Chapter 7 – Linear Regression

1. Cereals.

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

2. Horsepower.

$\widehat{mpg} = 46.87 - 0.084\,HP = 46.87 - 0.084(200) \approx 30.07$ mpg. According to the model, we expect a car with 200 horsepower to get about 30.07 miles per gallon.

3. More cereal.

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

4. Horsepower, again.

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

5. Another bowl.

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

6. More horsepower.

The model predicts that cars lose an average of 0.84 miles per gallon for each additional 10 horse power.

7. 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.

8. Another car.

$R^2 = (-0.869)^2 \approx 0.755$. About 75.5% of the variability in fuel economy is accounted for by the model.

9. Last bowl!

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

10. Last tank!

True fuel economy varies from the predicted amount with a standard deviation of 3.287 miles per gallon.

11. Residuals I.

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.

12. Residuals II.

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 heteroscedastic data. The models predicting power increases as the value of the explanatory variable increases.

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

13. What slope I?

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 it is in the right ballpark.

14. What slope II?

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.

15. 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.

16. True or false II.

a) True.

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

c) True.

17. Bookstore sales revisited.

a) $\widehat{Sales} = 8.1006 + 0.9134\,People$, where *Sales* is measured in thousands of dollars, and *People* is the number of sales people working.

b) The slope of the line of best fit, b_1, is 0.9134. The model predicts an increase in sales of 0.9134($1000), or $913.40, for each additional sales person working.

c) The intercept, b_0, is 8.1006. The model predicts that average sales would be approximately $8.1006($1000), or $8100.60, when there were no sales people working. This doesn't make sense in this context.

d) $\widehat{Sales} = 8.1006 + 0.9134\,People = 8.1006 + 0.9134(18) = 24.5418$
According to the model, we would expect sales to be approximately $24,540 when 18 sales people are working.

e) Residual $= sales - \widehat{sales} = \$25,000 - \$24,540 = \$460$

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

18. Disk drives again.

a) $\widehat{Price} = 18.617 + 103.929\,Capacity$, where *Price* is measured in dollars and *Capacity* is measured in terabytes.

b) The slope of the line of best fit, b_1, is 103.929. The model predicts an average increase of approximately $104 for each additional TB of storage.

c) The intercept, b_0, is 18.617. According to the model, the average cost of a drive with 0 TB capacity is expected to be $18.64. This could be thought of as the cost of the cabinet and basic electronics common to all hard drives.

d) $\widehat{Price} = 18.617 + 103.929\,Capacity = 18.617 + 103.929(3.0) = \$330.40$
According to the model, we would expect the price of a 3.0 TB drive to average approximately $330.40. (Answers may vary depending on the values used for the slope and the intercept.)

e) Residual $= Price - \widehat{Price} = 300 - 330.40 = -\$30.40$. This drive is a good buy. It costs $30.37 less than you expected to pay.

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

19. 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. This corresponds to 10 sales people working.

c) The residual with the smallest magnitude, 0.07, contributes least to the sum of the squared residuals. This corresponds to 2 sales people working.

20. Disk drives once more.

a) The drive with a capacity of 1.0 MB, with a residual of –23.55, 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 –23.55 indicates a drive that costs $23.55 less than we might expect.

21. Bookstore sales last time.

$R^2 = 93.2\%$. About 93% of the variability in sales can be accounted for by the regression between sales and the number of sales people working.

22. Disk drives encore.

$R^2 = 98.8\%$. About 98.8 % of the variability in the price of the disk drive can be accounted for by the regression between the price of the drive and the size of the drive.

23. Residual plots.

a) The residuals plot shows a curved pattern, indicating that the linearity assumption has been violated.

b) The residuals plot shows an outlier. The outlier condition has been violated. Note that an outlier may not be visible in a residuals plot if the point is an outlier in the x-direction. The regression line may be dragged toward the point, leaving a small residual for such an outlier.

c) The residuals plot shows a fanning pattern, with larger residuals for larger fitted values. This indicates a violation of the equal spread condition.

24. Disk drives last time.

a) There appears to be a curve in the residuals plot.

b) Since the residuals plot shows a curve, we should not use the linear regression.

25. 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.

26. Roller coaster.

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.

27. 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.)

28. Coasters again.

12.4% of the variability in duration can be accounted for by variability in initial drop. (In other words, 12.4% of the variability in duration can be accounted for by the linear model.)

29. 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.

