Chapter 3

Discussion Question Solutions

D1. **a.** Display 3.1 illustrates the commonsense idea that people born earlier typically become employed before people who were born later. A younger person in the company will then have less seniority than an older employee.

b. All the points are in the upper-left half of the plot because people cannot be hired until they are about 18 years old. Specifically, a person's age must be at least 18 years greater than their seniority so one would expect all points to be above the line $y = x + 18$. In fact, there is one point below this diagonal line, a person who was hired into the company at a very young age during the 1940s. Although there is a lot of variation, in general people were born 25 or 30 years before they were hired.

Note on D1b: As you discuss the graph in the context of a real situation, introduce the practice of writing models with variable names that are in the context of the problem. For example, the line $y = x + 18$ could be stated as year of *Age* = *Seniority* + 18. The practice of using intelligible variable names is consistent with many statistics software programs but not with calculators. Mixing words and symbols can be a problem at lower levels, but should not present difficulty for your students.

c. No, this is not correct. The ages plotted are not the ages of the employees when they were hired but their ages when layoffs began. This idea will be explored further in E7.

D2. a. For these data, the cases are the states and the variables are the number of people per thousand living in dorms and the proportion of the state population living in cities. The shape of the cluster is linear (roughly oval or elliptical), except for three points (VT, RI, and MA) that lie relatively far away from the main cloud of points. Vermont, a rural state, has a large number of colleges and a higher dorm proportion than would be anticipated. Rhode Island and Massachusetts also have a relatively high proportion living in dorms, but they are essentially urban states. The trend is negative, as a larger proportion of people living in cities tends to mean a smaller proportion in dorms. There is a lot of variability in dorm proportions for any particular proportion living in cities, and thus the strength of the association between the two variables is only moderate. Other than the three apparent outliers, the strength is relatively constant across all values of proportion of population living in cities. This scatterplot shows the data for all of the 50 states, so there is no larger population to generalize to. What you see is all there is for this particular year. However, it is reasonable to generalize to a previous or subsequent year. A possible explanation for this pattern is that states with a high proportion of the population living in cities also have a high proportion of their colleges located in cities. In an urban area, there is little need for students to live in dorms. They can commute from home or get off-campus housing in nearby apartments. Thus, for highly urbanized states, a lower proportion of students need to live in dorms.

b. The positive trend in the original data comes from the fact that states with a large number of people tend to have a large number of colleges and universities and a large number of people

living in cities. A possible explanation for the negative trend in the proportion data is given in part a.

D3. a. The slope of the line is 13.812. The slope tells us that the CPI tended to increase at the rate of approximately \$13.81 per year across this time span.

b. The y-intercept is -27102 and is the value of the model when $x = 0$. The given model would likely not apply to years much further back than 1970 (see D4) so the point $(0, -27102)$ cannot be interpreted as a valid year-CPI pair.

D4. a. The equation predicts a CPI of $-27102 + 13.812(1960) = -30.48$ in 1960. The error in this prediction is $88.7 - (-30.48) = 119.18$.

b. Looking at the graph in Display 3.21, the point furthest the line, and so the point having the greatest residual, is the point corresponding to 1975.
The residual is $161.2 - [-27102 + 13.812(1975)] = -15.5$.
In 1975, the United States was coming out of a two-year recession. This particular recession was notable as it was also a period of high inflation. When coming out of this recession during 1975, prices therefore did not rise as much as prior years or subsequent years after the economy improved.

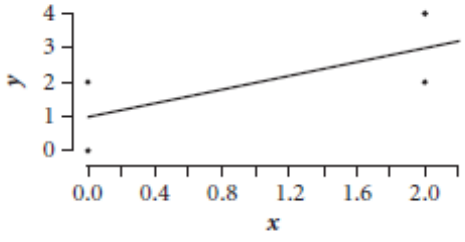
D5. a. Far below the line; the point lies on the line.

b. The fitted line is too low. It lies below all of the points. Move the line up so that there are both positive and negative residuals.

D6. The arithmetic is fine. The reasoning is an amusing example of the folly of extreme extrapolation. The equation is $length = 6460 - 1.375\ year$.

D7. a. The plot and a table that includes the absolute deviations are shown. This line fits best because it passes through the mean value of y at each value of x. It also minimizes the variation among the residuals.

x	y	$\hat{y}$	$\lvert y - \hat{y} \rvert$
0	0	1	1
0	2	1	1
2	2	3	1
2	4	3	1
			$\sum \lvert y - \hat{y} \rvert = 4$



b. Another such line is $y = 1.5 + 0.5x$. The residuals are $-1.5,\ 0.5,\ -0.5,$ and 1.5, which sum in absolute value to 4.

c. Another such line is $y = 0.2 + 1.6x$. The residuals are $-0.2,\ 1.8,\ -1.4,$ and 0.6, which sum in absolute value to 4.

d. Such a line is $y = 2.5 + 0.5x$. The residuals are $-2.5,\ -0.5,\ -1.5,$ and 0.5, which sum in absolute value to 5.

e. The original line, $y = 1 + x$, is the least squares line. The sum of squared residuals is 4. The sum of squared residuals for the lines in b, c and d are, respectively, 5.0, 5.6, and 9.0. The least squares line minimizes the sum of squared residuals among these lines.

f. The standard deviation of the residuals for the least squares line is 1.15. For the lines in b, c, and d, the standard deviations are, respectively, 1.2910, 1.3466, and 1.2910. The least squares line also minimizes the standard deviations of the residuals among these lines.

D8. a. $income = -9124.7 + 4.6381\ year$. Estimating the SE from the plot, you get about $(-2)^2 + 2^2 + 0^2 + (-2)^2 + (-3)^2 + 3^2 + 6^2 + 3^2 + 1^2 + (-11)^2 + 2^2 + 1^2$, or about 202. From the printout, you get 207.2.

b. The main difference in the output is that the StatPlus output gives the values to more decimal places than Minitab. StatPlus also returns an additional R^2 value and expresses all R^2 values as decimal numbers not percentages.

D9. a. 0.783 **b.** 0.999 **c.** 0.906

D10. a. Relationships II and III are positive. Relationship IV is negative because the more socks in a bag, the cheaper they tend to be per sock. Relationship II is the strongest, with $r = 1$. Relationship I is the weakest; r is almost 0.

b. For I, there should be no relationship between these two variables. That is, for any day of the month picked, range of haircut costs will be about the same. This range will depend on local prices, and knowing the day of the month a person was born tells you nothing further about the cost of a haircut. Thus, there is virtually no correlation between these two variables.
For II, y does vary with x, but for a fixed x, there will no variation in y. Each y will be exactly equal to πx. The correlation is 1.
For III, generally, the more socks in a bag, the higher the price of the bag. There will be some variation in price because some brands of socks are more expensive than others, so there would be a strong correlation, but not a perfect one. Knowing the number of socks in the package does assist you in predicting a price range for the bag.
For IV, there will be some variation. Generally, the more socks in a bag, the cheaper the cost per sock. Knowing how many socks are in a bag can assist you in your prediction of the price range per sock.
All else being equal, the larger the variation in y at each value of x, the lower the correlation.

D11. a. For Student A,

$$\frac{x-\bar{x}}{s_x}=\frac{3-11}{8.433}=-0.9487 \qquad \frac{y-\bar{y}}{s_y}=\frac{3.5-3.1}{0.386}=1.0363$$

$$\left(\frac{x-\bar{x}}{s_x}\right)\left(\frac{y-\bar{y}}{s_y}\right)=-0.9487\cdot 1.0363=-0.9831$$

or -0.98 rounded.

b. Student J, in the lower right corner of Quadrant IV, has the largest product $z_x \cdot z_y$, in absolute value.

c. Student B, lying on the horizontal y mean line between Quadrants II and III, has the smallest value of $|z_x \cdot z_y|$. A point will make a small contribution if it is either near $(\bar{x}, \bar{y})$ or near one of the lines $x=\bar{x}$ and $y=\bar{y}$.

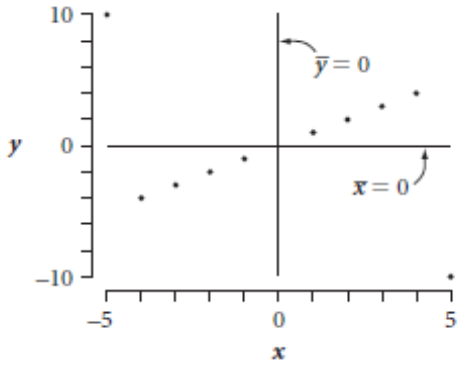
D12. a. The correlation measures the strength of a linear association by measuring how tightly packed the data points are about a straight line. Its size is affected most strikingly by points far away from $(\bar{x}, \bar{y})$ and points not near the new coordinate axes, $x=\bar{x}$ and $y=\bar{y}$.

b. Correlation is a unitless quantity because it is the average product of z-scores, which have no units. For example, the unit for the TV watching data is hours. When computing the z-scores, the units cancel out. For Student A, this would be

$$\frac{x-\bar{x}}{s_x}=\frac{3\,hr-11\,hr}{8.433\,hr}\approx -0.9487$$

c. No. Because it is based on a symmetric calculation, $z_x \cdot z_y = z_y \cdot z_x$, r does not depend on which variable is chosen as x and which as y.

D13. a. For well-behaved data, the correlation will be positive because most of the products $z_x \cdot z_y$ are positive. It is not necessarily the case that r is positive, however. The scatterplot here has more points in Quadrants I and III, but the correlation is negative, $r=-0.237$.



b. For well-behaved data, the correlation will be negative because most of the products $z_x \cdot z_y$ are

negative. As in part a, it is not necessarily the case that r is negative.

c. For well-behaved data, the correlation will be near 0.

