Chapter 7 – Linear Regression

Section 7.1

1. True or false.

a) False. The line usually touches none of the points. The line minimizes the sum of least squares.

b) True.

c) False. Least squares means the sum of all the squared residuals is minimized.

2. True or false II.

a) True.

b) False. Least squares means the sum of all the squared residuals is minimized.

c) True.

Section 7.2

3. Least square interpretations.

The weight of a newborn boy can be predicted as –5.94 kg plus 0.1875 kg per cm of length. This is a model fit to data. Parents should not be concerned if their newborn's length and weight don't fit this equation. No particular baby should be expected to fit this equation exactly.

4. Residual interpretations.

$\widehat{weight} = -5.94 + 0.1875 length = -5.94 + 0.1875(48) = 3.06$ kg

$\text{Residual} = weight - \widehat{weight} = 3 - 3.06 = -0.06$ kg

The newborn was slightly lighter than the weight predicted by his length.

Section 7.3

5. Bookstore sales revisited.

a) The slope of the line of best fit is $b_1 = r\dfrac{s_y}{s_x} = (0.965)\dfrac{5.34}{5.64} = 0.914$. (When using technology, the slope is 0.913. When calculated by hand, the standard deviations and correlation must be rounded, resulting in a slight inaccuracy.)

b) The model predicts an increase in sales of 0.914($1000), or $914, for each additional sales person working.

c) The intercept, b_0, is 8.09. (When using technology, the intercept is 8.10. When calculated by hand, the means and slope must be rounded, resulting in a slight inaccuracy.)

$$\hat{y} = b_0 + b_1 x \Rightarrow \overline{y} = b_0 + b_1 \overline{x} \Rightarrow 17.6 = b_0 + (0.914)(10.4) \Rightarrow b_0 = 8.09$$

d) The model predicts that average sales would be approximately $8.10($1000), or $810, when there were no sales people working. This doesn't make sense in this context.

e) $\widehat{Sales} = 8.09 + 0.914\, People$ (hand calculation) $\widehat{Sales} = 8.10 + 0.913\, People$ (technology)

f) $\widehat{Sales} = 8.09 + 0.914\, People = 8.09 + 0.914(18) = 24.542$; According to the model, we would expect sales to be approximately $24,540 when 18 sales people are working. ($24,530 using technology)

g) $\text{Residual} = sales - \widehat{sales} = \$25,000 - \$24,540 = \$460$ (Using technology, $470)

h) Since the residual is positive, we have underestimated the sales.

6. Disk drives 2016 again.

a) The slope of the line of best fit is $b_1 = r\frac{s_y}{s_x} = (0.9876)\frac{1418.67}{9.855} = 142.17$. (When using technology, the slope is 142.18. When calculated by hand, the standard deviations and correlation must be rounded, resulting in a slight inaccuracy.)

b) The model predicts an average increase of \$142.17 for each additional TB of storage.

c) The intercept, b_0, is –\$296.25. (When using technology, the intercept is –\$296.31. When calculated by hand, the means and slope must be rounded, resulting in a slight inaccuracy.

d) According to the model, the average cost of a drive with 0 TB capacity is expected to be –\$296.25. This makes no sense in this context.

$$\hat{y} = b_0 + b_1 x \Rightarrow \overline{y} = b_0 + b_1 \overline{x} \Rightarrow 785.819 = b_0 + (142.17)(7.6111) \Rightarrow b_0 = -296.25$$

e) $\widehat{Price} = -296.25 + 142.17\,Capacity$ (hand calculation)

$\widehat{Price} = -296.31 + 142.18\,Capacity$ (technology)

f) $\widehat{Price} = -296.25 + 142.17\,Capacity = -296.25 + 142.17(20) = \$2547.15$; According to the model, we would expect the price of a 20 TB drive to average approximately \$2547.15. (Answers may vary depending on the values used for the slope and the intercept.)

g) Residual = $Price - \widehat{Price} = 2017.86 - 2547.15 = -\$529.29$. This drive is a good buy. It costs \$529.29 less than you expected to pay.

h) Since the residual is negative, the model overestimates the price.

i) No, this does not mean the model is accurate. We saw from the scatterplot that the relationship is curved. The model may not be accurate, or even appropriate.

Section 7.4

7. Sophomore slump?

The winners may be suffering from regression to the mean. Perhaps they weren't really better than other rookie executives, but just happened to have a lucky year the first year. When their performance the following year landed them closer to the mean of the others, it looked like their performance had suffered.

8. Sophomore slump again?

Although on average, the performance of funds will cluster around the mean, we can't predict how any particular fund will do.

Section 7.5

9. Bookstore sales once more.

a) The residuals are measured in the same units as the response variable, thousands of dollars.

b) The residual with the largest magnitude, 2.77, contributes most to the sum of the squared residuals.

c) The residual with the smallest magnitude, 0.07, contributes least to the sum of the squared residuals.

10. Disk drives 2016, residuals.

a) The drive with a capacity of 12 TB, with a residual of –329.80, contributes the most to the sum of squared residuals, since it has the residual with the largest magnitude.

b) A negative residual means that the drive costs less than we might expect from this model and its capacity. For example, a residual of –245.00 indicates a drive that costs \$245.00 less than we might expect.

Section 7.6

11. Bookstore sales last time.

$R^2 = 93.12\%$; Approximately 93% of the variance in *Sales* can be accounted for by the regression of *Sales* on *Number of Sales Workers*.

12. Disk drives encore.

$R^2 = 97.54\%$; Approximately 97.54% of the variance in the price of these disk drives can be accounted for by the regression on *Capacity*.

Section 7.7

13. Residual plots

a) The residual plot has a clear curved pattern. The linearity assumption is violated.

b) One point on the residual plot has a much larger residual than the others. The outlier condition is violated.

c) The residual plot shows a fanned shape. The equal spread condition is violated.

14. Disk drives 2016, residuals again.

a) The residuals show a curved pattern, or possibly an outlier. Either one would violate the Linearity Assumption.

b) Possibly, we should consider a re-expression of the data to improve the straightness

Chapter Exercises

15. Cereals.

$\widehat{Potassium} = 38 + 27Fiber = 38 + 27(9) = 281\,\text{mg}$; According to the model, we expect cereal with 9 grams of fiber to have 281 milligrams of potassium.

16. Engine size.

$\widehat{CombinedMPG} = 33.46 - 3.23Displacement = 33.46 - 3.23(4) \approx 20.54$ mpg; According to the model, we expect a car with a 4-liter engine to get about 20.54 miles per gallon.

17. More cereal.

A negative residual means that the potassium content is actually lower than the model predicts for a cereal with that much fiber.

18. Engine size, again.

A positive residual means that the car gets better gas mileage than the model predicts for a car with that size engine.

19. Another bowl.

The model predicts that cereals will have approximately 27 more milligrams of potassium for each additional gram of fiber.

20. More engine size.

The model predicts that cars lose an average of 3.23 miles per gallon for each additional liter of engine size.

21. Cereal again.

$R^2 = (0.903)^2 \approx 0.815$; About 81.5% of the variability in potassium content is accounted for by the model.

22. Another car.

$R^2 = (-0.774)^2 \approx 0.599$; About 60% of the variability in fuel economy is accounted for by the model.

23. Last bowl!

True potassium contents of cereals vary from the predicted values with a standard deviation of 30.77 milligrams.

24. Last tank!

True fuel economy varies from the predicted amount with a standard deviation of 3.522 miles per gallon.