30. Another ride.

a) The correlation between drop and duration is $r = \sqrt{R^2} = \sqrt{0.124} = 0.352$. The positive value of the square root is used, since the relationship is believed to be positive.

b) The duration of a coaster whose initial drop is one standard deviation below the mean drop would be predicted to be about 0.352 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 1.056 (or 3×0.352) standard deviations above the mean duration.

31. 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) $\widehat{Price} = 47.82 + 0.061\,Size$

$\widehat{Price} = 47.82 + 0.061(3000)$

$\widehat{Price} = 230.82$

According to the linear model, a 3000 square-foot home is expected to have a price of \$230,820.

c) $\widehat{Price} = 47.82 + 0.061\,Size$

$\widehat{Price} = 47.82 + 0.061(1200)$

$\widehat{Price} = 121.02$

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.

32. Last ride.

a) According to the linear model, the duration of a coaster ride is expected to increase by about 0.180 seconds for each additional foot of initial drop.

b) $\widehat{Duration} = 64.232 + 0.180\,Drop$

$\widehat{Duration} = 64.232 + 0.180(200)$

$\widehat{Duration} = 100.232$

According to the linear model, a coaster with a 200 foot initial drop is expected to last 100.232 seconds.

c) $\widehat{Duration} = 64.232 + 0.180\,Drop$

$\widehat{Duration} = 64.232 + 0.180(150)$

$\widehat{Duration} = 91.232$

According to the linear model, a coaster with a 150 foot initial drop is expected to last 91.232 seconds. The advertised duration is longer, at 120 seconds.
120 seconds – 91.232 seconds = 28.768 seconds, a positive residual.

33. 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."

34. 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."

35. 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, as well, meaning his score will probably go up. If the goal is to get a higher score, your friend should try again.

36. 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, Tiger Woods, Serena Williams, and others.

37. Cigarettes.

a) A linear model is probably appropriate. The residuals plot shows some initially low points, but there is not clear curvature.

b) 92.4% of the variability in nicotine level is accounted for by variability in tar content. (In other words, 92.4% of the variability in nicotine level is accounted for by the linear model.)

38. Attendance 2010.

a) The linear model is appropriate. Although the relationship is not strong, it is reasonably straight, and the residuals plot shows no pattern. There may be a bit of thickening on the right, but not enough to keep us from trying the linear model.

b) 28.4% of the variability in attendance is accounted for by variability in the number of wins. (In other words, 28.4% of the variability is accounted for by the model.)

c) The residuals spread out. There is more variation in attendance as the number of wins increases.

d) The Yankees attendance was about 13,000 fans more than we might expect given the number of wins. This is a positive residual.

39. Another cigarette.

a) The correlation between tar and nicotine is $r = \sqrt{R^2} = \sqrt{0.924} = 0.961$. 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 1.922 (2×0.961) standard deviations below the mean nicotine content.

c) Cigarettes that are one standard deviation above average in nicotine content are expected to be about 0.961 standard deviations (in other words, *r* standard deviations) above the mean tar content.

40. Second inning 2010.

a) The correlation between attendance and wins is $r = \sqrt{R^2} = \sqrt{0.284} = 0.533$. 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 wins would be expected to have attendance that is 1.066 (or 2×0.533) 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 wins that is 0.533 standard deviations (in other words, *r* standard deviations) below the mean number of wins. 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.

41. Last cigarette.

a) $\widehat{Nicotine} = 0.15403 + 0.065052\,Tar$ is the equation of the regression line that predicts nicotine content from tar content of cigarettes.

b)

$$\widehat{Nicotine} = 0.15403 + 0.065052\,Tar$$
$$\widehat{Nicotine} = 0.15403 + 0.065052(4)$$
$$\widehat{Nicotine} = 0.414$$

The model predicts that a cigarette with 4 mg of tar will have about 0.414 mg of nicotine.

c) For each additional mg of tar, the model predicts an increase of 0.065 mg of nicotine.

d) The model predicts that a cigarette with no tar would have 0.154 mg of nicotine.

e)

$$\widehat{Nicotine} = 0.15403 + 0.065052\,Tar$$
$$\widehat{Nicotine} = 0.15403 + 0.065052(7)$$
$$\widehat{Nicotine} = 0.6094$$

The model predicts that a cigarette with 7 mg of tar will have 0.6094 mg of nicotine. If the residual is –0.5, the cigarette actually had 0.1094 mg of nicotine.