D14. When the correlation is small, the error in prediction will be larger than if the correlation were larger. A larger correlation (near 1 or -1) means the points are generally nearer the line, and predictions made using the line will be relatively close to the observed values. Even though the error in prediction is large when the correlation is small, having a regression line is better than having no line as long as there is a definite linear trend in the data. You could give an example from the quiz scores in Display 3.55. A student who scored 10 on the second quiz would be predicted to get about 15 on the third quiz, but we wouldn't be surprised if this prediction was off by 10 points or so. A student who scored 25 on the first quiz, would be expected to get about 23 on the second quiz, again with about the same estimated error. These lead to two different predictions for the two students: 15 ± 10 is different from 23 ± 10 even though there is some overlap in what we would think of as reasonable bounds on the predictions.
Note on D14: Many students will say that the prediction from the regression equation is better than no prediction at all. To lead students toward the idea of r^2 introduced on page 135, ask them what they mean by "no prediction." That is, what would their estimate be if they had full information about the scores on Quiz 2 and Quiz 3, but no information about the relationship between the two?

D15. a. Scenarios could be situations in which there is not a definite linear trend in the data (y appears to be unrelated to x), along with large variation in the y-values. Age versus *month of birth* for a large group of adults might be an example.

b. Scenarios could be situations in which there is a definite linear trend, but where there is much variation in the y-values at each level of x. *SAT math score* versus *score on the first college calculus test* for a group of college students could be an example. *Family grocery bill per month* versus *number of people in the family* could be another.

c. Scenarios could be situations in which the data points fit closely to a line but the cloud of points has considerable curvature, so as to make the straight line a poor measure of the center, or a poor description of the nature of the association between the variables. *Height* versus *age* for trees could be an example, because height levels off for older trees. You have seen other curved relationships earlier in this chapter.

d. Scenarios could be situations in which the data points fit closely to a line and the line has a slope that is not close to zero. *Height* versus *age* for growing children could be an example, as could height versus shoe size for adults.

D16. The growth rate will probably begin to slow down at some point, if it hasn't already. New blogs will continue to appear, but probably not at the same rate they did initially.

D17. An estimate of the slope is

$$b_1 = r\frac{s_y}{s_x} = 0.7 \cdot \frac{112}{116} \approx 0.68$$

To find the y-intercept, use the fact that the point (515, 502) is on the regression line:

$$y = slope \cdot x + y\text{-}intercept$$
$$502 = 0.68 \cdot 515 + y\text{-}intercept$$
$$y\text{-}intercept = 151.8$$

The equation is

$$critical\ reading = 151.8 + 0.68\ math\ .$$

D18. A student spending many hours watching TV instead of studying could lead to a low GPA so a causal relationship is possible. It is also possible that there is no direct link between the variables; both could be a result of the lurking variable of the student's study skills or the amount the student learned in class. For example, a student may become frustrated in their studies, give up and go watch TV.

D19. At first glance, it would seem that the more highly rated the university, the lower its acceptance rate. One possible explanation is that few students apply to the most selective of these universities unless they are pretty sure they will be admitted. You would not say that one variable causes the other, rather that they are both associated with the most savvy students. Another possible lurking variable is that the very best students apply to more colleges because they are shopping for the college that will offer them the best financial aid deal. As a result, the more highly rated colleges must accept a high percentage of the students that apply because they know the students have applied many places and are likely to go elsewhere even if they are admitted.

D20. a. Some might say that a high percentage of males "causes" higher salaries because men are more favorably treated than women when it comes to salary. On the other hand, the lurking variable may be how quantitative the subject is. There tend to be far fewer graduates in quantitative subjects, and business and industry want to hire them also; therefore, these faculty positions are harder to fill. This competition for fewer graduates may be the reason the people in quantitative subjects (who tend to be male) are more highly paid.

b. Some people might say that a high number of hate groups "causes" a large number of people on death row because members of hate groups tend to commit murders. On the other hand, an obvious lurking variable here is the size of the state's population. Larger states have more of everything. In fact, in percentage terms, hate crimes account for a small proportion of those on death row. To determine whether more hate groups in a state results in more people on death row because of hate crimes, you might look at a scatterplot of the percentage of people on death row because of hate crimes against the percentage of people in the state who belong to hate groups.

c. Some people might say that a high rate of gun ownership "causes" lower rates of violent crime by making criminals reluctant to commit violent crime for fear the person will protect himself or herself with a gun. On the other hand, the explanation may be the lurking variable of how rural the state is. Rural areas have higher rates of gun ownership, presumably for hunting, and have lower reported crime.

D21. a. The best prediction would be the mean IQ of 101.

b. Approximately 1 point.

c. $0.997 \cdot 54 + 45 = 98.8$. You should not have much faith in this prediction because the variation around the line is so great.

d. Approximately 2% of the variation is accounted for by taking head circumference into account. The regression equation is not of much practical help in making the prediction.

e. Answers will vary. It seems likely that the regression line would become less steep and the correlation would decrease, indicating little to no correlation between head circumference and IQ.

D22. Rewrite the ratio as follows: $r^2 = \dfrac{\text{SST} - \text{SSE}}{\text{SST}} = 1 - \dfrac{\text{SSE}}{\text{SST}}$ Because SSE is less than or equal to SST and both are positive, the ratio $\dfrac{\text{SSE}}{\text{SST}}$ will always be between 0 and 1 inclusive. Thus, r^2 will always be between 0 and 1 inclusive. So r must always be between -1 and 1 inclusive.

D23. The number of "heating units" used by a house varies from year to year. A good predictor of how many units that will be is temperature. The investigator is saying that temperature accounts for 70% $(r^2 = 0.7)$ of the year-to-year variation.

D24. The regression line is a "line of means" because it attempts to go through the mean value of the *y*'s at each fixed value of *x*. That is, the regression line estimates the mean value of *y* for each fixed value of *x*.

D25. Two older sisters 1 inch apart in height will, on the average, have two younger sisters who are only 0.337 inches apart. For the line $y = x$, the interpretation would be that if one older sister is 1 inch taller than another, her younger sister also tends to be 1 inch taller than the younger sister of the other older sister. The latter interpretation is what most people would expect. But the element of chance involved results in the regression effect that the younger sister is not as tall as expected.

Note on D26–28: Students might work in groups of four, with each group calculating the correlation coefficient and the regression equation for one of the four Anscombe data sets. Groups can then share their results. Seeing similar summaries for quite different graphs will reinforce the fact that numerical summaries are not sufficient to describe any data set, either univariate or bivariate.

D26. Plot I shows a positive linear trend that is moderately strong. Plot II shows points that lie along a curve. Plot III shows all points lying on a straight line except for the one point near the right end. (This point will have the effect of raising that end of the regression line.) Plot IV has all but one of the points stacked up at the same value of x, although the outlier gives it a slight

positive trend. (The regression line will go through the middle of the points on the left and through the isolated point on the right.)

a. A straight line is a good summary only for plot I. However, in all four plots the regression line has a slope of about 0.5 and an intercept of about 3.0. (See D27.)

b. The correlation is about 0.8 for plot I and should not be used to describe the others because it is a measure of the strength of a linear relationship.

D27. There is no way to tell which plot produced the given summary statistics. In fact, the summary statistics are essentially the same for all four plots. The moral? Draw a picture before you summarize data!

D28. a. Plots III and IV both have influential observations, but plot IV contains the more influential point. The influential point in plot IV is more isolated from the data and completely controls the slope of the regression line.

b. This influential point lies on the regression line through these data.

c. If the influential point (the isolated point) is removed from plot IV, all of the data points will stack up at the same value of *x*. Thus, a regression slope and correlation cannot be computed.

D29. Of the two points aligned vertically at the far left of the plot, Student A corresponds to the higher one located at the point (3, 0.20). Student J produces the point at the far right of the plot at the point $(29, -0.14)$.

D30. a. A—I; B—IV; C—II; D—III

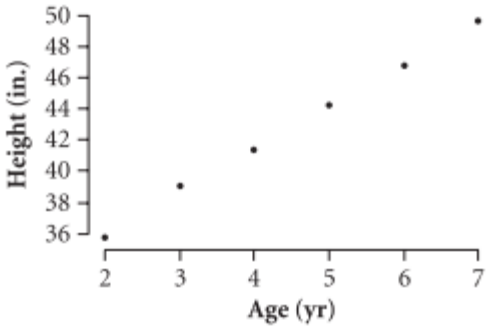
b. The scatterplot shows the actual values of *y* and upward or downward trend but may obscure patterns in the residuals (or at least appear to diminish them a bit). The residual plots do not show the values of *y* or trend in the original data, but they do show the values of the residuals, and they make departures from linearity easier to see.

D31. The vertical axis in both plots is the same while the horizontal axis is labeled according to the values of *x* in the first and the predicted values of *y* in the second. The key difference to note is the pattern of the data points in the second graph is the mirror image of the pattern in the first graph. For example, the two points at the far left in the first plot where x = 3, correspond to the two point at the far right in the second plot where $\hat{y} = 3.30$. This mirroring occurs because the slope of the regression line is negative, so the increasing values of x correspond to decreasing values of $\hat{y}$. If the regression slope is positive the scatterplot patterns would appear the same.

Practice Problem Solutions

P1. Plot a shows a positive relationship that is strong and linear. There is fairly uniform variation across all values of *x*.
Plot b shows a negative relationship that is strong and linear, again with fairly uniform variation across all values of *x*.
Plot c shows a positive relationship that is moderate and linear with fairly uniform variation across all values of *x*. There is one point that lies a short distance from the bulk of the data.
Plot d shows a negative relationship that is moderate and linear with fairly uniform variation across all values of *x*. Again, there is one outlier.
Plot e shows a positive relationship that is strong and linear except for the outlier. As students will learn, the one outlier has dramatic influence on the strength of this relationship. There is fairly uniform variation across all values of *x*.
Plot f shows a negative relationship that is very strong and curved. Again, one point lies in the general pattern but far away from the remainder of the data, which accentuates the strong relationship. Another outlier lies below the bulk of the data on the left.
Plot g shows a negative relationship that is strong and curved. The two points at either end of the array accentuate the curvature. There is a bit more variability among values of *y* for smaller values of *x* than for larger values of *x*.
Plot h shows a positive relationship that is strong and curved. Again, the outlier on the extreme right accentuates the curved pattern and would have dramatic influence on where a trend line might be placed. The variability in *y* is fairly constant across all values of *x*.