25. Regression equations.

	$\bar{x}$	s_x	$\bar{y}$	s_y	r	$\hat{y} = b_0 + b_1 x$
a)	10	2	20	3	0.5	$\hat{y} = 12.5 + 0.75x$
b)	2	0.06	7.2	1.2	–0.4	$\hat{y} = 23.2 - 8x$
c)	12	6	152	30	–0.8	$\hat{y} = 200 - 4x$
d)	2.5	1.2	25	100	0.6	$\hat{y} = -100 + 50x$

a) $b_1 = r\dfrac{s_y}{s_x}$ $\quad$ $\hat{y} = b_0 + b_1 x$

$b_1 = (0.5)\dfrac{3}{2}$ $\quad$ $\bar{y} = b_0 + b_1 \bar{x}$

$b_1 = 0.75$ $\quad$ $20 = b_0 + 0.75(10)$

$b_0 = 12.5$

b) $b_1 = r\dfrac{s_y}{s_x}$ $\quad$ $\hat{y} = b_0 + b_1 x$

$b_1 = (-0.4)\dfrac{1.2}{0.06}$ $\quad$ $\bar{y} = b_0 + b_1 \bar{x}$

$b_1 = -8$ $\quad$ $7.2 = b_0 - 8(2)$

$b_0 = 23.2$

c) $\hat{y} = b_0 + b_1 x$ $\quad$ $b_1 = r\dfrac{s_y}{s_x}$

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

$\bar{y} = 200 - 4(12)$ $\quad$ $-4 = (-0.8)\dfrac{s_y}{6}$

$\bar{y} = 152$ $\quad$ $s_y = 30$

d) $\hat{y} = b_0 + b_1 x$ $\quad$ $b_1 = r\dfrac{s_y}{s_x}$

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

$\bar{y} = -100 + 50(2.5)$ $\quad$ $50 = r\dfrac{100}{1.2}$

$\bar{y} = 25$ $\quad$ $r = 0.6$

26. More regression equations.

	$\bar{x}$	s_x	$\bar{y}$	s_y	r	$\hat{y} = b_0 + b_1 x$
a)	30	4	18	6	–0.2	$\hat{y} = 27 - 0.3x$
b)	100	18	60	10	0.9	$\hat{y} = 10 + 0.5x$
c)	4	0.8	50	15	0.8	$\hat{y} = -10 + 15x$
d)	6	1.2	18	4	–0.6	$\hat{y} = 30 - 2x$

a) $b_1 = r\dfrac{s_y}{s_x}$ $\quad$ $\hat{y} = b_0 + b_1 x$

$b_1 = (-0.2)\dfrac{6}{4}$ $\quad$ $\bar{y} = b_0 + b_1 \bar{x}$

$b_1 = -0.3$ $\quad$ $18 = b_0 - 0.3(30)$

$b_0 = 27$

b) $b_1 = r\dfrac{s_y}{s_x}$ $\quad$ $\hat{y} = b_0 + b_1 x$

$b_1 = (0.9)\dfrac{10}{18}$ $\quad$ $\bar{y} = b_0 + b_1 \bar{x}$

$b_1 = 0.5$ $\quad$ $60 = b_0 + 0.5(100)$

$b_0 = 10$

26. (continued)

c) $\hat{y} = b_0 + b_1 x$
$\overline{y} = b_0 + b_1 \overline{x}$
$50 = -10 + 15(\overline{x})$
$\overline{x} = 4$

$b_1 = r\dfrac{s_y}{s_x}$
$15 = r\dfrac{15}{0.8}$
$r = 0.8$

d) $\hat{y} = b_0 + b_1 x$
$\overline{y} = b_0 + b_1 \overline{x}$
$18 = 30 - 2(\overline{x})$
$\overline{x} = 6$

$b_1 = r\dfrac{s_y}{s_x}$
$-2 = (-0.6)\dfrac{4}{s_x}$
$s_x = 1.2$

27. Residuals.

a) The scattered residuals plot indicates an appropriate linear model.

b) The curved pattern in the residuals plot indicates that the linear model is not appropriate. The relationship is not linear.

c) The fanned pattern indicates that the linear model is not appropriate. The model's predicting power decreases as the values of the explanatory variable increase.

28. Residuals.

a) The curved pattern in the residuals plot indicates that the linear model is not appropriate. The relationship is not linear.

b) The fanned pattern indicates uneven spread. The models predicting power increases as the value of the explanatory variable increases.

c) The scattered residuals plot indicates an appropriate linear model.

29. Real estate.

a) The explanatory variable (x) is size, measured in square feet, and the response variable (y) is price measured in thousands of dollars.

b) The units of the slope are thousands of dollars per square foot.

c) The slope of the regression line predicting price from size should be positive. Bigger homes are expected to cost more.

30. Coasters 2015, revisited.

a) The explanatory variable (x) is initial drop, measured in feet, and the response variable (y) is duration, measured in seconds.

b) The units of the slope are seconds per foot.

c) The slope of the regression line predicting duration from initial drop should be positive. Coasters with higher initial drops probably provide longer rides.

31. What slope?

The only slope that makes sense is 300 pounds per foot. 30 pounds per foot is too small. For example, a Honda Civic is about 14 feet long, and a Cadillac DeVille is about 17 feet long. If the slope of the regression line were 30 pounds per foot, the Cadillac would be predicted to outweigh the Civic by only 90 pounds! (The real difference is about 1500 pounds.) Similarly, 3 pounds per foot is too small. A slope of 3000 pounds per foot would predict a weight difference of 9000 pounds (4.5 tons) between Civic and DeVille. The only answer that is even reasonable is 300 pounds per foot, which predicts a difference of 900 pounds. This isn't very close to the actual difference of 1500 pounds, but at least it is in the right ballpark.

32. What slope again?

The only slope that makes sense is 1 foot in height per inch in circumference. 0.1 feet per inch is too small. A trunk would have to increase in circumference by 10 inches for every foot in height. If that were true, pine trees would be all trunk! 10 feet per inch (and, similarly 100 feet per inch) is too large. If pine trees reach a maximum height of 60 feet, for instance, then the variation in circumference of the trunk would only be 6 inches. Pine tree trunks certainly come in more sizes than that. The only slope that is reasonable is 1 foot in height per inch in circumference.

33. Real estate again.

71.4% of the variability in price can be accounted for by variability in size. (In other words, 71.4% of the variability in price can be accounted for by the linear model.)

34. Coasters 2015 again.

29.4% of the variability in duration can be accounted for by variability in initial drop. (In other words, 29.4% of the variability in duration can be accounted for by the linear model.)

35. Misinterpretations.

a) R^2 is an indication of the strength of the model, not the appropriateness of the model. A scattered residuals plot is the indicator of an appropriate model.

b) Regression models give predictions, not actual values. The student should have said, "The model predicts that a bird 10 inches tall is expected to have a wingspan of 17 inches."

36. More misinterpretations.

a) R^2 measures the amount of variation accounted for by the model. Literacy rate determines 64% of *the variability* in life expectancy.

b) Regression models give predictions, not actual values. The student should have said, "The slope of the line shows that an increase of 5% in literacy rate *is associated with an expected* 2-year improvement in life expectancy."

37. Real estate redux.

a) The correlation between size and price is $r = \sqrt{R^2} = \sqrt{0.714} = 0.845$. The positive value of the square root is used, since the relationship is believed to be positive.

b) The price of a home that is one standard deviation above the mean size would be predicted to be 0.845 standard deviations (in other words r standard deviations) above the mean price.

c) The price of a home that is two standard deviations below the mean size would be predicted to be 1.69 (or 2×0.845) standard deviations below the mean price.

38. Another ride.

a) The correlation between drop and duration is $r = \sqrt{R^2} = \sqrt{0.294} = 0.542$. The positive value of the square root is used, since the relationship is positive.

b) The duration of a coaster whose initial drop is one standard deviation below the mean drop would be predicted to be about 0.542 standard deviations (in other words, r standard deviations) below the mean duration.

c) The duration of a coaster whose initial drop is three standard deviation above the mean drop would be predicted to be about 1.63 (or 3×0.542) standard deviations above the mean duration.

39. ESP.

a) First, since no one has ESP, you must have scored 2 standard deviations above the mean by chance. On your next attempt, you are unlikely to duplicate the extraordinary event of scoring 2 standard deviations above the mean. You will likely "regress" towards the mean on your second try, getting a lower score. If you want to impress your friend, don't take the test again. Let your friend think you can read his mind!

b) Your friend doesn't have ESP, either. No one does. Your friend will likely "regress" towards the mean score on his second attempt, meaning his score will probably go up. If the goal is to get a higher score, your friend should try again.

