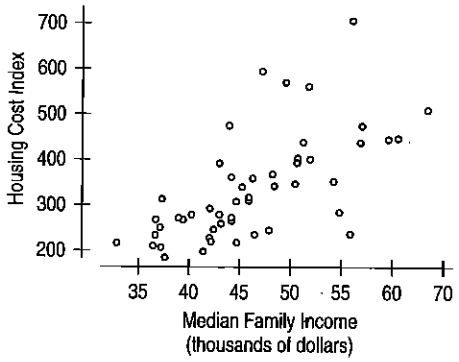
47. Last cigarette Take another look at the regression analysis of tar and nicotine content of the cigarettes in Exercise 43.

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.

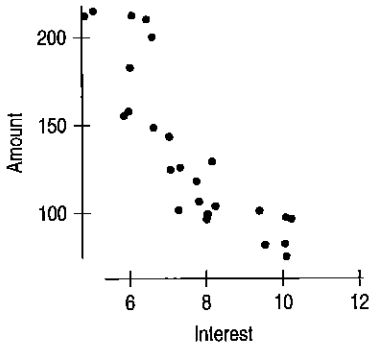
49. Income and housing revisited In Chapter 6, Exercise 39, we learned that the Office of Federal Housing Enterprise Oversight (OFHEO) collects data on various aspects of housing costs around the United States. Here's a scatterplot (by state) of the *Housing Cost Index* (HCI) versus the *Median Family 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?

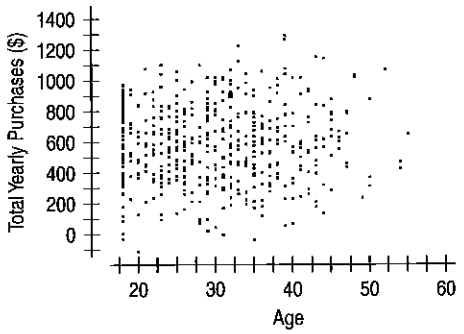
e) If we standardized both variables, what would be the regression equation that predicts standardized HCI from standardized MFI?
f) If we standardized both variables, what would be the regression equation that predicts standardized MFI from standardized HCI?

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:*

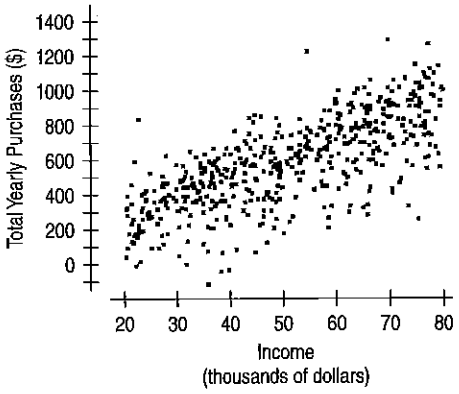


The correlation between *Total Yearly Purchases* and *Age* is $r = 0.037$. Summary statistics for the two variables are:

	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:



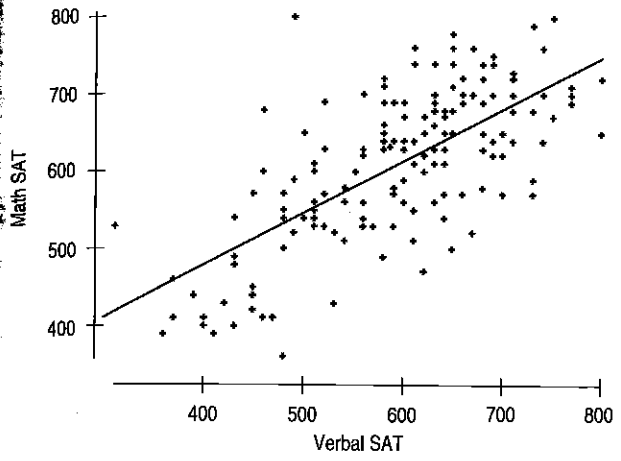
The correlation between *Total Yearly Purchases* and *Income* is 0.722. Summary statistics for the two variables are:

	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.

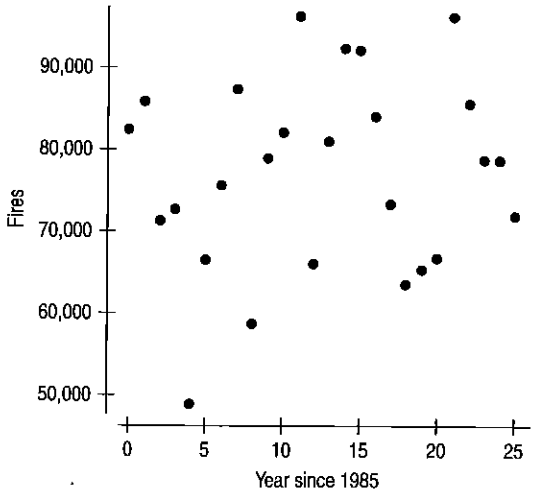
a) Write the equation of the regression line.
b) Explain what the y-intercept of the regression line indicates.
c) Interpret the slope of the regression line.
d) Predict the GPA of a freshman who scored a combined 1400.
e) Based upon these statistics, how effective do you think SAT scores would be in 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.