P2. a. You may have to remind students that a scatterplot without labels and units on the axes is meaningless. Emphasize the importance of appropriate labeling.
Here is the scatterplot.



b. These data are not very interesting to describe. The *x*-axis shows ages 2 to 7 years, and the *y*-axis shows an median height of children at each age. The shape is linear, the trend is positive, and the strength is very strong. That is, the scatterplot shows a very strong positive linear trend. A typical child grows about 2.7 inches per year.

c. The linear trend could reasonably be expected to hold for another year. However, *median height* could not be expected to increase at this rate to age 50, as people typically stop growing around age 20.

P3. a. The relationship is a positive trend, has moderate strength, and slight curvature with more variability at larger values of x.

b. Trees of same age at another location may have similar pattern; older trees may not fit this pattern.

c. As trees age, growth slows and variability among tree sizes is more pronounced.

P4. a. The points (40, 91) and (240, 88.3) lie on or near the regression line, so the slope is about –0.0135. Each day, the eraser tended to lose around 0.0135 gm of weight, on average.

b. The *y*-intercept is 91.48 grams. This is a prediction of the weight of the eraser at the beginning of the year.

P5. a. About 0.8.

b. If one student has a hand length that is 1 in. longer than that of another student, the first student's hand tends to be 0.8 in. wider.

c. The *y*-intercept is 1.69. This is a prediction of the width of a hand with no length. This is not reasonable.

d. The students in the lower cluster did not spread their fingers. The slope would increase. In fact, if these points were removed, the regression line would move up slightly at the end for smaller hand lengths and move up a bit more at the end for longer hand lengths.

P6. a. The slope is 0.072. This is an estimate of average thickness per sheet (in mm).

b. The *y*-intercept is 2.7 mm. This is the predicted thickness of a stack comprised of 0 sheets; i.e., the thickness of the cover. This seems reasonable.

c. $\hat{y} = 2.7 + 0.072(175) = 15.3mm$

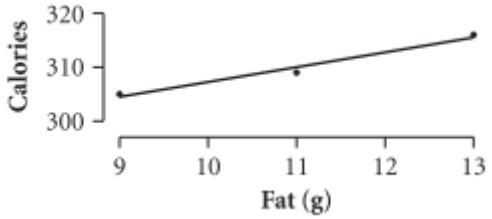
d. The biggest residual occurs for 100 sheets; it is positive.

e. The residuals are $y - \hat{y}$. Evaluating for all data points, we have the following:

Number of sheets x	y	Regression Line $\hat{y} = 2.7 + 0.072x$	Residual $y - \hat{y}$
50	6.0	6.3	-0.3
100	11.0	9.9	1.1
150	12.5	13.5	-1.0
200	17.0	17.1	-0.1
250	21.0	20.7	0.3

The sum of the values in the rightmost column is 0.

P7. a.



b. The equation is $\hat{y} = 279.75 + 2.75x$. The slope and y-intercept are found using the table below.

Pizza	x	y	$x-\bar{x}$	$y-\bar{y}$	$(x-\bar{x})\cdot(y-\bar{y})$	$(x-\bar{x})^2$
1	9	305	−2	−5	10	4
2	11	309	0	−1	0	0
3	13	316	2	6	12	4
Sum	**33**	**930**	**0**	**0**	**22**	**8**
Mean	**11**	**310**				

So, $\text{slope} = \frac{22}{8} = 2.75$, $\quad y-\text{intercept} = 310 - 2.75(11) = 279.75$.

P8. a. (*adult weight*) = -28.81 + 13.22 (*birth weight*)

b. For the llama, y = 140 and so $\hat{y} = -28.81 + 13.22(11) = 116.61$. So, the residual is $y - \hat{y} = 23.39$.

c.

Birth Weight	y	Regression Line $\hat{y} = 28.81 + 13.22x$	Residual $y - \hat{y}$
36	475	504.73	-29.73
37	434	517.95	-83.95
11	140	174.23	-34.23
11.5	100	180.84	-80.84
7.21	62	124.13	-62.13
5.74	50	104.69	-54.69

d. Square all values in the rightmost column in (c) and sum to obtain $\sum(y-\hat{y})^2 = 2586.2$.

e. $-28.81 + 13.22(\bar{x}) = -28.81 + 13.22(18.08) = 210.2 = \bar{y}$

P9. Yes, the regression equation of $\hat{y} = 279.75 + 2.75x$ and the mean of the response variable, 310, are the same as we computed by hand. The SSE is $0.5^2 + 1^2 + (-0.5)^2 = 1.5$ and is found in the Analysis of Variance table in row "Error," column "Sum of Squares."

P10. a. The slope is 1.13; on the average, the area increase by about 1.13 inches per year
b. The y-intercept is -3.70; it predicts the area of a tree at year 0 – this is not a practical interpretation.

P11. a. –0.5 **b.** 0.5 **c.** 0.95 **d.** 0 **e.** –0.95

P12. a. About 0.95 **b.** 0.99

P13. a. Note that

$$\overline{x} = 0,\ \overline{y} = 0.40,\ s_x = \sqrt{2.5} \approx 1.581,\ s_y = \sqrt{0.8} \approx 0.894\,.$$

Next, consider the following table of values:

x	$z_x = \frac{x-0}{1.581}$	y	$z_y = \frac{y-0.40}{0.883}$	$z_x \cdot z_y$
-2	-1.265	-1	-1.566	1.981
-1	-0.633	1	0.671	-0.4247
0	0	0	-0.447	0
1	0.633	1	0.671	0.4247
2	1.265	1	0.671	0.8488

So, $r = \frac{1}{5-1}\sum z_x \cdot z_y = 0.7075.$

b. Similarly, note that

$$\overline{x} = 0,\ \overline{y} = 3.0,\ s_x = \sqrt{2} \approx 1.414,\ s_y = \sqrt{1} = 1\,.$$

Next, consider the following table of values:

x	$z_x = \frac{x-0}{1.414}$	y	$z_y = \frac{y-3.0}{1}$	$z_x \cdot z_y$
-2	-1.414	2	-1	1.414
0	0	2	-1	0
0	0	3	0	0
0	0	4	1	0
2	1.414	4	1	1.414

So, $r = \frac{1}{5-1}\sum z_x \cdot z_y = 0.707.$

P14. 0.99; all but one of the products are positive.

P15. a. Positive

b. The point at the extreme upper right of the plot at about (5, 5) will make the largest positive contribution to the correlation because it is farthest away from the new origin $(\overline{x}, \overline{y})$ and from the new coordinate axes $(x = \overline{x})$ and $(y = \overline{y})$ and so has a large $z_x \cdot z_y$.

c. In Quadrants I and III; 20 have a positive product.

d. In Quadrants II and IV; 7 have a negative product.

P16. The plot on the top has a strong curvature. A line would not be appropriate here. The plot on the bottom is linear. The cloud of points is roughly elliptical. A line would be appropriate for this plot.

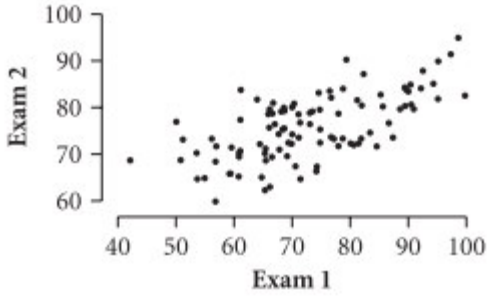
P17. a. The correlation is 0.650:

$$b_1 = r \cdot \frac{s_y}{s_x} \Rightarrow 0.368 = r \cdot \frac{7}{12.37} \Rightarrow r = 0.650$$

b. The regression equation is *exam* 2 = 48.94 + 0.368 *exam 1*. The predicted *exam 2* score is 78.38.

c. The regression equation is *exam 1* = –132.99 + 2.72 *exam 2*.

d.



P18. a. The size of the city's population.
b. You should divide each number by the population of the city to get the number of fast–food franchises per person and the proportion of the people who get stomach cancer.

P19. An obvious lurking variable is the age of the child. Parents tend to give higher allowances to older children, and vocabulary is larger for older children than for younger.

P20. A careless conclusion would be that people are too busy watching television to have babies. The lurking variable is how affluent the people in the country are. More affluent people tend to have more televisions and have fewer children.

P21. a. The formula relating these quantities is

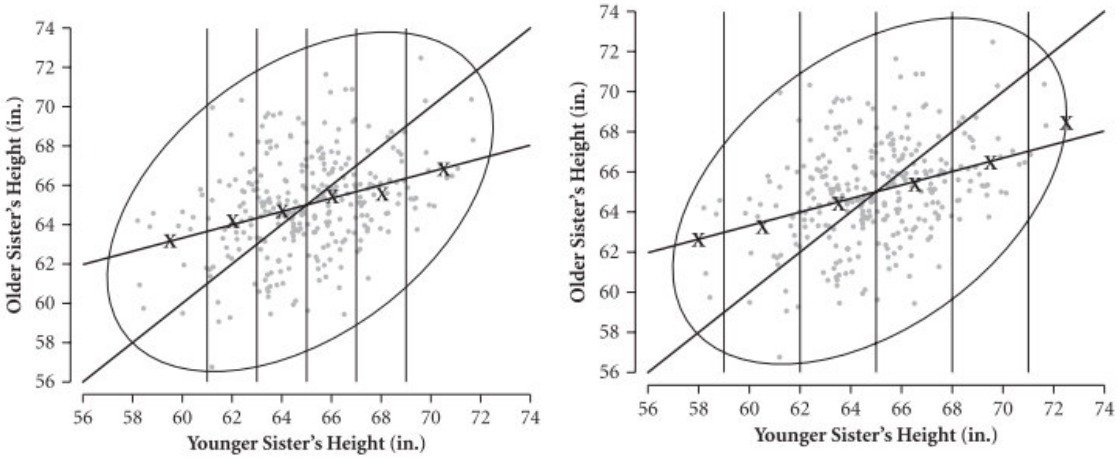
$$r^2 = \frac{\text{SST} - \text{SSE}}{\text{SST}} = \frac{480.25 - 212.37}{480.25} \approx 0.5578$$

(as given in the output) so $r = \pm 0.747$. Because the slope of the regression line is negative (the scatterplot goes downhill), $r = -0.747$.

b. The slope of –0.621 tells us that if one state has a high school graduation rate that is one percentage point higher than another, we expect its poverty rate to be lower by 0.621 percentage points, on average.