40. *SI* jinx.

Athletes, especially rookies, usually end up on the cover of Sports Illustrated for extraordinary performances. If these performances represent the upper end of the distribution of performance for this athlete, future performance is likely to regress toward the average performance of that athlete. An athlete's average performance usually isn't notable enough to land the cover of SI. Of course, there are always exceptions, like Michael Jordan, Lebron James, Serena Williams, and others.

41. More real estate.

a) According to the linear model, the price of a home is expected to increase \$61 (0.061 thousand dollars) for each additional square-foot in size.

b) $\widehat{Price} = 47.82 + 0.061 Size = 47.82 + 0.061(3000) = 230.82$; According to the linear model, a 3000 square-foot home is expected to have a price of \$230,820.

c) $\widehat{Price} = 47.82 + 0.061 Size = 47.82 + 0.061(1200) = 121.02$; According to the linear model, a 1200 square-foot home is expected to have a price of \$121,020. The asking price is \$121,020 – \$6000 = \$115,020. \$6000 is the (negative) residual.

42. Last ride.

a) According to the linear model, the duration of a coaster ride is expected to increase by about 0.389 seconds for each additional foot of initial drop.

b) $\widehat{Duration} = 87.22 + 0.389\, Drop = 87.22 + 0.389(200) = 165.02$; According to the linear model, a coaster with a 200 foot initial drop is expected to last 165.02 seconds.

c) $\widehat{Duration} = 87.22 + 0.389\, Drop = 87.22 + 0.389(150) = 145.57$; According to the linear model, a coaster with a 150 foot initial drop is expected to last 145.57 seconds. The advertised duration is longer, at 150 seconds. 150 seconds – 145.57 seconds = 4.43 seconds, a positive residual.

43. Cigarettes.

a) A linear model is probably appropriate. The residuals plot shows residuals that are larger than others, but there is no clear curvature.

b) 81.4% of the variability in nicotine level is accounted for by variability in tar content. (In other words, 81.4% of the variability in nicotine level is accounted for by the linear model.)

44. Attendance 2016, revisited.

a) The linear model is appropriate. Although the relationship is not strong, it is reasonably straight, and the residuals plot shows no pattern.

b) 21.0% of the variability in attendance is accounted for by variability in the number of runs. (In other words, 21.0% of the variability is accounted for by the model.)

c) There is no clear pattern in the residuals.

d) The Dodgers attendance was about 15,000 fans more than we might expect given the number of runs. This is a positive residual.

45. Another cigarette.

a) The correlation between tar and nicotine is $r = \sqrt{R^2} = \sqrt{0.814} = 0.902$. The positive value of the square root is used, since the relationship is believed to be positive. Evidence of the positive relationship is the positive coefficient of tar in the regression output.

b) The average nicotine content of cigarettes that are two standard deviations below the mean in tar content would be expected to be about 1.804 (2×0.902) standard deviations below the mean nicotine content.

c) Cigarettes that are one standard deviation above average in nicotine content are expected to be about 0.902 standard deviations (in other words, *r* standard deviations) above the mean tar content.

46. Attendance 2016, revisited.

a) The correlation between attendance and number of runs is $r = \sqrt{R^2} = \sqrt{0.210} = 0.458$. The positive value of the square root is used, since the relationship is positive.

b) A team that is two standard deviations above the mean in number of runs would be expected to have attendance that is 0.916 (or 2×0.458) standard deviations above the mean attendance.

c) A team that is one standard deviation below the mean in attendance would be expected to have a number of runs that is 0.458 standard deviations (in other words, *r* standard deviations) below the mean number of runs. The correlation between two variables is the same, regardless of the direction in which predictions are made. Be careful, though, since the same is NOT true for predictions made using the slope of the regression equation. Slopes are valid only for predictions in the direction for which they were intended.

47. Last cigarette.

a) $\widehat{Nicotine} = 0.148305 + 0.062163\,Tar$ is the equation of the regression line that predicts nicotine content from tar content of cigarettes.

b) $\widehat{Nicotine} = 0.148305 + 0.062163\,Tar = 0.148305 + 0.062163(4) = 0.397$; The model predicts that a cigarette with 4 mg of tar will have about 0.397 mg of nicotine.

c) For each additional mg of tar, the model predicts an increase of 0.062 mg of nicotine.

d) The model predicts that a cigarette with no tar would have 0.1483 mg of nicotine.

e) $\widehat{Nicotine} = 0.148305 + 0.062163\,Tar = 0.148305 + 0.062163(7) = 0.583$; The model predicts that a cigarette with 7 mg of tar will have 0.583 mg of nicotine. If the residual is –0.05, the cigarette actually had 0.533 mg of nicotine.

48. Attendance 2016, last inning.

a) $\widehat{Attendance} = -12882.2 + 59.389\,Runs$ is the equation of the regression line that predicts attendance from the number of runs scored by American League baseball teams.

b) $\widehat{Attendance} = -12882.2 + 59.389\,Runs = -12882.2 + 59.389(750) = 31659.55$; The model predicts that a team with 750 runs will have attendance of approximately 31,660 people.

c) For each additional run scored, the model predicts an average increase in attendance of 59.389 people.

d) A negative residual means that the team's actual attendance is lower than the attendance model predicts for a team with as many runs.

49. Income and housing revisited.

a) Yes. Both housing cost index and median family income are quantitative. The scatterplot is Straight Enough, although there may be a few outliers. The spread increases a bit for states with large median incomes, but we can still fit a regression line.

b) Using the summary statistics given in the problem, calculate the slope and intercept.

$$b_1 = r\frac{s_{HCI}}{s_{MFI}}$$
$$b_1 = (0.624)\frac{119.07}{7003.55}$$
$$b_1 = 0.0106$$

$$\hat{y} = b_0 + b_1 x$$
$$\overline{y} = b_0 + b_1\overline{x}$$
$$342.3 = b_0 + 0.0106(46210)$$
$$b_0 = -147.526$$

The regression equation that predicts HCI from MFI is $\widehat{HCI} = -147.526 + 0.0106MFI$. (From technology $\widehat{HCI} = -148.15 + 0.0106MFI$.)

c) $\widehat{HCI} = -147.526 + 0.0106MFI = -147.526 + 0.0106(44993) = 329.40$; The model predicts that a state with median family income of \$44993 have an average housing cost index of 329.40.

d) The prediction is 218.62 too low. Washington has a positive residual.

e) The correlation is the slope of the regression line that relates *z*-scores, so the regression equation would be $\widehat{z_{HCI}} = 0.624z_{MFI}$.

f) The correlation is the slope of the regression line that relates *z*-scores, so the regression equation would be $\widehat{z_{MFI}} = 0.624z_{HCI}$.

50. Interest rates and mortgages 2015 again.

a) Both interest rate and total mortgages are quantitative, but the scatterplot is not straight. We should probably not use a linear model.

b) Using the summary statistics given in the problem, calculate the slope and intercept.

$$b_1 = r\frac{s_{MortAmt}}{s_{IntRate}}$$
$$b_1 = (-0.845)\frac{4.527}{2.139}$$
$$b_1 = -1.788$$

$$\hat{y} = b_0 + b_1 x$$
$$\overline{y} = b_0 + b_1\overline{x}$$
$$8.207 = b_0 - 1.788(6.989)$$
$$b_0 = 20.703$$

The regression equation that predicts total mortgage amount from interest rate is $\widehat{MortAmt} = 20.703 - 1.788IntRate$. (From technology $\widehat{MortAmt} = 20.71 - 1.789\,IntRate$.)

c) $\widehat{MortAmt} = 20.703 - 1.788IntRate = 20.703 - 1.788(13) = -2.54$; If interest rates were 13%, the model predicts –\$2.54 trillion in total mortgages.

d) A negative mortgage amount makes no sense. We should be very cautious in making a prediction about an interest rate of 13%. It is well outside the range of our original *x*-variable, and care should always be taken when extrapolating. This prediction is not be appropriate.

e) The correlation is the slope of the regression line that relates *z*-scores, so the regression equation would be $\widehat{z_{MortAmt}} = -0.845z_{IntRate}$.

f) The correlation is the slope of the regression line that relates *z*-scores, so the regression equation would be $\widehat{z_{IntRate}} = -0.845z_{MortAmt}$.