42. Last inning 2010.

a) $\widehat{Attendance} = -6760.5 + 431.22\,Wins$ is the equation of the regression line that predicts attendance from the number of games won by American League baseball teams.

b)

$$\widehat{Attendance} = -6760.5 + 431.22\,Wins$$
$$\widehat{Attendance} = -6760.5 + 431.22(50)$$
$$\widehat{Attendance} = 14,800.5$$

The model predicts that a team with 50 wins will have attendance of approximately 14,800 people.

c) For each additional win, the model predicts an increase in attendance of 431.22 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 wins.

e)

$$\widehat{Attendance} = -6760.5 + 431.22(Wins)$$
$$\widehat{Attendance} = -6760.5 + 431.22(92)$$
$$\widehat{Attendance} = 32,911.74$$

The predicted attendance for the Cardinals was 32,911.74. The actual attendance of 41,736 gives a residual of 41,736 – 32,911.74 = 8,824.26.
The Cardinals had over 8,800 more people attending on average than the model predicted.

43. 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:

$$b_1 = \frac{rs_{HCI}}{s_{MFI}}$$
$$b_1 = \frac{(0.65)(116.55)}{7072.47}$$
$$b_1 = 0.0107$$

$$\hat{y} = b_0 + b_1 x$$
$$\overline{y} = b_0 + b_1 \overline{x}$$
$$338.2 = b_0 + 0.0107(46234)$$
$$b_0 = -156.50$$

The regression equation that predicts HCI from MFI is $\widehat{HCI} = -156.50 + 0.0107\,MFI$

c)

$$\widehat{HCI} = -156.50 + 0.0107\,MFI$$
$$\widehat{HCI} = -156.50 + 0.0107(44993)$$
$$\widehat{HCI} = 324.93$$

The model predicts that a state with median family income of \$44993 have an average housing cost index of 324.93.

d) The prediction is 223.09 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 $\hat{z}_{HCI} = 0.65 z_{MFI}$.

f) The correlation is the slope of the regression line that relates *z*-scores, so the regression equation would be $\hat{z}_{MFI} = 0.65 z_{HCI}$.

44. Interest rates and mortgages again.

a) Yes. Both interest rate and total mortgages are quantitative, and the scatterplot is Straight Enough. The spread is fairly constant, and there are no outliers.

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

$$b_1 = \frac{rs_{MortAmt}}{s_{IntRate}}$$
$$b_1 = \frac{(-0.86)(47.36)}{1.79}$$
$$b_1 = -22.754$$

$$\hat{y} = b_0 + b_1 x$$
$$\overline{y} = b_0 + b_1 \overline{x}$$
$$121.8 = b_0 - 22.754(7.74)$$
$$b_0 = 297.92$$

The regression equation that predicts total mortgage amount from interest rate is $\widehat{MortAmt} = 297.9 - 22.75\,IntRate$

(From the original data, $\widehat{MortAmt} = 309.13 - 22.79\,IntRate$)

c)

$$\widehat{MortAmt} = 297.9 - 22.75\,IntRate$$
$$\widehat{MortAmt} = 297.9 - 22.75(13)$$
$$\widehat{MortAmt} = 2.15$$

If interest rates were 13%, we would expect there to be \$2.15 million in total mortgages. (\$12.79 million if you worked with the actual data.)

d) 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 may not be appropriate.

e) The correlation is the slope of the regression line that relates z-scores, so the regression equation would be $\hat{z}_{MortAmt} = -0.86 z_{IntRate}$.

f) The correlation is the slope of the regression line that relates z-scores, so the regression equation would be $\hat{z}_{IntRate} = -0.86 z_{MortAmt}$.

45. Online clothes.

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

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

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

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)

$$\widehat{Total} = 539.803 + 1.103\,Age$$
$$\widehat{Total} = 539.803 + 1.103(18)$$
$$\widehat{Total} = 559.66$$

The model predicts that an 18 year old will have \$559.66 in total yearly online clothing purchases.

$$\widehat{Total} = 539.803 + 1.103\,Age$$
$$\widehat{Total} = 539.803 + 1.103(50)$$
$$\widehat{Total} = 594.95$$

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.

46. Online clothes II.

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

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

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

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)

$\widehat{Total} = 28.73 + 0.0108\,Income$
$\widehat{Total} = 28.73 + 0.0108(20,000)$
$\widehat{Total} = 244.73$

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.)

$\widehat{Total} = 28.73 + 0.0108\,Income$
$\widehat{Total} = 28.73 + 0.0108(80,000)$
$\widehat{Total} = \$892.73$

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.)

d) R^2 $(0.722)^2$ 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.