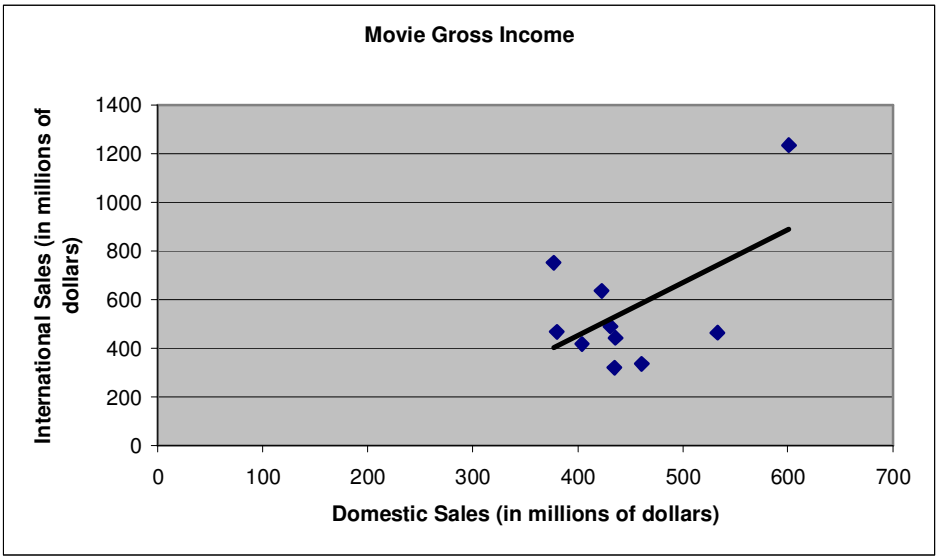
c. x is in percentage of high school graduates, y is in percentage of families living in poverty, b_1 is in percentage of families living in poverty per percentage of high school graduates, and r has no units.

P22. The plot should look similar to one of the two options shown here. The regression line is flatter than the line connecting the endpoints of the ellipse. This plot shows the regression effect as well. This time, it is the older sisters of the taller younger sisters who tend to be less tall than their younger sisters.



P23. There is evidence of regression to the mean. An ellipse around the cloud of points will have a major axis that is steeper than the regression line. The slope of the regression line is only about 0.4, much less than 1. Also, for the vertical strip containing exam 1 scores above 95, the mean exam 2 score is about 93. For exam 1 scores less than 70 the mean exam 2 score is about 76.

P24. a. Consider the following scatterplot with regression line:



The cluster of points has a decreasing trend except for one extreme value in upper right.

b. *International* = -417 + 2.17*Domestic*; $r = 0.56$

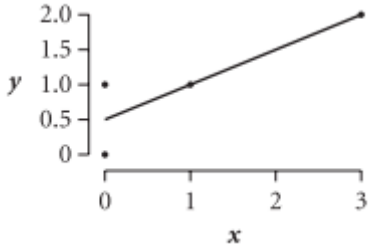
c. *International* = 979 – 1.16*Domestic*; $r = -0.4$. Removing the point allows the line to fit the cloud of points better, but reduces the r.

P25. a. The student did not predict very well. The estimates were consistently low.

b. (180, 350) appears to be the most influential point. It is an outlier in both variables and is not aligned with the other points.

c. With the point (180, 350) the regression equation is $actual = 12.23 + 1.92 \cdot estimate$, and $r = 0.975$. Without this point, the equation is $actual = -27.10 + 3.67 \cdot estimate$, with $r = 0.921$. This point pulls the right end of the regression line down, decreasing the slope and increasing the correlation.

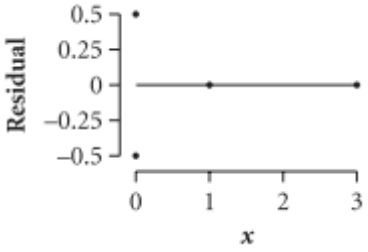
P26. a. The scatterplot with the regression line is shown next.



b. The table is as follows:

x	y	Predicted Value	Residual
0	0	0.5	−0.5
0	1	0.5	0.5
1	1	1	0
3	2	2	0

c. The residual plot is shown here.



d. The residual plot straightens out the tilt in the scatterplot so that the residuals can be seen as deviations above and below zero rather than above and below a tilted line. The symmetry of the residuals in this example shows up better on the residual plot.

P27. a. Population density appears to increase linearly, with little variation around a straight line.

b. Residuals show that the rate of increase in population density is not constant. In the 1800s the rate is a little lower than average, from 1900 to 1930 the rate is higher than average, lower again from 1930 to 1960 and higher since 1960.

P28. a. A–IV; B–II; C–I; D–III

b. Curve with increasing slope: The residual plot will open upward, as in plot II.

Unequal variation in the responses: The residual plot has a fan shape, as in plot I.

Curve with decreasing slope: The residual plot will open downward, like an inverted cup. No plot in this example shows this pattern.

Two linear patterns with different slopes: The residual plot will be V–shaped, as in plot III. The pattern is more clear if you ignore the point with a residual of about 10.

c. Plot D and residual plot III show a scatterplot that looks as though it should be modeled by two different straight lines. The V shape can be seen in the scatterplot, but it may be more obvious to many in the residual plot because the overall linear trend (tilt of the V) has been removed.

Exercise Solutions

E1. a. Positive and strong: As eggs get bigger, both length and width increase proportionally.
b. Positive and moderate: Most students tend to score relatively high on both parts of the exam, middling on both parts, or relatively low on both parts of the exam.
c. Positive and strong: Trees produce one new ring each year.
d. Negative and moderate: People tend to lose flexibility with age.
e. Positive and strong: The number of representatives is roughly proportional to the population of a state.
f. Positive and weak: Large countries tend to have large populations, but there are notable exceptions such as Canada and Australia. Also, some small countries in area have very large populations, such as Indonesia.
g. Negative and strong (but curved): Winning times tend to improve (get smaller) over the years.

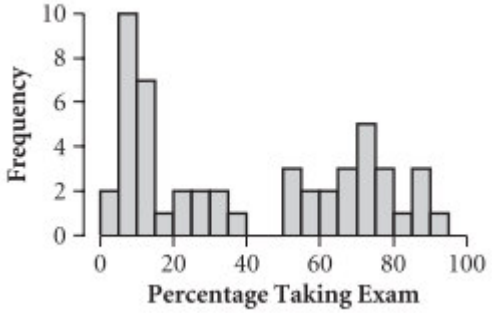
E2. A. II
B. IV
C. III
D. I (Heavier cars tend to get lower gas mileage.)

E3. a. The worst record for baggage handling during this period is Delta, and Northwest has the lowest on-time percentage among the airlines.
b. Airlines with a high percentage of on-time arrivals and a low rate of mishandled baggage would fall in the upper left of the plot. Thus, United is "best" for mishandled baggage; America West is best for percentage on-time arrivals.
c. False. American's baggage mishandling rate of 6.5 mishandled bags per thousand was not twice Southwest's rate of a little under 4.5. It appears to be more than twice because the scale on the *x*-axis starts at 3.75, not 0.
d. The relationship between the two variables is negative and moderate. The negative relationship shows that an airline that is "bad" on one variable tends to be "bad" on the other as well.
e. Yes, since an airline that is on-time a higher percentage of the time probably pays attention to other details as well (like handling baggage).

E4. a. Iowa 5% and Illinois 10%; about 87% of Maine students took the SAT, and they averaged 463.

b. The overall trend is negative, moderately strong, and curved. The gap between 40 and 50% in the middle of the scatterplot suggests two groups of states—one with low percentages and high average scores and another with high percentages and low average scores. There is one state that stands out a bit from the rest—West Virginia, with only 20% taking the SAT and a relatively low average of 511.

c. Yes, the distribution of the percentage taking the SAT looks bimodal because there is a cluster of percentages around 10 and a second around 65 to 75. Ask students to visualize all of the points dropping onto the *x*-axis so that they can see this distribution. A histogram of percentages is shown here. It makes a nice demonstration to display the scatterplot on Fathom and remove the *y* attribute and watch the dots drop to the x-axis.



The distribution of SAT scores also looks like it may be bimodal. There appear to be a cluster of scores around 510 and another cluster around 560. This time, ask students to visualize all the points dropping onto the *y*-axis so that they can see this distribution. (Again, the Fathom file makes a nice demonstration.)

The histogram of the average SAT scores is shown next. Even though you can see two peaks around 510 and 565, the shape is more skewed toward the larger values than bimodal.

d. There are no more states to add, so this is the complete picture for the given year. What you see is all there is. You might generalize, however, to the previous year and the next year. In fact, the plots for the last 20 or so years look similar to this one. These numbers do not change rapidly from year to year.

e. The Midwestern states are predominantly ACT states. In these states, only small percentages of students take the SAT, and these tend to be the better students who are trying for admission to exclusive colleges, perhaps outside the Midwest. If only a few students in a state are taking the SAT, they are probably the best students in the state and their average scores would then be higher than the average scores for other states. Thus, as the percentage of students taking the SAT increases, the average score tends to decrease. Although this explanation makes sense, we cannot be sure from these data alone.

E5. a. Plots A, B, and C are the most linear. Plot D is not linear because of the seven universities in the lower right, which may be different from the rest. The first plot, of *graduation rate* versus *alumni giving rate,* gives some impression of downward curvature. However, if you disregard the point in the upper right, the impression of any curvature disappears.