51. Online clothes.

a) Using the summary statistics given in the problem, calculate the slope and intercept.

$$b_1 = r\frac{s_{Total}}{s_{Age}}$$
$$b_1 = (0.037)\frac{253.62}{8.51}$$
$$b_1 = 1.1027$$

$$\hat{y} = b_0 + b_1 x$$
$$\overline{y} = b_0 + b_1\overline{x}$$
$$572.52 = b_0 + 1.1027(29.67)$$
$$b_0 = 539.803$$

The regression equation that predicts total online clothing purchase amount from age is $\widehat{Total} = 539.803 + 1.103\,Age$

b) Yes. Both total purchases and age are quantitative variables, and the scatterplot is Straight Enough, even though it is quite flat. There are no outliers and the plot does not spread throughout the plot.

c) $\widehat{Total} = 539.803 + 1.103\,Age = 539.803 + 1.103(18) = 559.66$; The model predicts that an 18 year old will have \$559.66 in total yearly online clothing purchases.

$\widehat{Total} = 539.803 + 1.103\,Age = 539.803 + 1.103(50) = 594.95$; The model predicts that a 50 year old will have \$594.95 in total yearly online clothing purchases.

d) $R^2 = (0.037)^2 \approx 0.0014 = 0.14\%$.

e) This model would not be useful to the company. The scatterplot is nearly flat. The model accounts for almost none of the variability in total yearly purchases.

52. Online clothes II.

a) Using the summary statistics given, calculate the slope and intercept.

$$b_1 = r\frac{s_{Total}}{s_{Income}}$$
$$b_1 = (0.722)\frac{253.62}{16952.50}$$
$$b_1 = 0.01080157$$

$$\hat{y} = b_0 + b_1 x$$
$$\overline{y} = b_0 + b_1\overline{x}$$
$$572.52 = b_0 + 0.01080157(50343.40)$$
$$b_0 = 28.73$$

The regression equation that predicts total online clothing purchase amount from income is $\widehat{Total} = 28.73 + 0.0108\,Income$. (Since the mean income is a relatively large number, the value of the intercept will vary, based on the rounding of the slope. Notice that it is very close to zero in the context of yearly income.)

b) The assumptions for regression are met. Both variables are quantitative and the plot is straight enough. There are several possible outliers, but none of these points are extreme, and there are 500 data points to establish a pattern. The spread of the plot does not change throughout the range of income.

c) $\widehat{Total} = 28.73 + 0.0108\,Income = 28.73 + 0.0108(20{,}000) = 244.73$; The model predicts that a person with \$20,000 yearly income will make \$244.73 in online purchases. (Predictions may vary, based on rounding of the model.)

$\widehat{Total} = 28.73 + 0.0108\,Income = 28.73 + 0.0108(80{,}000) = \$892.73$; The model predicts that a person with \$80,000 yearly income will make \$892.73 in online purchases. (Predictions may vary, based on rounding of the model.)

52. (continued)

d) $R^2 = (0.722)^2 \approx 0.521 = 52.1\%$.

e) The model accounts for a 52.1% of the variation in total yearly purchases, so the model would probably be useful to the company. Additionally, the difference between the predicted purchases of a person with \$20,000 yearly income and \$80,000 yearly income is of practical significance.

53. SAT scores.

a) The association between SAT Math scores and SAT Verbal Scores was linear, moderate in strength, and positive. Students with high SAT Math scores typically had high SAT Verbal scores.

b) One student got a 500 Verbal and 800 Math. That set of scores doesn't seem to fit the pattern.

c) $r = 0.685$ indicates a moderate, positive association between SAT Math and SAT Verbal, but only because the scatterplot shows a linear relationship. Students who scored one standard deviation above the mean in SAT Math were expected to score 0.685 standard deviations above the mean in SAT Verbal. Additionally, $R^2 = (0.685)^2 = 0.469225$, so 46.9% of the variability in math score was accounted for by variability in verbal score.

d) The scatterplot of verbal and math scores shows a relationship that is straight enough, so a linear model is appropriate.

$$b_1 = r\frac{s_{Math}}{s_{Verbal}}$$
$$b_1 = (0.685)\frac{96.1}{99.5}$$
$$b_1 = 0.661593$$

$$\hat{y} = b_0 + b_1 x$$
$$\overline{y} = b_0 + b_1\overline{x}$$
$$612.2 = b_0 + 0.661593(596.3)$$
$$b_0 = 217.692$$

The equation of the least squares regression line for predicting SAT Math score from SAT Verbal score is $\widehat{Math} = 217.692 + 0.662\,Verbal$.

e) For each additional point in verbal score, the model predicts an increase of 0.662 points in math score. A more meaningful interpretation might be scaled up. For each additional 10 points in verbal score, the model predicts an increase of 6.62 points in math score.

f) $\widehat{Math} = 217.692 + 0.662\,Verbal = 217.692 + 0.662(500) = 548.692$; According to the model, a student with a verbal score of 500 was expected to have a math score of 548.692.

g) $\widehat{Math} = 217.692 + 0.662\,Verbal = 217.692 + 0.662(800) = 747.292$; According to the model, a student with a verbal score of 800 was expected to have a math score of 747.292. She actually scored 800 on math, so her residual was 800 – 747.292 = 52.708 points

54. Success in college

a) A scatterplot showed the relationship between combined SAT score and GPA to be reasonably linear, so a linear model is appropriate.

$$b_1 = r\frac{s_{GPA}}{s_{SAT}}$$
$$b_1 = (0.47)\frac{0.56}{123}$$
$$b_1 \approx 0.0021398$$

$$\hat{y} = b_0 + b_1 x$$
$$\overline{y} = b_0 + b_1\overline{x}$$
$$2.66 = b_0 + 0.0021398(1222)$$
$$b_0 \approx 0.045$$

The regression equation predicting GPA from SAT score is: $\widehat{GPA} = 0.045 + 0.00214\,SAT$

54. (continued)

b) The model predicts that a student with an SAT score of 0 would have a GPA of 0.045. The *y*-intercept is not meaningful, since an SAT score of 0 is impossible.

c) The model predicts that students who scored 100 points higher on the SAT tended to have a GPA that was 0.2140 higher.

d) $\widehat{GPA} = 0.045 + 0.00214\,SAT = 0.045 + 0.00214(1400) \approx 3.04$; According to the model, a student with an SAT score of 1400 is expected to have a GPA of 3.04.

e) According to the model, SAT score is not a very good predictor of college GPA. $R^2 = (0.47)^2 = 0.2209$, which means that only 22.1% of the variability in GPA can be accounted for by the model. The rest of the variability is determined by other factors.

f) A student would prefer to have a positive residual. A positive residual means that the student's actual GPA is higher than the model predicts for someone with the same SAT score.

55. SAT, take 2.

a) $r = 0.685$; The correlation between SAT Math and SAT Verbal is a unitless measure of the degree of linear association between the two variables. It doesn't depend on the order in which you are making predictions.

b) The scatterplot of verbal and math scores shows a relationship that is straight enough, so a linear model is appropriate.

$$b_1 = r\frac{s_{Verbal}}{s_{Math}}$$
$$b_1 = (0.685)\frac{99.5}{96.1}$$
$$b_1 = 0.709235$$

$$\hat{y} = b_0 + b_1 x$$
$$\overline{y} = b_0 + b_1 \overline{x}$$
$$596.3 = b_0 + 0.709235(612.2)$$
$$b_0 = 162.106$$

The equation of the least squares regression line for predicting SAT Verbal score from SAT Math score is: $\widehat{Verbal} = 162.106 + 0.709\,Math$

c) A positive residual means that the student's actual verbal score was higher than the score the model predicted for someone with the same math score.

d) $\widehat{Verbal} = 162.106 + 0.709\,Math = 162.106 + 0.709(500) = 516.606$; According to the model, a person with a math score of 500 was expected to have a verbal score of 516.606 points.

e) $\widehat{Math} = 217.692 + 0.662\,Verbal = 217.692 + 0.662(516.606) = 559.685$; According to the model, a person with a verbal score of 516.606 was expected to have a math score of 559.685 points.

f) The prediction in part (e) does not cycle back to 500 points because the regression equation used to predict math from verbal is a different equation than the regression equation used to predict verbal from math. One was generated by minimizing squared residuals in the verbal direction, the other was generated by minimizing squared residuals in the math direction. If a math score is one standard deviation above the mean, its predicted verbal score regresses toward the mean. The same is true for a verbal score used to predict a math score.