47. 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) $r = \sqrt{R^2} = \sqrt{0.469} = 0.685$ 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, $R^2 = 0.469$, 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.

The equation of the least squares regression line for predicting SAT Math score from SAT Verbal score is $\widehat{Math} = 209.5542 + 0.67507\,Verbal$.

e) For each additional point in verbal score, the model predicts an increase of 0.675 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 6.75 points in math score.

f)

$\widehat{Math} = 209.5542 + 0.67507\,Verbal$

$\widehat{Math} = 209.5542 + 0.67507(500)$

$\widehat{Math} = 547.0892$

According to the model, a student with a verbal score of 500 was expected to have a math score of 547.09.

g)

$\widehat{Math} = 209.5542 + 0.67507\,Verbal$

$\widehat{Math} = 209.5542 + 0.67507(800)$

$\widehat{Math} = 749.6102$

According to the model, a student with a verbal score of 800 was expected to have a math score of 749.61. The student actually scored 800 on math, so the residual was 800 – 749.61 = 50.39 points

48. 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.

The regression equation predicting GPA from SAT score is:

$\widehat{GPA} = -1.262 + 0.00214\,SAT$

b) The model predicts that a student with an SAT score of 0 would have a GPA of –1.262. The y-intercept is not meaningful, since an SAT score of 0 is impossible, and a GPA can't be negative.

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)

$\widehat{GPA} = -1.262 + 0.00214\,SAT$

$\widehat{GPA} = -1.262 + 0.00214\,(2100)$

$\widehat{GPA} \approx 3.232$

According to the model, a student with an SAT score of 2100 is expected to have a GPA of 3.23.

e) According to the model, SAT score is not a very good predictor of college GPA. $R^2 = 22.1\%$, 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.

49. SAT, take 2.

a) $r = \sqrt{R^2} = \sqrt{0.469} = 0.685$ 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.

The equation of the least squares regression line for predicting SAT Verbal score from SAT Math score is: $\widehat{Verbal} = 171.333 + 0.6943\,Math$

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)

$\widehat{Verbal} = 171.333 + 0.6943\,Math$

$\widehat{Verbal} = 171.333 + 0.6943\,(500)$

$\widehat{Verbal} = 518.483$

According to the model, a person with a math score of 500 was expected to have a verbal score of 518.5 points.

e)

$$\widehat{Math} = 209.5542 + 0.67507\,Verbal$$
$$\widehat{Math} = 209.5542 + 0.67507(518.483)$$
$$\widehat{Math} = 559.567$$

According to the model, a person with a verbal score of 516.606 was expected to have a math score of 559.567 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.

50. Success, part 2.

The typical GPA prediction misses the actual GPA by about 0.275 points.

51. Wildfires 2010.

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 an increase of an average of about 210 wildfires per year.

c) It seems reasonable to interpret the intercept. The model predicts about 74,487 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 11920 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 1.9% 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.

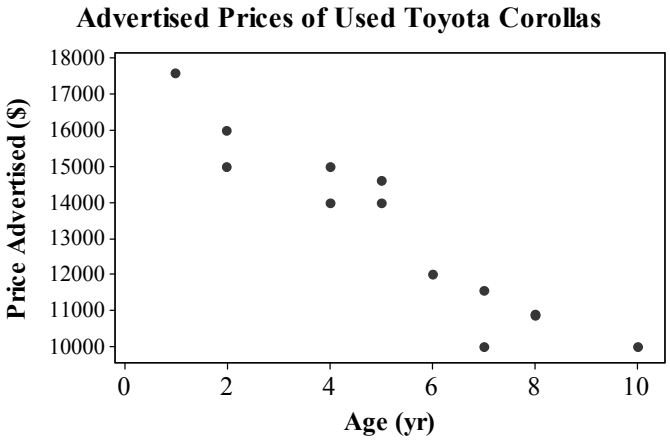
52. Wildfires 2010 – sizes.

a) The regression model isn't appropriate the association between the number of acres per fire and the year. The scatterplot is not straight, and the residuals plot shows a curved pattern, indicating that there is still information contained in the model that is not accounted for by the linear model.

b) R^2 is not of any use when the model is curved, since it is based on *r*, which is a measure of linear association.

53. Used cars 2011.

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 at the right.



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. $\sqrt{0.891} = 0.944$. Since association between age and price is negative, $r = -0.944$.

e) 89.1% 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.