Plots A, B, and C, all have just one cluster. However, Plot D, the plot of *graduation rate* versus *top 10% in high school,* has two clusters. Most of the points follow the upward linear trend, but the cluster of seven points in the lower right with the highest percentage of freshmen in the top 10% shows little relationship with the graduation rate.

Plots A and C have possible outliers. Plot A, the plot of *graduation rate* versus *alumni giving rate* has a possible outlier in the upper right. It is below the general trend and its x-value (but not its y-value) is unusually large. In Plot C, the plot of *graduation rate* versus *SAT 75th percentile,* the points toward the upper left and the middle right should be examined because they are farther from the general trend than the other points, although neither their x-values nor their y-values are unusual. The point in the lower right of Plot D should also be examined, along with the other six points nearby.

b. The plots of *graduation rate* versus *alumni giving rate* and *graduation rate* versus *SAT 75th percentile* have similar moderate positive linear trends. The plot of *graduation rate* versus *top 10% in high school* shows wide variation in both variables, with little or no trend. The plot of *graduation rate* versus *student/faculty ratio* is the only plot that shows a negative trend.

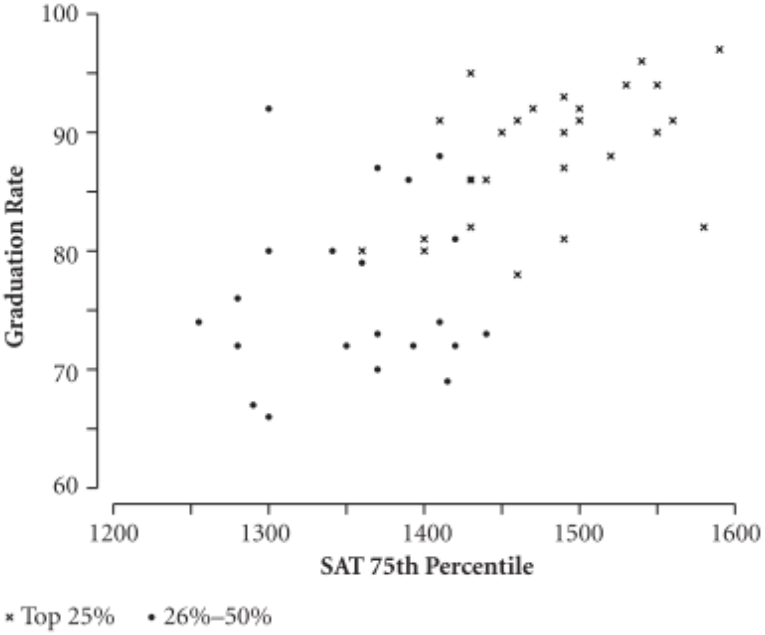
c. Among these four variables, it appears that the alumni giving rate is the best predictor of the graduation rate and SAT scores (as measured by the 75th percentile) is second best. However, both of these relationships are moderate and neither is a strong predictor of graduation rate. Ranking in high school class (as measured by the top 10%) is almost useless as a predictor of college graduation rate. The plot of *graduation rate* versus *alumni giving rate* owes part of the impression of a strong relationship to the point in the upper right. This plot shows some heteroscedasticity, with the graduation rate

varying more with smaller alumni giving rates.

Even though the relationship between, say, the graduation rate and the student/faculty ratio is negative, that's not what makes the student/faculty ratio a poor predictor of the graduation rate. Given a specific student/faculty ratio, we can predict a graduation rate. The problem is that there is a great deal of uncertainty about how close the actual rate would be to the predicted rate because the range of graduate rates is large for any given student/faculty ratio.

d. The relationships could change considerably when looking at all universities because these are highly rated universities, so the values of all variables tend to be "good." With a larger collection of universities, there may be more spread in the values of all variables and, most likely, more pronounced patterns.

Consider the scatterplot given below. On this plot, the **x**'s represent the top 25 universities and the closed circles represent the next 25 most highly rated universities. Note that if you look at either group, there is very little upward trend. However, putting the two groups together gives a stronger linear trend. If another group of 25 universities were added, the trend probably would be stronger. This phenomenon is sometimes called the *effect of a restricted range*.

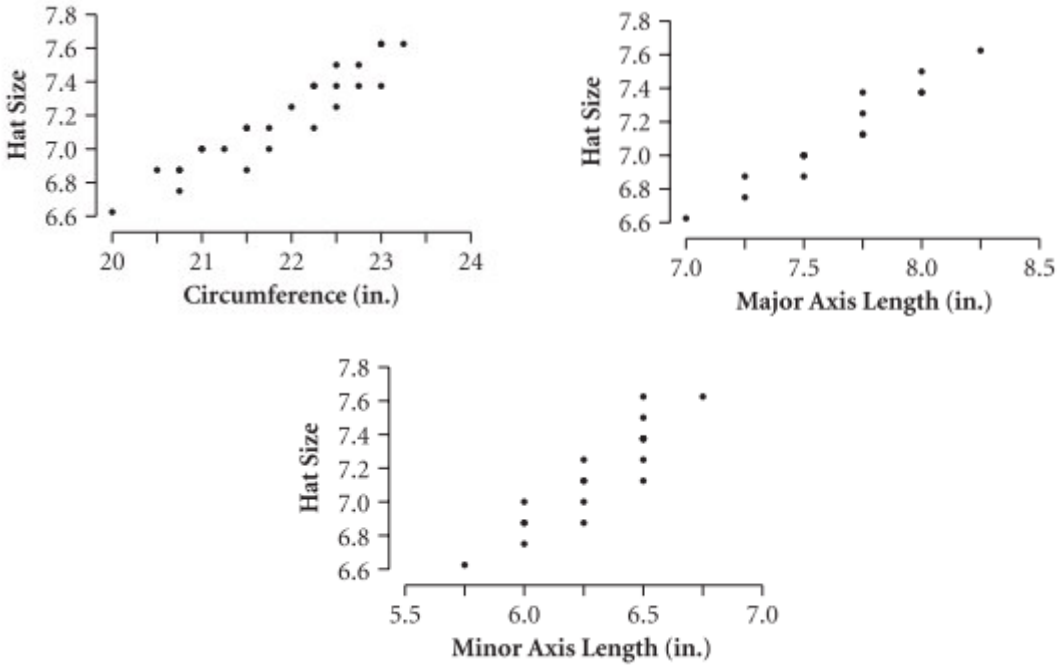


e. Graduation rates may increase as SAT scores increase because better prepared students may be more successful in college courses, so a university with a greater number of prepared students will, on average, graduate a higher percentage of students. Alumni giving rates may increase as graduation rates increase because the university has produced happy alumni. These data do not "prove" this claim, however, because there are other possible explanations. These types of observational studies cannot prove claims; the proof of a claim requires an experiment, which is one of the topics of the next chapter.

Note: These relationships hold for *universities* and that the data provide no evidence about whether the relationships hold for *individuals.*

E6. All three have positive association, but *circumference* appears to have the strongest positive association with *hat size.* Students may discover this rule: *hat size* is equal to

Note that the measurements tend to come aligned in vertical strips, indicating that the students made the measurements only to the nearest quarter of an inch.



E7. a. This plot does not help us decide this question. Because so many people who were hired in the early years (and even recently) would have retired, we do not know whether older people were hired then or not. To determine whether age discrimination in hiring may have existed, we need a plot of the age at hire of all people hired, not just those who remained at the time of layoffs.

b. From this plot, it appears that people hired earliest were more subject to layoff, not necessarily the older employees. Everyone with 30 years seniority was laid off, but not all of the older people were laid off. Perhaps, then, it was higher salaries because of seniority or obsolete job skills that resulted in a greater proportion of older employees being laid off.

E8. a. The basic trend is positive in that as the banking angle is increased, the TRAS generally increases (with the exception of two data points). For banking angles between 8 and 18 degrees, there are two clusters separated by 40 mph.

b. Generally, for a given banking angle, the TRAS is larger for longer tracks than it is for shorter ones.

c. The possible outlier is the track with banking angle of about 36 degrees and a TRAS of about 100.

E9. a. The scatterplot comprised of data which exhibit less variance is the one that provides a stronger relationship. In this sense, PPE does better, but neither does well.

b. Utah has a high graduation rate and low PPE.

c. The eastern states cluster to the left, high graduation rates and low student–teacher ratios; the Western states cluster to the right, high student–teacher ratios and large variation in graduation rates; the Midwestern states cluster in the upper middle while the Southern states cluster in the lower middle of the plot.

d. Mention that Delaware is at the median in graduation rate, in the upper quartile in PPE, and a little above the median on student-teacher ratio. Further reducing the student-teacher ratio can possibly help increase graduation rate.

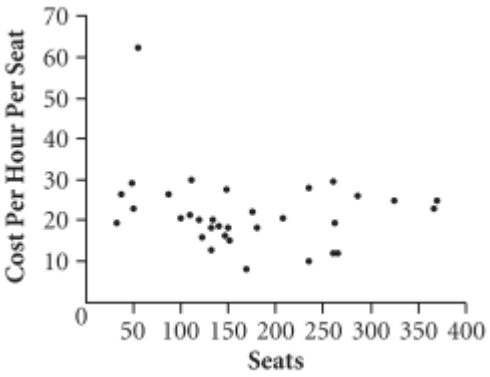
E10. a. *Cost per hour*
i. Students might point out that all of the associations are positive. As the variable on the *x*-axis increases, so does the cost per hour. Although all relationships are strong, the strongest are *cost per hour* versus *fuel consumption per hour* and versus *number of seats.* The other three relationships are weaker and show less constant strength. The scatterplot of *cost per hour* against *cargo space* is the most fan-shaped.

Relationships of *cost per hour* to *fuel per hour* and to *number of seats* are the most linear. The relationship of *cost per hour* to *speed* is the most curved. As *speed* increases, the *cost per hour* stays relatively constant up to about 460 miles per hour and then increases rapidly with increasing speed.