56. Success, part 2.

Using the summary statistics given, calculate the slope and intercept.

$$b_1 = r\frac{s_{SAT}}{s_{GPA}}$$
$$b_1 = (0.47)\frac{123}{0.56}$$
$$b_1 = 103.232$$

$$\hat{y} = b_0 + b_1 x$$
$$\overline{y} = b_0 + b_1 \overline{x}$$
$$1222 = b_0 + 103.232(2.66)$$
$$b_0 = 947.403$$

The regression equation to predict SAT score from GPA is: $\widehat{SAT} = 947.403 + 103.232GPA$.

$\widehat{SAT} = 947.403 + 103.232(3) = 1257.1$; The model predicts that a student with a GPA of 3.0 is expected to have an SAT score of 1257.1.

57. Wildfires 2015.

a) The scatterplot shows a roughly linear relationship between the year and the number of wildfires, so the linear model is appropriate. The relationship is very weak, however.

b) The model predicts a decrease of an average of about 222 wildfires per year.

c) It seems reasonable to interpret the intercept. The model predicts about 78,792 wildfires in 1985, which is within the scope of the data, although it isn't very useful since we know the actual number of wildfires in 1985. There isn't much need for a prediction.

d) The standard deviation of the residuals is 12,397 fires. That's a large residual, considering that these years show between 60,000 and 90,000 fires per year. The association just isn't very strong.

e) The model only accounts for about 2.7% of the variability in the number of fires each year. The rest of the variability is due to other factors that we don't have data about.

58. Wildfires 2015 – sizes.

a) The scatterplot is curved, but since 1990, the relationship is fairly straight. We should proceed with caution.

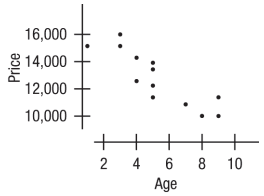
b) The number of Acres per fire has increased, on average, about 274,000 acres per fire per year since 1985.

c) The model estimates that fires had an average size of 21.7 acres in 1985.

d) The model accounts for 45.45% of the variation in the number of acres per fire. This indicates a moderately weak model.

59. Used cars 2014.

a) We are attempting to predict the price in dollars of used Toyota Corollas from their age in years. A scatterplot of the relationship is shown below.



b) There is a strong, negative, linear association between price and age of used Toyota Corollas.

c) The scatterplot provides evidence that the relationship is straight enough. A linear model will likely be an appropriate model.

d) Since $R^2 = 0.891$, simply take the square root to find $r = \sqrt{0.752} = 0.867$. Since association between age and price is negative, $r = -0.867$.

59. (continued)

e) 75.2% of the variability in price of a used Toyota Corolla can be accounted for by variability in the age of the car.

f) The relationship is not perfect. Other factors, such as options, condition, and mileage explain the rest of the variability in price.

60. Drug abuse revisited.

a) The scatterplot shows a positive, strong, linear relationship. It is straight enough to make the linear model the appropriate model.

b) 87.3% of the variability in percentage of other drug usage can be accounted for by percentage of marijuana use.

c) $R^2 = 0.873$, so $r = \sqrt{0.873} = 0.93434$ (since the relationship is positive).

$$b_1 = r\frac{s_O}{s_M}$$
$$b_1 = (0.93434)\frac{10.2}{15.6}$$
$$b_1 = 0.61091$$

$$\hat{y} = b_0 + b_1 x$$
$$\overline{y} = b_0 + b_1 \overline{x}$$
$$11.6 = b_0 + 0.61091(23.9)$$
$$b_0 = -3.001$$

The regression equation used to predict the percentage of teens that use other drugs from the percentage who have used marijuana is: $\widehat{Other} = -3.001 + 0.611\,Marijuana$. (Using technology, the model is $\widehat{Other} = -3.068 + 0.615\,Marijuana$.)

d) According to the model, each additional percent of teens using marijuana is expected to add 0.611 percent to the percentage of teens using other drugs.

e) The results do not *confirm* marijuana as a gateway drug. They do indicate an *association* between marijuana and other drug usage, but association does not imply causation.

61. More used cars 2014.

a) The scatterplot from the Exercise 59 shows that the relationship is straight, so the linear model is appropriate. The regression equation to predict the price of a used Toyota Corolla from its age is $\widehat{Price} = 17674 - 844.5\,Years$. The computer regression output used is shown below.

```
Predictor      Coef  SE Coef      T      P
Constant   17674.0    836.2  21.14  0.000
Age         -844.5    146.1  -5.78  0.000


S = 1224.82   R-Sq = 75.2%   R-Sq(adj) = 73.0%
```

b) According to the model, for each additional year in age, the car is expected to drop $844.5 in price.

c) The model predicts that a new Toyota Corolla (0 years old) will cost $17,674.

d) $\widehat{Price} = 17674 - 844.5\,Years = 17674 - 844.5(7) = 11762.5$; According to the model, an appropriate price for a 7-year old Toyota Corolla is $11,762.50.

e) Buy the car with the negative residual. Its actual price is lower than predicted.

f) $\widehat{Price} = 17674 - 844.5 Years = 17674 - 844.5(10) = 9229$; According to the model, a 10-year-old Corolla is expected to cost $9229. The car has an actual price of $8500, so its residual is $8500 – $9229 = –$729

g) The model would not be useful for predicting the price of a 25-year-old Corolla. The oldest car in the list is 9 years old. Predicting a price after 25 years would be an extrapolation.

62. Veggie burgers 2014.

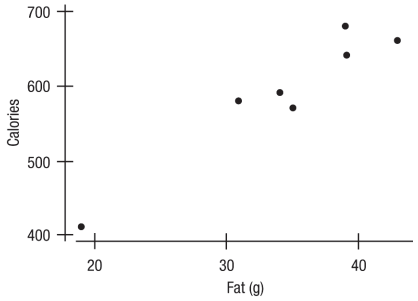
a) $\widehat{Fat} = 8.4 + 0.91\,Protein = 8.4 + 0.91(21) = 27.51$; According to the model, a burger with 21 grams of protein is expected to have 27.51 grams of fat.

b) From the package, the actual fat content of the veggie burger is 8 grams. The residual is 8 – 27.51 = –19.51 grams of fat. The veggie burgers have about 19.51 fewer grams of fat than is predicted by the model for a regular burger with a similar protein content.

c) The new veggie burger has 21 grams of protein and 8 grams of fat. The veggie burger has about 19.5 fewer grams of fat than a typical BK menu item with a similar protein content.

63. Burgers revisited.

a) The scatterplot of calories vs. fat content in fast food hamburgers is shown below. The relationship appears linear, so a linear model is appropriate.