54. Drug abuse.

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).

$$b_1 = \frac{rs_O}{s_M} \qquad \hat{y} = b_0 + b_1 x$$

$$b_1 = \frac{(0.93434)10.2}{15.6} \qquad \overline{y} = b_0 + b_1 \overline{x}$$

$$b_1 = 0.61091 \qquad 11.6 = b_0 + 0.61091(23.9)$$

$$b_0 = -3.001$$

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 the data set, and technology: $\widehat{Other} = -3.068 + 0.615\, Marijuana$)

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.

55. More used cars 2011.

a) The scatterplot from the previous exercise 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 $\widehat{Price} = 17767 - 862\,Years$.
The computer regression output used is at the right.

Dependent variable is: **Price Advertised ($)**
No Selector
R squared = 89.1% R squared (adjusted) = 88.1%
s = 845.3 with 13 - 2 = 11 degrees of freedom

Source	Sum of Squares	df	Mean Square	F-ratio
Regression	64.4816e6	1	64.4816e6	90.3
Residual	7.85899e6	11	714454	

Variable	Coefficient	s.e. of Coeff	t-ratio	prob
Constant	17767	535.7	33.2	≤ 0.0001
Age (yr)	-862.055	90.74	-9.5	≤ 0.0001

b) According to the model, for each additional year in age, the car is expected to drop $862 in price.

c) The model predicts that a new Toyota Corolla (0 years old) will cost $17,767.

d)

$\widehat{Price} = 17767 - 862\,Years$
$\widehat{Price} = 17767 - 862(7)$
$\hat{Price} = 11733$

According to the model, an appropriate price for a 7-year old Toyota Corolla is $11,733.

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

f) $\widehat{Price} = 17767 - 862\,Years$
$\widehat{Price} = 17767 - 862(10)$
$\widehat{Price} = 9147$

According to the model, a 10-year-old Corolla is expected to cost $9147. The car has an actual price of $8500, so its residual is $8500 — $9147 = — $647

g) The model would not be useful for predicting the price of a 25-year-old Corolla. The oldest car in the list is 10 years old. Predicting a price after 25 years would be an extrapolation.

56. Birthrates 2009.

a) A scatterplot of the live birth rates in the US over time is at the right. The association is negative, strong, and appears to be curved, 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 is slightly curved, it is straight enough to try a linear model. The linear regression output from a computer program is shown below:

Dependent variable is: **Rate**
No Selector
R squared = 72.8% R squared (adjusted) = 69.4%
s = 1.051 with 10 - 2 = 8 degrees of freedom

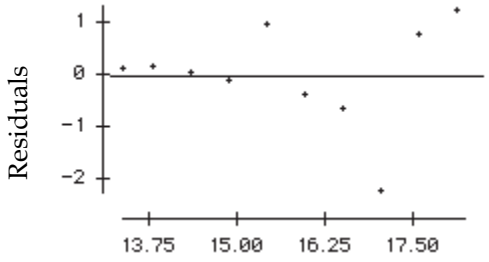
Source	**Sum of Squares**	**df**	**Mean Square**	**F-ratio**
Regression	23.7142	1	23.7142	21.5
Residual	8.84183	8	1.10523	

Variable	**Coefficient**	**s.e. of Coeff**	**t-ratio**	**prob**
Constant	25.1935	2.072	12.2	≤ 0.0001
Year-1900	-0.108393	0.0234	-4.63	0.0017

The linear regression model for predicting birth rate from year is:

$\widehat{Birthrate} = 25.1935 - 0.108393(Year - 1900)$

c) The residuals plot, at the right, shows a slight curve. 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 0.11 births per 1000 women.

e) $\widehat{Birthrate} = 25.1935 - 0.108393(Year - 1900)$

$\widehat{Birthrate} = 25.1935 - 0.108393(78)$

$\widehat{Birthrate} = 16.739$

The model predicts about 16.739 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 − 16.739 = −1.739 births per 1000 women. This means that the model predicted 1.74 births higher than the actual rate.

g) $\widehat{Birthrate} = 25.1935 - 0.108393(Year - 1900)$

$\widehat{Birthrate} = 25.1935 - 0.108393(110)$

$\widehat{Birthrate} = 13.27$

According to the model, the birth rate in 2010 is predicted to be 13.27 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 72.8% of the variability in birth rate. Don't place too much faith in this prediction.

h)