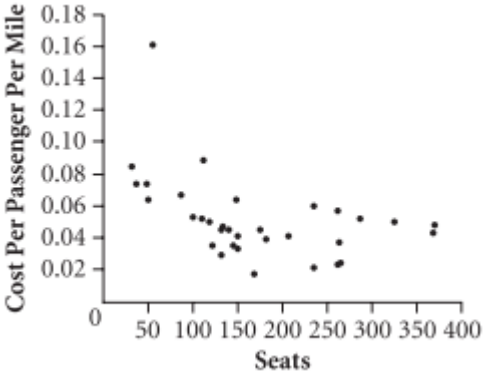
There appear to be no outliers in *cost per hour.* There is, however, one plane that is an outlier in *flight length,* the B747-400.

ii. It is true that the patterns in the plots show that bigger planes cost more per hour to operate, but this may be perfectly reasonable given that they carry more people and more cargo (and go faster).

One variable that might measure cost efficiency for the airplanes is *cost per hour per seat* (*cost/seats*). (Students may come up with others.) A plot showing this variable plotted against its denominator is shown in the text. Notice *cost per hour per seat* remains somewhat constant (but with a lot of variability) across the number of passengers carried. That is, larger planes tend to cost about the same to fly a passenger for an hour as smaller planes. However, larger planes also tend to go faster and take less time to travel the same number of miles. Considering that, larger planes may be more efficient.



The next scatterplot shows *cost per passenger mile* (*cost/hr,* divided by *speed* (in mph) divided by *seats* or *number of passengers carried*) versus *number of seats.* Now we see that larger planes do tend to be somewhat more cost efficient per passenger mile. However, the relationship is weak.



b. *Flight length*

i. Here are some things students might discover from their scatterplots or from the scatterplots in the fourth column of the scatterplot matrix:

Number of seats has a stronger relationship with *flight length* than does *cargo.* These are all passenger aircraft, and their primary purpose is to carry passengers. Cargo is added as space (and weight) is available. The weak relationship in the *cargo* versus *flight length* plot is due primarily to two aircraft, the large A300-600 (Airbus), used primarily to carry passengers and cargo over the short routes in Europe, and the B747-400, used to carry large numbers of passengers (and relatively less freight) over the very long international routes.

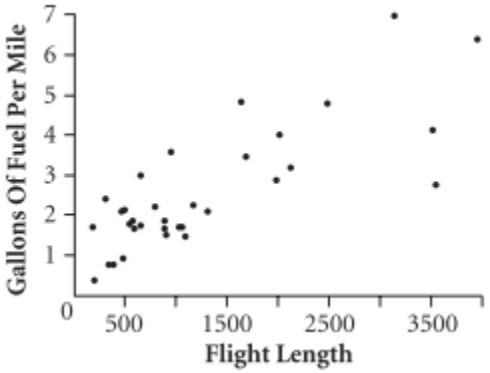
Planes with the longest flight lengths (the B747s) have the most seats but are not at the top of the cargo carriers. This can be seen in the plots of *seats* and *cargo* versus *length,* but you have to look back at the data to identify the planes.

A description of the scatterplot of *speed* versus *flight length* follows.

Cases and variables: The cases are the planes in the data set, and the variables are the airborne speed in miles per hour and the length of flight in miles.
Shape: The shape shows a single cluster of points in a thin, curved array that opens downward, with one point (the B747-400) as an outlier on the *length* axis.
Trend: The direction of the relationship is positive.
Strength: The relationship is very strong; a pattern is quite obvious.
Generalization: It seems reasonable that a similar pattern might appear even if other planes were added to the study. The general pattern of *speed* versus *flight length* should not depend entirely on the specific planes being studied here.
Explanation: As the typical length of flight increases, airlines tend to use faster planes because it is inefficient to use a fast plane on a short flight and inconvenient for travelers to use a slow plane on a long flight. But there is a maximum speed that can be achieved by the designs used for commercial aircraft so that few planes fly much over 500 miles per hour. This causes the leveling off of the plot for the longer flights.

ii. The faster planes used on the longer flights use more fuel per hour, but they cover many more miles in an hour than do the slower planes. So perhaps they use less fuel per mile.

To compute *gallons of fuel used per mile,* we must divide *fuel* in gallons per hour by *speed* in miles per hour. This plot shows *fuel consumption* in gallons per mile plotted against *flight length.*



Apparently, the planes that are capable of flying longer distances use more gallons of fuel per mile than do planes that fly shorter distances. This isn't surprising, as they carry more passengers and cargo. Further, flying faster may take more gallons per mile (as with an automobile).

c. *Speed, seats, and cargo*

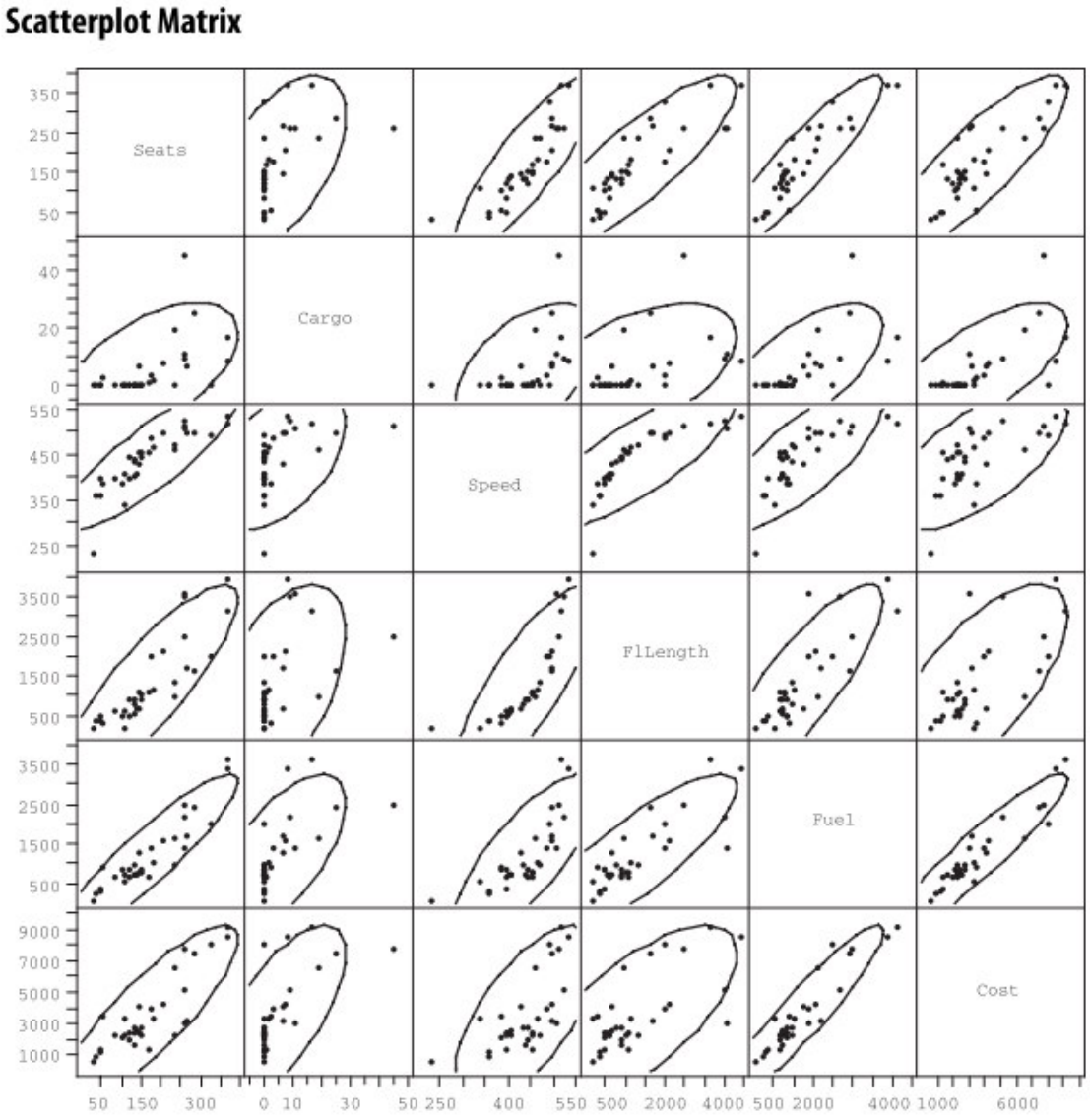
i. Here are some things you may discover:
The curvature is more pronounced in the relationship between *speed* and *cargo.* The slower (and smaller) planes carry little cargo, and the plane that carries the biggest cargo has only about medium speed. The planes that carry the biggest cargo carry only a

moderate number of passengers.

The A300-600 (Airbus) is unusually slow, both for the amount of cargo it carries (it has the largest cargo capacity of all the planes on the list) and the number of seats it has.

ii. The flat part to the left of the plot of *cargo* versus *seats* reveals that passenger planes with a relatively small number of seats (up to nearly 200) carry very little cargo. The fan shape to the right reveals that as the planes get bigger and the number of seats increases beyond 200, the amount of cargo carried by these planes also increases. However, the variation in the cargo-carrying capacity of a plane also increases as the number of seats increases.

Consider the following scatterplot matrix:



E11. a. I – E; II – C; III – A; IV – D; V – B
b. I – A;II – E; III – B; IV – D; V – C

E12. a. Pizza Hut's Hand Tossed and Little Caesar's Original Round have the fewest calories. They also have the least fat. The right side of the graph contains pizzas with the most fat.

b. i. E **ii.** D **iii.** A **iv.** C **v.** B

c. i.A: The line lies above all the points.
ii. E: The line lies below most of the points.
iii. B: The line lies over most points on the left and under most points on the right.
iv. D: The line lies under most points on the left and over most points on the right.
v. C: This line fits best overall, going through the middle of the points on both the left and right.

E13. a. The slope is about (67 – 37)/(14 – 2), or 2.5 inches/year. Or, 2.47 based on the regression line.

On the average, boys tend to grow about 2.5 inches per year from the ages of 2 to 14.

b. The *y*-intercept of 31.6 would mean that an average newborn is 31.6 inches long. Because this is clearly too long, this extrapolation is not valid.