Dependent variable is: **Calories**
No Selector
R squared = 92.3% R squared (adjusted) = 90.7%
s = 27.33 with 7 - 2 = 5 degrees of freedom

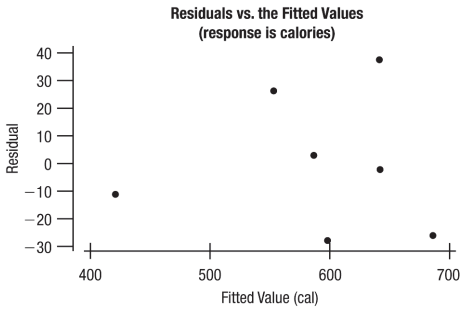
Source	Sum of Squares	df	Mean Square	F-ratio
Regression	44664.3	1	44664.3	59.8
Residual	3735.73	5	747.146	

Variable	Coefficient	s.e. of Coeff	t-ratio	prob
Constant	210.954	50.10	4.21	0.0084
Fat	11.0555	1.430	7.73	0.0006

b) From the computer regression output, $R^2 = 92.3\%$. 92.3% of the variability in the number of calories can be explained by the variability in the number of grams of fat in a fast food burger.

c) From the computer regression output, the regression equation that predicts the number of calories in a fast food burger from its fat content is $\widehat{Calories} = 210.954 + 11.0555\,Fat$.

d) The residuals plot shown below shows no pattern. The linear model appears to be appropriate.



63. (continued)

e) The model predicts that a fat free burger would have about 211 calories. Since there are no data values close to 0, this extrapolation isn't of much use.

f) For each additional gram of fat in a burger, the model predicts an increase of 11.06 calories.

g) $\widehat{Calories} = 210.954 + 11.0555\,Fat = 210.954 + 11.0555(28) = 520.508$; The model predicts a burger with 28 grams of fat will have 520.508 calories. If the residual is +33, the actual number of calories is 520.508 + 33 $\approx$ 553.5 calories.

64. Chicken.

a) The scatterplot is fairly straight, so the linear model is appropriate.

b) The correlation of 0.947 indicates a strong, linear, positive relationship between fat and calories for chicken sandwiches.

c) Using the summary statistics given, calculate the slope and intercept.

$$b_1 = r\frac{s_{Cal}}{s_{Fat}}$$

$$b_1 = (0.947)\frac{144.2}{9.8}$$

$$b_1 = 13.934429$$

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

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

$$472.7 = b_0 + 13.934429(20.6)$$

$$b_0 = 185.651$$

The linear model for predicting calories from fat in chicken sandwiches is $\widehat{Calories} = 185.651 + 13.934\,Fat$.

d) For each additional gram of fat, the model predicts an increase in 13.934 calories.

e) According to the model, a fat-free chicken sandwich would have 185.651 calories. This is probably an extrapolation, although without the actual data, we can't be sure.

f) In this context, a negative residual means that a chicken sandwich has fewer calories than the model predicts.

g) For the chicken sandwich: $\widehat{Calories} = 185.651 + 13.934\,Fat = 185.651 + 13.934(35) = 673.341$

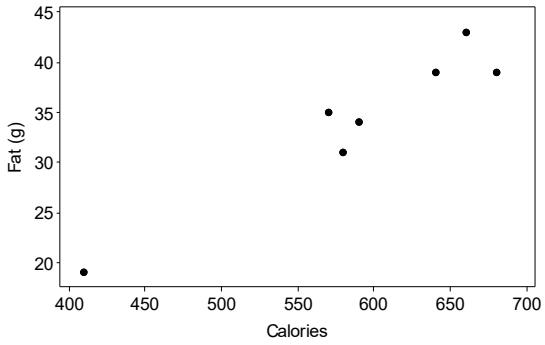
For the burger: $\widehat{Calories} = 210.954 + 11.056\,Fat = 210.954 + 11.056(35) = 597.914$

A chicken sandwich with 35 grams of fat is predicted to have more calories than a burger with 35 grams of fat.

65. A second helping of burgers.

a) The model from Exercise 63 was for predicting number of calories from number of grams of fat. In order to predict grams of fat from the number of calories, a new linear model needs to be generated.

b) The scatterplot below shows the relationship between number fat grams and number of calories in a set of fast food burgers. The association is strong, positive, and linear. Burgers with higher numbers of calories typically have higher fat contents. The relationship is straight enough to apply a linear model.



The linear model for predicting fat from calories is: $\widehat{Fat} = -14.9622 + 0.083471 Calories$. The model predicts that for every additional 100 calories, the fat content is expected to increase by about 8.3 grams. The residuals plot shows no pattern, so the model is appropriate. $R^2 = 92.3\%$, so 92.3% of the variability in fat content can be accounted for by the model.

$\widehat{Fat} = -14.9622 + 0.083471 Calories = -14.9622 + 0.083471(600) \approx 35.1$; According to the model, a burger with 600 calories is expected to have 35.1 grams of fat.

66. Cost of living 2016.

a) The association between the cost of a cappuccino and a third of a liter of water is, moderate, positive, and mostly linear, but curving at the highest water prices. The scatterplot is Straight Enough, indicating that the linear model is appropriate.

b) $R^2 = (0.597)^2 = 0.356$; This means that 35.6% of the variability in the cost of a cappuccino can be explained by the price of a third of a liter of water.

c) $\widehat{CappuccinoCost} = 1.636 + 0.9965 WaterCost = 1.636 + 0.9965(2) = 3.629$; According to the model, Christchurch is predicted to have a water cost of \$3.63. Christchurch actually had a water cost of \$3.37. Christchurch's residual was about –\$0.26.

67. New York bridges 2016.

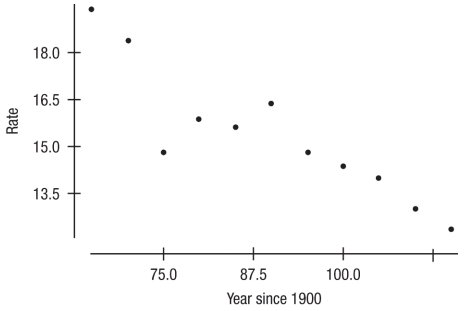
a) Overall, the model predicts the condition score of new bridges to be 5.0112, close to the cutoff of 5. The negative slope means that most bridges are predicted to have condition less than 5. The model is not a very encouraging one in regards to the conditions of New York City bridges.

b) According to the model, the condition of the bridges in New York City is decreasing by an average of 0.00513 per year. This is less rapid than the bridges in Tompkins County.

c) We shouldn't place too much faith in the model. R^2 of 3.9% is very low, and the standard deviation of the residuals, 0.6912, is quite high in relation to the scope of the data values themselves. This association is very weak.

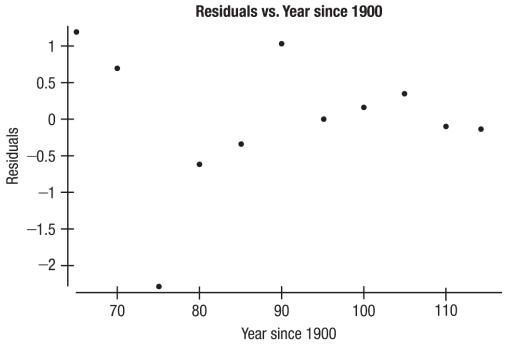
68. Birthrates 2015.

a) A scatterplot of US birthrates over time is shown below. The association is negative, strong, and reasonably linear, with one low outlier, the rate of 14.8 live births per 1000 women age 15 – 44 in 1975. Generally, as time passes, the birth rate is getting lower.



b) Although the association appears curved, it is straight enough to try a linear model. The equation of the linear model is $\widehat{Birthrate} = 25.649 - 0.1142(Year - 1900)$.

c) The scatterplot of the residuals vs. the predicted values is shown below. There appears to be some curvature to the plot, indicating that a linear model may not be appropriate. With so few data points, it's hard to differentiate between pattern and random fluctuation. Additionally, the scatterplot shows a low outlier for the year 1975. We may want to investigate further. At the very least, be cautious when using this model.



d) The model predicts that each passing year is associated with a decline in birth rate of 0.1142 births per 1000 women.

e) $\widehat{Birthrate} = 25.649 - 0.1142(Year - 1900) = 25.649 - 0.1142(1978 - 1900) = 16.7414$; The model predicts about 16.7414 births per 1000 women in 1978.

f) If the actual birth rate in 1978 was 15.0 births per 1000 women, the model has a residual of 15.0 – 16.7414 = –1.7414 births per 1000 women. This means that the model predicted 1.74 births higher than the actual rate.

g) $\widehat{Birthrate} = 25.649 - 0.1142(Year - 1900) = 25.649 - 0.1142(2020 - 1900) = 11.945$; According to the model, the birth rate in 2020 is predicted to be 11.945 births per 1000 women. This prediction seems a bit low. It is an extrapolation outside the range of the data, and furthermore, the model only explains 80.17% of the variability in birth rate. Don't place too much faith in this prediction.

h) $\widehat{Birthrate} = 25.649 - 0.1142(Year - 1900) = 25.649 - 0.1142(2050 - 1900) = 8.519$; According to the model, the birth rate in 2050 is predicted to be 8.519 births per 1000 women. This prediction is an extreme extrapolation outside the range of the data, which is dangerous. No faith should be placed in this prediction.