$\widehat{Birthrate} = 25.1935 - 0.108393(Year - 1900)$

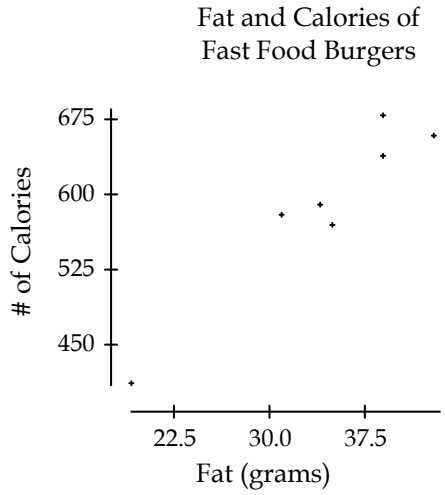
$\widehat{Birthrate} = 25.1935 - 0.108393(125)$

$\widehat{Birthrate} = 11.64$

According to the model, the birth rate in 2025 is predicted to be 11.64 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.

57. Burgers.

a) The scatterplot of calories vs. fat content in fast food hamburgers is at the right. 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

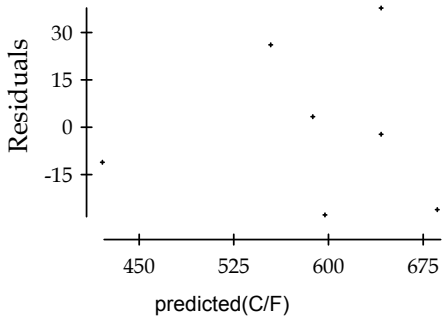
Source	Sum of Squares	df	Mean Square	F-ratio
Regression	44664.3	1	44664.3	59.8
Residual	3735.73	5	747.146	

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 at the right shows no pattern. The linear model appears to be appropriate.

e) The model predicts that a fat free burger would have 210.954 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 11.056 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 ≈ 553.5 calories.

58. Chicken.

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

b) $r = \sqrt{R^2} = \sqrt{0.932} = 0.965$. The correlation of 0.965 indicates a strong, linear, positive relationship between fat and calories for chicken sandwiches.

c) The linear model for predicting calories from fat in chicken sandwiches is:

$\widehat{Calories} = 12.7234 + 21.2171\,Fat$

d) For each additional gram of fat, the model predicts an increase of 21.22 calories.

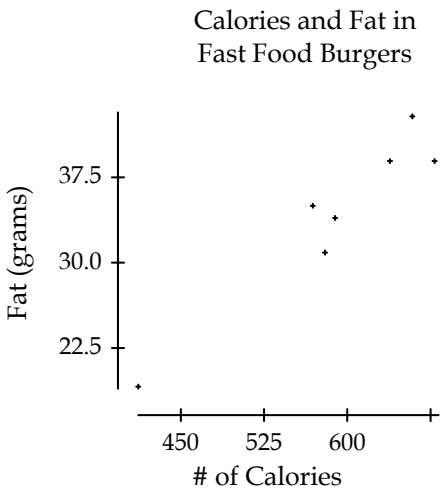
e) According to the model, a fat-free chicken sandwich would have 12.72 calories. This is an extrapolation, and has no practical significance.

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

59. A second helping of burgers.

a) The model from the previous 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 at the right 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.



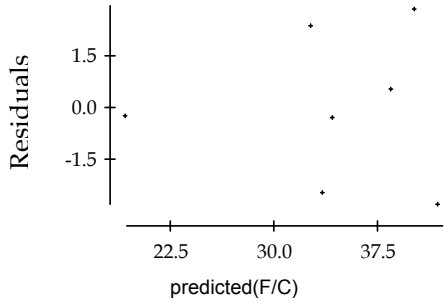
Dependent variable is: **Fat**
No Selector
R squared = 92.3% R squared (adjusted) = 90.7%
s = 2.375 with 7 - 2 = 5 degrees of freedom

Source	Sum of Squares	df	Mean Square	F-ratio
Regression	337.223	1	337.223	59.8
Residual	28.2054	5	5.64109	

Variable	Coefficient	s.e. of Coeff	t-ratio	prob
Constant	-14.9622	6.433	-2.33	0.0675
Calories	0.083471	0.0108	7.73	0.0006

The linear model for predicting fat from calories is: $\widehat{Fat} = -15.0 + 0.083\,Calories$

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.

$\widehat{Fat} = -15.0 + 0.083\,Calories$

$\widehat{Fat} = -15.0 + 0.083(600)$

$\widehat{Fat} \approx 34.8$

According to the model, a burger with 600 calories is expected to have 34.8 grams of fat.