E14. a. The calorie prediction for a pizza with 10.5 grams of fat is about 270 calories. the calorie prediction for a pizza with 15 grams of fat is about 335 calories.

b. The slope of the line is approximately 14.9 (from the regression line), which is the average increase in calories per unit increase in grams of fat.

c. The estimated slope is quite a bit higher than 9. There must be other ingredients that add calories and also increase as the fat content increases.

E15. a. The predictor variable is the arsenic concentration in the well water. The response variable is the concentration of arsenic in the toenails of people who use the well water.

b. There is a moderate positive linear relationship between arsenic concentration in the toenails of well water users and the arsenic concentration in the well water. There is a cluster around well water arsenic concentrations of 0 to 0.005 parts per million.

c. The residual for the person with highest concentration of arsenic in the well water is about 0.3 ppm.

d. This person has a concentration of about 0.076 ppm in their well water. The concentration of arsenic in this person's toenails is about 0.4 parts per million.

e. Seven of the 21 wells are above this standard.

E16. a. Fuel consumption rate is the explanatory variable. Operating cost is the response variable.

b. The slope is approximately 2.5. The operating cost tends to increase about $2.50 per hour, on average, for every one gallon per hour increase in fuel consumption. This could be the cost of one gallon of fuel.

c. This means that if an aircraft used no fuel, the cost per hour would be $470 per hour. While it doesn't make sense for an aircraft to use no fuel, it does make sense that there are costs besides fuel; the *y*-intercept would represent the cost per hour of running an aircraft in addition to fuel costs.

d. The cost per hour for a plane that consumes 1500 gallons per hour of fuel is approximately 470 + 2.50(1500) or $4220.

E17. a. In the order listed, Bristol, Martinsville, Richmond, New Hampshire, Daytona.

b. The residual is -64.13. It tells you that the track's TQS is much smaller than would be predicted from data on all tracks.

c. Talladega has a positive residual, its data point lies above the line. Richmond has a negative residual, its data point lies below the line

E18. a.

–40.58	Pizza Hut's Pan
–17.66	Domino's Deep Dish
–15.95	Pizza Hut's Hand Tossed
–1.03	Little Caesar's Original Round
14.28	Domino's Hand Tossed
26.44	Little Caesar's Deep Dish
34.50	Pizza Hut's Stuffed Crust

b. The residual of -40.58 shows that Pizza Hut's Pan Pizza has more than 40 fewer calories that would be predicted for a pizza with that pizza's amount of fat.

c. You can see that both are negative because both points are below the regression line.

E19. a. Observe that

$\bar{x} = 2,\ \bar{y} = 26$

$$b_1 = \frac{\sum (x-\bar{x})(y-\bar{y})}{\sum (x-\bar{x})^2} = \frac{(1-2)(31-26)+(2-2)(28-26)+(3-2)(19-26)}{(1-2)^2+(2-2)^2+(3-2)^2} = \frac{(-5)+0+(-7)}{1+0+1} = -6$$

$b_0 = \bar{y} - b_1\bar{x} = 26 - {}^{-}6 \cdot 2 = 38$

So, the equation is $\hat{y} = 38 - 6x$, where $\hat{y}$ is the predicted AQI and x is the number of years after 2000.

b. The AQI in Detroit tended to decrease 6 points per year on average.

c. The residuals for each year are calculated in the table below:

x	y	$\hat{y}$	Residual $= y - \hat{y}$
1	31	$38 - 6 \cdot 1 = 32$	-1
2	28	$38 - 6 \cdot 2 = 26$	2
3	19	$38 - 6 \cdot 3 = 20$	-1

The largest residual is for $x = 2$, which represents the year 2002. The residual is 2.

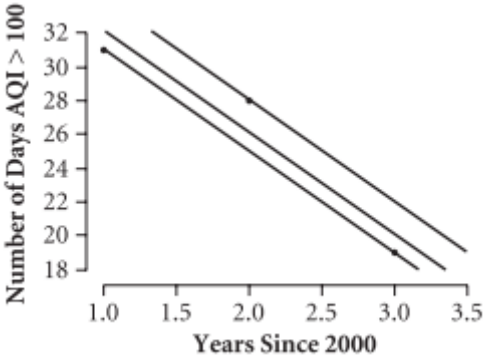
d. $\text{SSE} = (-1)^2 + 2^2 + (-1)^2 = 6$

e. $-1 + 2 + -1 = 0$

f. The equation for this line is y = 40 – 6x. (Since the residual for this point was 2, simply shift the line up two units.) SSE for this line would be $(-3)^2 + 0^2 + (-3)^2 = 18$. This is larger than the SSE for the least squares line, so according to the least squares approach the first line was better. Students should agree that the least squares line fits better because it passes through the middle of the set of points. The line through the point for 2002 is too high.

g. This equation would be y = 37 – 6x. The fitted value for 2002 would be $37 - 6 \cdot 2 = 25$. The only nonzero residual would be for this point, so SSE would be $3^2 = 9$.

h. As the plot below shows, the least squares line is a better indicator of the trend of all the points than either of the others because it represents the overall trend of the data. For both of the other lines, all points are on or to the same side of the line.



E20. a. Observe that $\overline{x} = 13.1$, $\overline{y} = 307.143$. Also,

x (Fat)	y (Calories)	$x - \overline{x}$	$y - \overline{y}$	$(x - \overline{x})(y - \overline{y})$	$(x - \overline{x})^2$
9.0	230	−4.1	−77.143	316.2863	16.81
19.5	385	6.4	77.857	498.2848	40.96
14.0	280	0.9	−27.143	−24.4287	0.81
12.0	305	−1.1	−2.143	2.3573	1.21
8.0	230	−5.1	−77.143	393.4299	26.01
14.2	350	1.1	42.857	47.1427	1.21
15.0	370	1.9	62.857	119.4283	3.61
				$\sum(x - \overline{x})(y - \overline{y})$ ≈ 1352.5	$\sum(x - \overline{x})^2$ ≈ 90.62

Using this information, we obtain

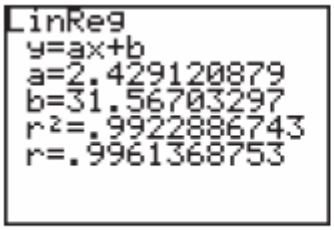
$$b_1 = \frac{\sum (x-\overline{x})(y-\overline{y})}{\sum (x-\overline{x})^2} = \frac{1352.5}{90.62} = 14.92$$

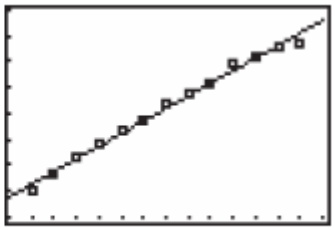
$$\overline{y} = b_0 + b_1\overline{x} \;\Rightarrow\; 307.143 = b_0 + 14.92 \cdot 13.1 \;\Rightarrow\; b_0 = 111.7.$$

So the equation is $calories = 111.7 + 14.92\,fat$.

b. From the scatterplot, Pizza Hut's Pan pizza has a residual of about 40. From this point alone, the SSE must be at least $40^2 = 1600$. Thus, 4307 is the only possible SSE.

E21. a. The equation is $height = 31.6 + 2.43 \cdot age$. Here is the plot.





[1, 15, 1, 30, 7, 5]

b. The residual appears to be positive because the point appears to be above the line. The actual residual is 59.5 – 58.33 = 1.17, which is positive. (The residual is 1.21 with no rounding.)

c. The mean age is 8 years, and the mean height is 51 inches. Substitute into the regression equation to see that this point is on the regression line:

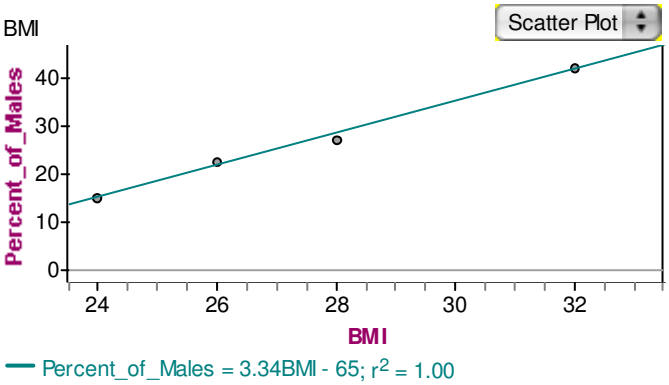
$$51 \neq 31.6 + 2.43 \cdot 8 = 51.04$$

This small discrepancy is due to rounding error. To avoid this, use the stored values for a and b from your calculator. For the TI-83 Plus and TI-84 Plus they can be found in the variables menu (VARS). Select **5:Statistics…** then right arrow to **EQ**. The **a** and **b** listed are the values from the most recently calculated regression equation. When those values

are used, the right side of the equation comes out to 51 exactly.

d. This equation is very close to that for boys obtained earlier.

E22. a. & b.

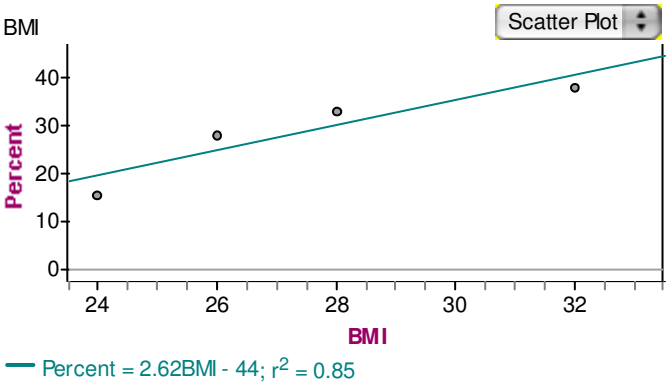


The regression line is (*Percent of Males*) = 3.34(BMI) – 65. The slope of this line, 3.34, is the increase in percentage of males with hypertension per unit increase in BMI.

c. Note that $\bar{x} = 27.5$, $\bar{y} = 26.475$. Substituting $\bar{x}$ in for BMI and $\bar{y}$ in for *Percent of Males* yields a true statement.

d. Substituting BMI = 30 into the equation in (a) yields *Percent of Males* = 35.2. So, we predict that 35.2% of males with a BMI of 30 to have hypertension. Similarly, substituting BMI = 36 into this equation yields *Percent of Males* = 55.24%. Given how tightly this line fits the data, this wouldn't be unreasonable. But, keep in mind that the further you get from values used to generate the line, the less confident you can be in the predictive value of the regression line.

e.