69. Climate change 2016.

a) The correlation between CO_2 level and mean global temperature anomaly is $r = \sqrt{R^2} = \sqrt{0.897} = 0.9471$.

b) 89.7% of the variability in mean global temperature anomaly can be accounted for by variability in CO_2 level.

c) Since the scatterplot of CO_2 level and mean global temperature anomaly shows a relationship that is straight enough, use of the linear model is appropriate. The linear regression model that predicts mean global temperature anomaly from CO_2 level is: $\widehat{TempAnomaly} = -3.17933 + 0.0099179\,CO_2$.

d) The model predicts that an increase in CO_2 level of 1 ppm is associated with an increase of 0.0099°C in mean global temperature anomaly.

e) According to the model, the mean global temperature anomaly is predicted to be –3.179 °C when there is no CO_2 in the atmosphere. This is an extrapolation outside of the range of data, and isn't very meaningful in context, since there is always CO_2 in the atmosphere. We want to use this model to study the change in CO_2 level and how it relates to the change in temperature.

f) The residuals plot shows no apparent patterns. The linear model appears to be an appropriate one.

g) $\widehat{TempAnomaly} = -3.17933 + 0.0099179\,CO_2 = -3.17933 + 0.0099179(450) = 1.283725$; According to the model, the mean global temperature anomaly is predicted to be 1.28 °C when the CO_2 level is 450 ppm.

h) No, this does not mean that when the CO_2 level hits 450 ppm, the temperature anomaly will be 1.28 °C. First, this is an extrapolation. The model can't say what will happen under other circumstances. Secondly, even for predictions within our set of data, the model can only give a general idea of what to expect.

70. Climate change 2016, revisited.

a) The correlation between Dow Jones Industrial Average and mean global temperature anomaly is $r = \sqrt{R^2} = \sqrt{0.802} = 0.8955$.

b) 80.2% of the variability in mean global temperature anomaly can be accounted for by variability in DJIA.

c) The scatterplot of DJIA and mean global temperature anomaly shows a violation of the Straight Enough condition, so the use of the linear model is not appropriate. We will answer the rest of the questions anyway. The linear regression model that predicts mean global temperature anomaly from DJIA is $\widehat{TempAnomaly} = 0.084216 + 0.0000444440\,DJIA$.

d) The model predicts that an increase in DJIA of 1000 points is associated with an increase of 0.044 °C in mean global temperature anomaly.

e) According to the model, the mean global temperature anomaly is predicted to be 0.084216 °C when there is a DJIA of zero. This is an extrapolation outside of the range of data, and isn't very meaningful in context, since the DJIA has never been zero. We want to use this model to study the change DJIA and how it relates to the change in temperature, and we can simply think of the intercept as the starting point of our model.

f) The residuals plot shows a clear curved pattern. The linear model is not appropriate.

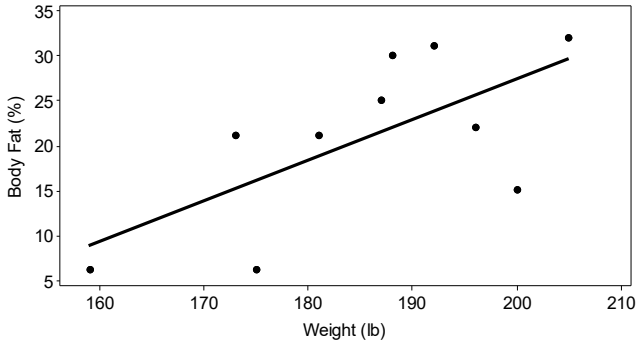
g) $\widehat{TempAnomaly} = 0.084216 + 0.0000444440\,DJIA = 0.084216 + 0.0000444440(25{,}000) = 1.195316$; According to the model, the mean global temperature anomaly is predicted to be 1.195 °C when the DJIA is 25,000.

70. (continued)

h) No, this does not mean that when the DJIA hits 25,000, the temperature anomaly will be 1.195 °C. First, this is a large extrapolation, but more important, this model does not show that the DJIA affects the temperature. It simply says that we can observe an association between the variables. It is likely that the association is due to a third, lurking, variable. For example, both the temperature and the DJIA have been increasing over time, making time a lurking variable here. Finally, the linear model we made isn't even justified. The scatterplot was curved, and so was the residuals plot. We can place no faith at all in this prediction.

71. Body fat.

a) The scatterplot of % body fat and weight of 20 male subjects, shown below, shows a strong, positive, linear association. Generally, as a subject's weight increases, so does % body fat. The association is straight enough to justify the use of the linear model.



The linear model that predicts % body fat from weight is $\widehat{\%Fat} = -27.4 + 0.25\,Weight$.

b) The residuals plot shows no apparent pattern. The linear model is appropriate.

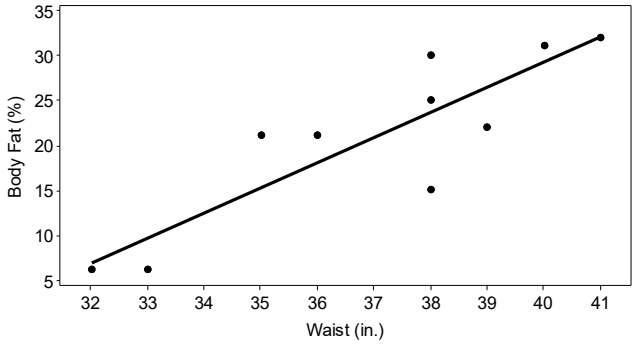
c) According to the model, for each additional pound of weight, body fat is expected to increase by about 0.25%.

d) Only 48.5% of the variability in % body fat can be accounted for by the model. The model is not expected to make predictions that are accurate.

e) $\widehat{\%Fat} = -27.3763 + 0.249874\,Weight = -27.3763 + 0.249874(190) = 20.09976$; According to the model, the predicted body fat for a 190-pound man is 20.09976%. The residual is 21 – 20.09976 ≈ 0.9%.

72. Body fat, again.

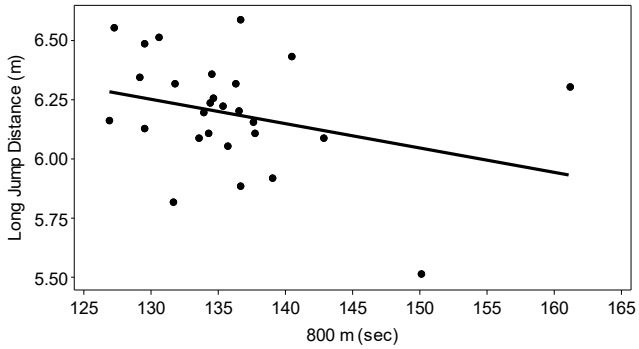
The scatterplot of percent body fat and waist size is shown below. The association is strong, linear, and positive. As waist size increases, percent body fat has a tendency to increase, as well. The scatterplot is straight enough to justify the use of the linear model.



The linear model for predicting % body fat from waist size is $\widehat{\%Fat} = -62.6 + 2.22\,Waist$. For each additional inch in waist size, the model predicts an increase of 2.22% body fat. 78.7% of the variability in % body fat can be accounted for by waist size. A residuals plot shows no apparent pattern. The residuals plot and the relatively high value of R^2 indicate an appropriate model with more predicting power than the model based on weight.