60. A second helping of chicken.

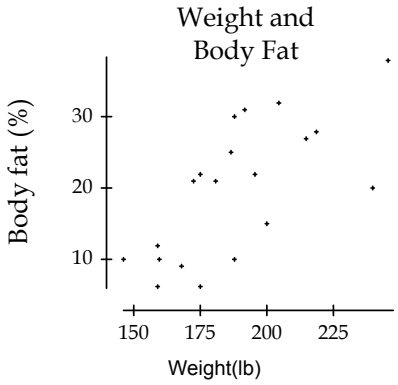
a) The model from the previous exercise was for predicting number of calories from number of grams of fat. In order to predict grams of fat from the number of calories, a new linear model needs to be generated.

b) The Quiznos large mesquite sandwich appears to be strengthening the association. Without it, the relationship would appear weaker than it currently appears.

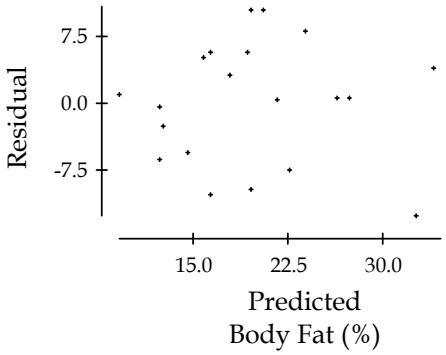
61. Body fat.

a) The scatterplot of % body fat and weight of 20 male subjects, at the right, 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: $\widehat{\%Fat} = -27.3763 + 0.249874\,Weight$



b) The residuals plot, at the right, 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)

$\widehat{\%Fat} = -27.3763 + 0.249874\,Weight$

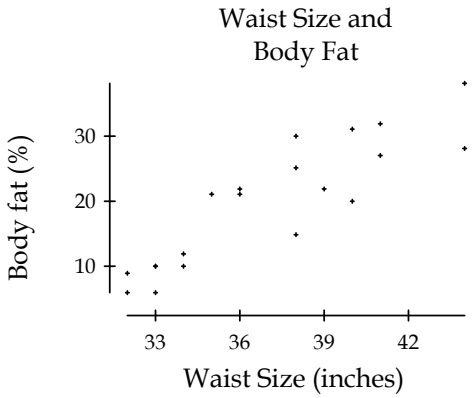
$\widehat{\%Fat} = -27.3763 + 0.249874(190)$

$\widehat{\%Fat} = 20.09976$

According to the model, the predicted body fat for a 190-pound man is 20.09976%.
The residual is $21 - 20.09976 \approx 0.9\%$.

62. Body fat, again.

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

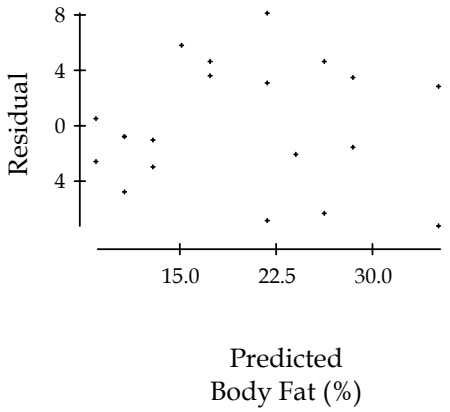


The linear model for predicting % body fat from waist size is :

$\widehat{\%Fat} = -62.557 + 2.222\,Waist$.

For each additional inch in waist size, the model predicts an increase of 2.222% body fat.

78.7% of the variability in % body fat can be accounted for by waist size. The residuals plot, at right, 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.



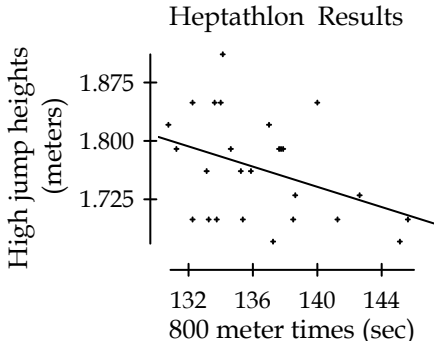
63. Heptathlon 2004.

a) Both high jump height and 800 meter time are quantitative variables, the association is straight enough to use linear regression.