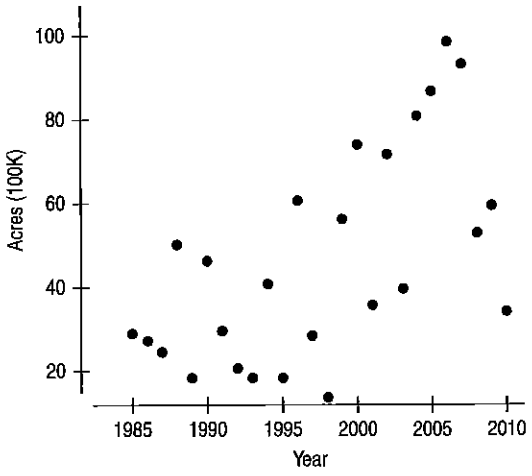
Dependent variable is Fires
R-squared = 1.9%
s = 11920

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?

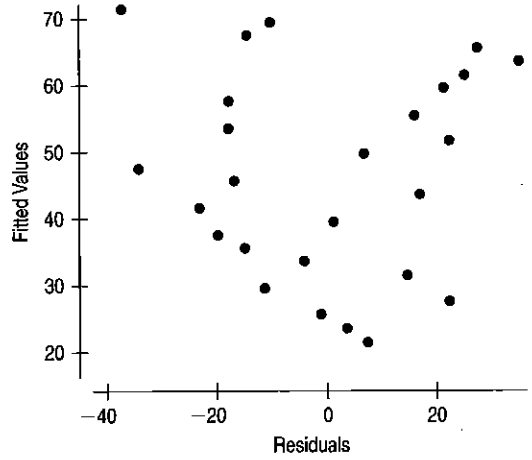
58. Wildfires 2010—sizes We saw in Exercise 57 that the number of fires was nearly constant. But has the damage they cause remained constant as well? Here's a regression that examines the trend in *Acres per Fire,*

(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



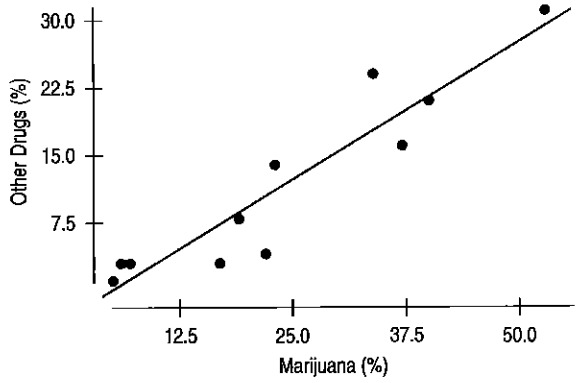
a) Is the regression model appropriate for these data? Explain.
b) What interpretation (if any) can you give for the R^2 in the regression table?

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.

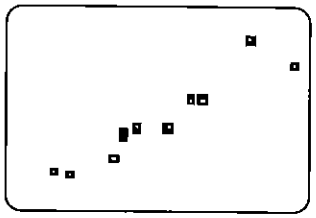
63. Burgers In the last chapter, you examined the association between the amounts of *Fat* and *Calories* in fast-food hamburgers. Here are the data:

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	

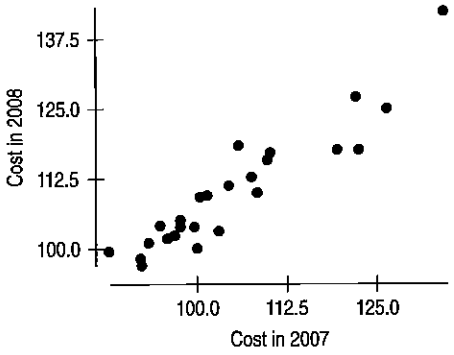


a) Do you think a linear model is appropriate in this situation?
b) Describe the strength of this association.
c) Write the equation of the regression line.
d) Explain the meaning of the slope.
e) Explain the meaning of the *y*–intercept.
f) What does it mean if a certain sandwich has a negative residual?
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.

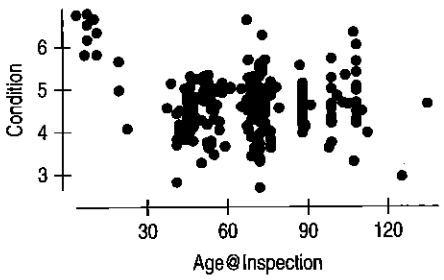
65. A second helping of burgers In Exercise 63, you created a model that can estimate the number of *Calories* in a burger when the *Fat* content is known.
a) Explain why you cannot use that model to estimate the fat content of a burger with 600 calories.
b) Using an appropriate model, estimate the fat 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

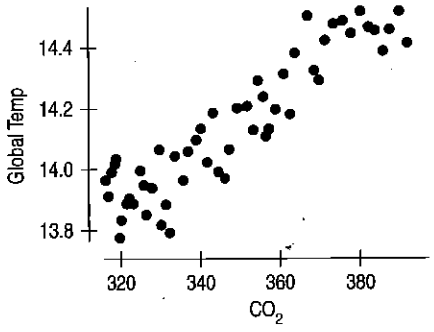
Variable	Coefficient
Intercept	4.95147
Age@Inspection	−0.00481

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.

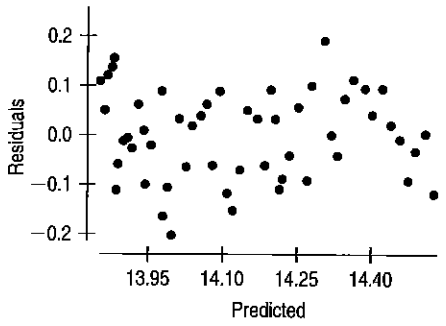


A regression predicting *Temperature* from CO_2 produces the following output table (in part):

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

Variable	Coefficient
Intercept	11.0276
CO_2 (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?

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.

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