73. Women's heptathlon revisited.

a) Both long jump height and 800-meter time are quantitative variables, the association is straight enough to use linear regression, though there are two outliers that may be influential points.



The regression equation to predict long jump from 800m results is: $\widehat{Longjump} = 7.595 - 0.01\,Time$. According to the model, the predicted long jump decreases by an average of 0.010 meters for each additional second in 800-meter time.

b) $R^2 = 9.39\%$; This means that 9.39% of the variability in long jump distance is accounted for by the variability in 800-meter time.

c) Yes, good long jumpers tend to be fast runners. The slope of the association is negative. Faster runners tend to jump long distances, as well.

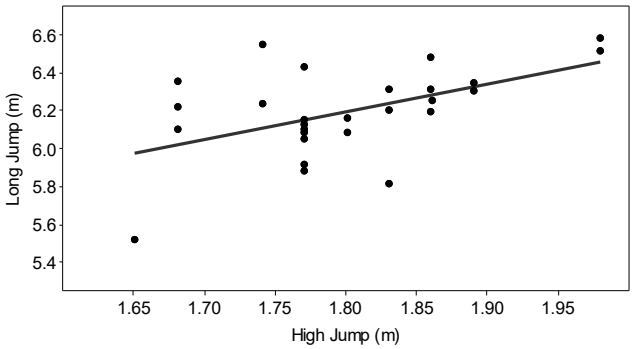
d) A residuals plot shows two outliers. One of these athletes, Akela Jones, had a long jump that was much longer than her 800m time would predict, making this point particularly influential. This model is not appropriate.

73. (continued)

e) The linear model is not particularly useful for predicting long jump performance. First of all, there are outliers and influential points. Additionally, only 9.39% of the variability in long jump distance is accounted for by the variability in 800-meter time, leaving 90.61% of the variability accounted for by other variables. Finally, the residual standard deviation is 0.23 meters, which is not much smaller than the standard deviation of all long jumps, 0.2475 meters. Predictions are not likely to be accurate.

74. Heptathlon revisited again.

a) Both high jump height and long jump distance are quantitative variables, the association is straight enough, and there are no outliers. It is appropriate to use linear regression.



The regression equation to predict long jump from high jump results is $\widehat{Longjump} = 3.403 + 1.538 Highjump$. According to the model, the predicted long jump increases by an average of 1.538 meters for each additional meter in high jump height.

b) $R^2 = 24.7\%$; This means that only 24.7% of the variability in long jump distance is accounted for by the variability in high jump height.

c) Yes, good high jumpers tend to be good long jumpers. The slope of the association is positive. Better high jumpers tend to be better long jumpers, as well.

d) A residuals plot, and the scatterplot itself, shows one jumper whose jumps were both quite short compared to the performance of other jumpers. This point is influential in the regression.

e) The linear model is not particularly useful for predicting long jump performance. First of all, only 24.7% of the variability in long jump distance is accounted for by the variability in high jump height, leaving 75.3% of the variability accounted for by other variables. Secondly, the residual standard deviation is 0.219 meters, which is not much smaller than the standard deviation of all long jumps, 0.247 meters. Finally, there is an influential point in the regression. Predictions are not likely to be accurate.

75. Hard water.

a) There is a fairly strong, negative, linear relationship between calcium concentration (in ppm) in the water and mortality rate (in deaths per 100,000). Towns with higher calcium concentrations tended to have lower mortality rates.

b) The linear regression model that predicts mortality rate from calcium concentration is $\widehat{Mortality} = 1676 - 3.23 Calcium$.

c) The model predicts a decrease of 3.23 deaths per 100,000 for each additional ppm of calcium in the water. For towns with no calcium in the water, the model predicts a mortality rate of 1676 deaths per 100,000 people.

75. (continued)

d) Exeter had 348.6 fewer deaths per 100,000 people than the model predicts.

e) $\widehat{Mortality} = 1676 - 3.23\,Calcium = 1676 - 3.23(100) = 1353$; The town of Derby is predicted to have a mortality rate of 1353 deaths per 100,000 people.

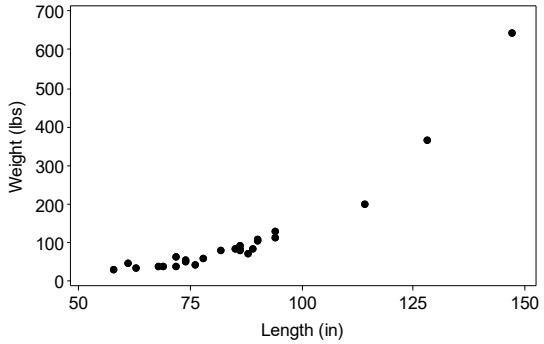
f) 43% of the variability in mortality rate can be explained by variability in calcium concentration.

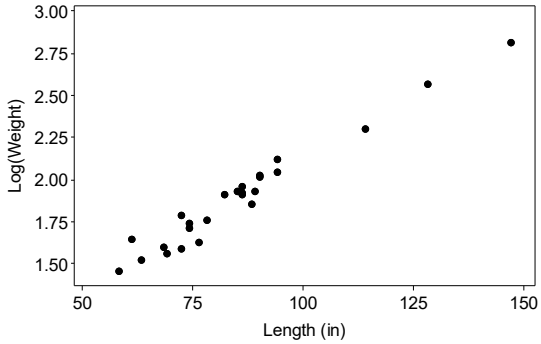
76. Gators.

a) Weight is the proper dependent variable. The researchers can estimate length from the air, and use length to predict weight, as desired.

b) The correlation between an alligator's length and weight is $r = \sqrt{R^2} = \sqrt{0.836} = 0.914$.

c) The scatterplot of alligator weights versus lengths is below to the left. The relationship is curved. The linear model is not appropriate.





d) Re-expressing the weights with logarithms seems to work well. The scatterplot is shown above to the right. The equation of the improved model is $\widehat{Log(Weight)} = 0.58 + 0.015\,Length$.

e) The re-expressed model would work better. Its scatterplot is much straighter, and the residuals plot shows random scatter.

f) The predictions made from the logarithmic model should be reasonably accurate. Since $R^2 = 0.96$, the length of the alligator accounts for 96% of the variation in the log of the weight of the gators. However, the residuals for the longer gators are large. They don't look large on the graph, but remember that this is a logarithmic scale. If we make a prediction for our largest gator, at 147 inches and 640 pounds, we will see that the prediction is off by about 30 pounds, since $\widehat{Log(Weight)} = 0.58 + 0.015(147) = 2.785$, and $\widehat{Weight} = 10^{2.785} \approx 610$.

77. Least squares.

If the 4 x-values are plugged into $\hat{y} = 7 + 1.1x$, the 4 predicted values are $\hat{y}$ = 18, 29, 51 and 62, respectively. The 4 residuals are –8, 21, –31, and 18. The squared residuals are 64, 441, 961, and 324, respectively. The sum of the squared residuals is 1790. Least squares means that no other line has a sum lower than 1790. In other words, it's the best fit.

78. Least squares.

If the 4 x-values are plugged into $\hat{y} = 1975 - 0.45x$, the 4 predicted values are $\hat{y}$ = 1885, 1795, 1705, and 1615, respectively. The 4 residuals are 65, –145, 95, and –15. The squared residuals are 4225, 21025, 9025, and 225, respectively. The sum of the squared residuals is 34,500. Least squares means that no other line has a sum lower than 34,500. In other words, it's the best fit.

79. Fuel Economy 2016 revisited.

a) The distribution of the slopes is unimodal and slightly skewed to the low end. 95% of the random slopes fell between –4.4 and –2.2.

b) The slope for the sample from Chapter 6 is just barely inside the interval that holds the middle 95% of the randomly generated slopes, so by that definition, it is not unusual.

c) Answers will vary.

80. Receivers 2015.

a) The distribution of the slopes is unimodal and symmetric. 95% of the random sample slopes are between 8.79 and 15.67.

b) The slope tells how much is gained by a typical pass reception. It is in units of yards per reception.

c) Somewhere between 8.79 and 15.67 yards per reception

d) Answers will vary.