Dependent variable is: **High Jump**
No Selector
R squared = 16.4% R squared (adjusted) = 12.9%
s = 0.0617 with 26 - 2 = 24 degrees of freedom

Source	Sum of Squares	df	Mean Square	F-ratio
Regression	0.017918	1	0.017918	4.71
Residual	0.091328	24	0.003805	

Variable	Coefficient	s.e. of Coeff	t-ratio	prob
Constant	2.68094	0.4225	6.35	< 0.0001
800m	-6.71360e-3	0.0031	-2.17	0.0401



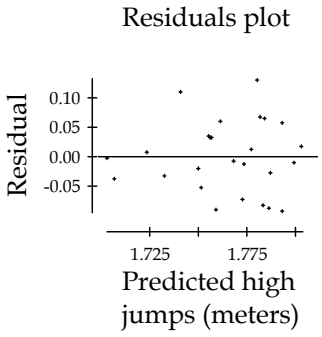
The regression equation to predict high jump from 800m results is: $\widehat{Highjump} = 2.681 - 0.00671 Time$.

According to the model, the predicted high jump decreases by an average of 0.00671 meters for each additional second in 800 meter time.

b) $R^2 = 16.4\%$. This means that 16.4% of the variability in high jump height is accounted for by the variability in 800 meter time.

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

d) The residuals plot is fairly patternless. The scatterplot shows a slight tendency for less variation in high jump height among the slower runners than the faster runners. Overall, the linear model is appropriate.



e) The linear model is not particularly useful for predicting high jump performance. First of all, 16.4% of the variability in high jump height is accounted for by the variability in 800 meter time, leaving 83.6% of the variability accounted for by other variables. Secondly, the residual standard deviation is 0.062 meters, which is not much smaller than the standard deviation of all high jumps, 0.066 meters. Predictions are not likely to be accurate.

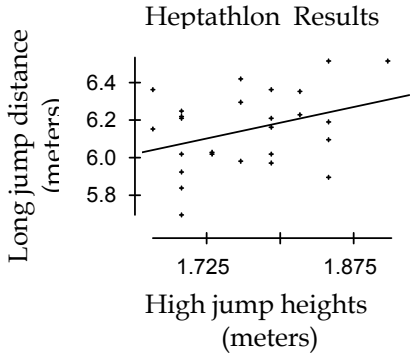
64. Heptathlon 2004 again.

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.

Dependent variable is: **Long Jump**
No Selector
R squared = 12.6% R squared (adjusted) = 9.0%
s = 0.1960 with 26 - 2 = 24 degrees of freedom

Source	Sum of Squares	df	Mean Square	F-ratio
Regression	0.133491	1	0.133491	3.47
Residual	0.922375	24	0.038432	

Variable	Coefficient	s.e. of Coeff	t-ratio	prob
Constant	4.20053	1.047	4.01	0.0005
High Jump	1.10541	0.5931	1.86	0.0746



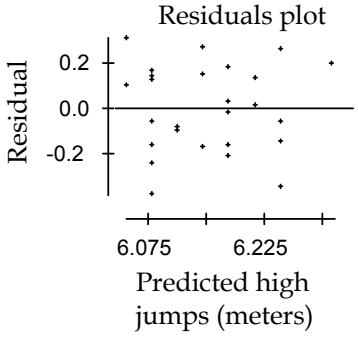
The regression equation to predict long jump from high jump results is:
$\widehat{Longjump} = 4.20053 + 1.10541\,Highjump$.

According to the model, the predicted long jump increases by an average of 1.1054 meters for each additional meter in high jump height.

b) $R^2 = 12.6\%$. This means that only 12.6% 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.

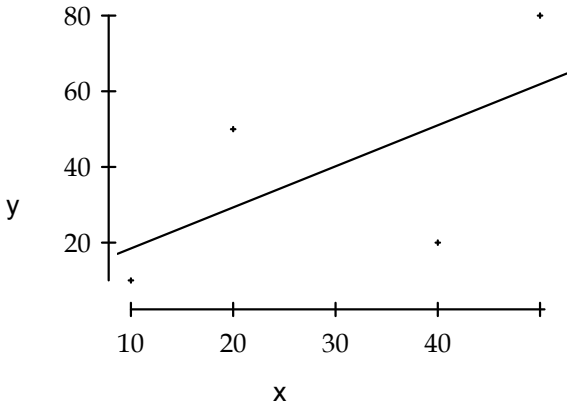
d) The residuals plot is fairly patternless. The linear model is appropriate.



e) The linear model is not particularly useful for predicting long jump performance. First of all, only 12.6% of the variability in long jump distance is accounted for by the variability in high jump height, leaving 87.4% of the variability accounted for by other variables. Secondly, the residual standard deviation is 0.196 meters, which is about the same as the standard deviation of all long jumps jumps, 0.206 meters. Predictions are not likely to be accurate.

53. 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.



54. 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.