The strength of the linear relationship is still strong, but it does seem as though an upside down wide U-shaped curve would fit better. However, the curved nature of the data might simply be due to noise in measurement.

E23. a. *Age* is the explanatory variable and *area* is the response variable.

b. The line has slope close to 1; in comparing 10 and 20 year-old trees, there areas are also about 10 and 20 square inches.

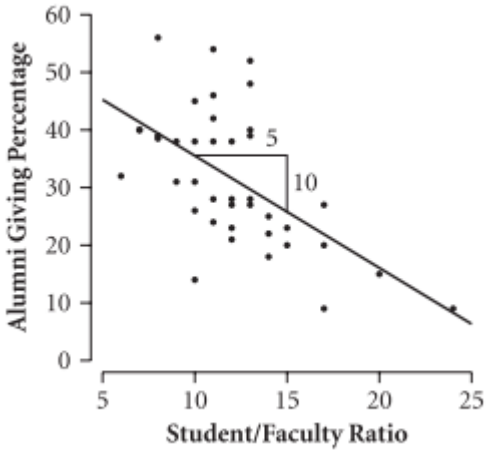
c. 30 year-olds would have larger error because there is more variation in the residuals near 30.

d. Near 0 (0.38); its predicted area almost equals its actual area.

e. Age 38 and the area is about 20; it has a much smaller area compared to other trees of its age.

E24. a. The alumni giving rate (in percent) is the response variable, and the student/faculty ratio is the predictor or explanatory variable.

b. As you can see from the plot shown here, a run of 4 in the direction of the *x*-axis corresponds to a drop of about 8 percentage points in the direction of the *y-axis*.



c. The *y*-intercept would mean that a university that has a student/faculty ratio of 0 (i.e., no students) would have a giving rate of 55%. This makes no sense. Extrapolation is not reasonable in this case.

d. The giving rate would be about 31%. The error probably is rather large because the points are not clustered closely about the line. There is quite a lot of variation around the line, especially for the universities with smaller student/faculty ratios. On the average, a prediction would be off by about 7 percentage points.

e. The point is just above the line, and the residual is about 1 or 2. The residual for the point with the highest giving rate is about 56 – 40 = 16.

f. $y = 55 - 2(6) = 43$. The residual is $32 - 43 = -11$.

g. Because the line has a negative slope, the largest possible predicted giving rate occurs when the student/faculty ratio is as small as possible. The smallest the student/faculty ratio could be is 0 (if there were no students). This ratio gives a predicted giving rate of 55 – 2(0) = 55. This is the largest possible predicted giving rate. The rate at Piranha State is larger than 55%, so the residual for Piranha State will be positive.

E25. a. $height = 31.5989 + 2.47418 \cdot age$. Answers will vary according to students' original estimates. It should be fairly close.

b. SSE = 2.4. This does seem reasonable from the scatterplot. The residuals are all quite small.

E26. a. $percent\ alumni\ giving = 54.979 - 1.9455 \cdot student\ /\ faculty\ ratio$**.**

b. The largest residual is 22.31. The student/faculty ratio for that university is 13.

c. The fit should be $54.979 - 1.9455 \cdot 13 = 29.6875$. The table has the fit calculated correctly. The actual *y*-value (also given in the table) is 52. The residual, 52 – 29.69 = 22.31, is calculated correctly in the table.

d. SSE = 3578.5. The value is large because there are many points and they are scattered widely around the regression line.

E27. For these proofs to make sense, you will need the following properties of finite sums:
(i) For a constant *a*, $\sum a = n \cdot a$, since that means adding up *n* instances of *a*,
(ii) The mean of a set of numbers is a constant, so $\sum \overline{x} = n \cdot \overline{x}$ where $\overline{x}$, in turn, is equal to the sum of the individual *x* values $\sum x_i$

a. A horizontal line has an equation of the form *y* = *a*. The residual for a point (x_i, y_i) would be $y_i - a$. Adding these up for *n* points, we obtain

$$\sum (y_i - a) = \sum y_i - \sum a = \sum y_i - na.$$

We want this sum to be zero.

$$\sum y_i - n \cdot a = 0$$
$$\sum y_i = n \cdot a$$
$$\frac{\sum y_i}{n} = \frac{n \cdot a}{n}$$
$$\frac{\sum y_i}{n} = a$$

And, of course, the left side of this equation is the mean of the *y*-values. This horizontal line has equation $y = \overline{y}$ so it passes through $(\overline{x}, \overline{y})$.

Conversely, if we start with the fact that the horizontal line passes through the point $(\bar{x}, \bar{y})$, the equation of the line is $y = \bar{y}$. The residual for a point (x_i, y_i) would be $y_i - \bar{y}$. Summing these residuals we obtain

$$\sum (y_i - \bar{y}) = \sum y_i - \sum \bar{y} = \sum y_i - n \cdot \bar{y} = \sum y_i - n \frac{\sum y_i}{n} = \sum y_i - \sum y_i = 0.$$

b. For a point (x_i, y_i), the predicted *y*-value is $a + b \cdot x_i$. So the residual is $y_i - a - b \cdot x_i$. We want the sum of these residuals to be zero.

$$\sum (y_i - a - b \cdot x_i) = 0$$
$$\sum y_i - n \cdot a - \sum b \cdot x_i = 0$$
$$\sum y_i = n \cdot a + b \cdot \sum x_i$$
$$\frac{\sum y_i}{n} = \frac{n \cdot a}{n} + \frac{b \cdot \sum x_i}{n}$$
$$\bar{y} = a + b \cdot \bar{x}$$

This means that the point $(\bar{x}, \bar{y})$ must satisfy the equation of the line if the residuals are to sum to zero.

Conversely, if we start with the fact that the line passes through $(\bar{x}, \bar{y})$, that means the equation $\bar{y} = a + b \cdot \bar{x}$ must be true. So $a = \bar{y} - b \cdot \bar{x}$. The residual for the point (x_i, y_i) would be $y_i - (\bar{y} - b \cdot \bar{x}) - b \cdot x_i = y_i - \bar{y} + b \cdot \bar{x} - b \cdot x_i$. The sum of these residuals then is

$$\sum y_i - n \cdot \bar{y} + n \cdot b \cdot \bar{x} - b \cdot \sum x_i = n \cdot \bar{y} - n \cdot \bar{y} + n \cdot b \cdot \bar{x} - n \cdot b \cdot \bar{x} = 0$$

c. As shown in parts a and b, *any* line that passes through the point $(\bar{x}, \bar{y})$ makes the sum of the residuals equal to zero.

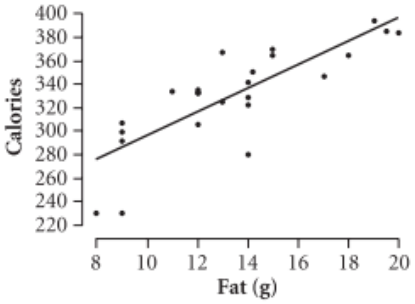
E28. a. Yes; the ratio of price to gallons is the price per gallon, which is the same for all four purchasers. Another way to look at this is to observe that the price, *y,* increases at a constant rate for each additional gallon of gas purchased.

b. The relationship between *average speed x, time y,* and *distance* is (*speed*)(*time*) = *distance,* or $x \cdot y = 80$ in this scenario. Thus, plotting *y* (*time*) against *x* (*speed*) means plotting $y = \frac{80}{x}$, or $time = \frac{80}{speed}$, which will be a curve (it's a hyperbola) opening upward. Plotting $y^* = \frac{1}{y}$ against *x* results in a straight line because

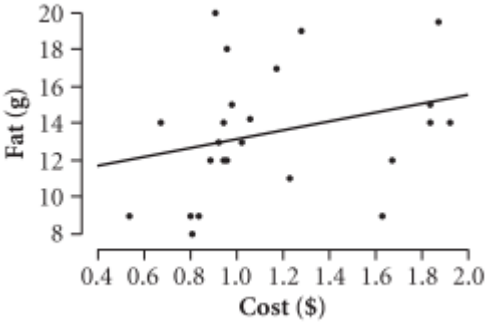
$$80 = xy$$
$$\frac{1}{y} = \frac{1}{80} x$$
$$y^* = \frac{1}{80} x$$

which is a linear equation.

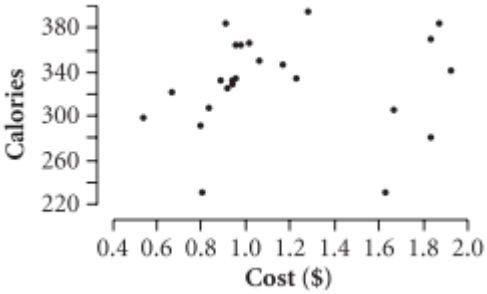
E29. a. The plot is shown here. This association shows a positive trend that is moderately strong. Even though a few points lie relatively far from the pattern, fat content could be used as a reasonably good predictor of calories. The equation of the regression line is $\hat{y} = 195 + 10.05x$, where $\hat{y}$ is the predicted number of calories and x is the number of grams of fat. The slope means that for every additional gram of fat, the pizzas tend to have 10.05 additional calories, on average.



b. The plot is shown next. There is possibly a very weak positive association between *fat* and *cost.* The equation of the regression line is $\hat{y} = 10.7 + 2.41x$, where $\hat{y}$ is the predicted number of grams of fat and x is the cost. The slope means that for every additional dollar in cost, the number of grams of fat tends to increase by 2.41 grams, on average. (You will be able to check to see if this association is "real" in a subsequent chapter.)



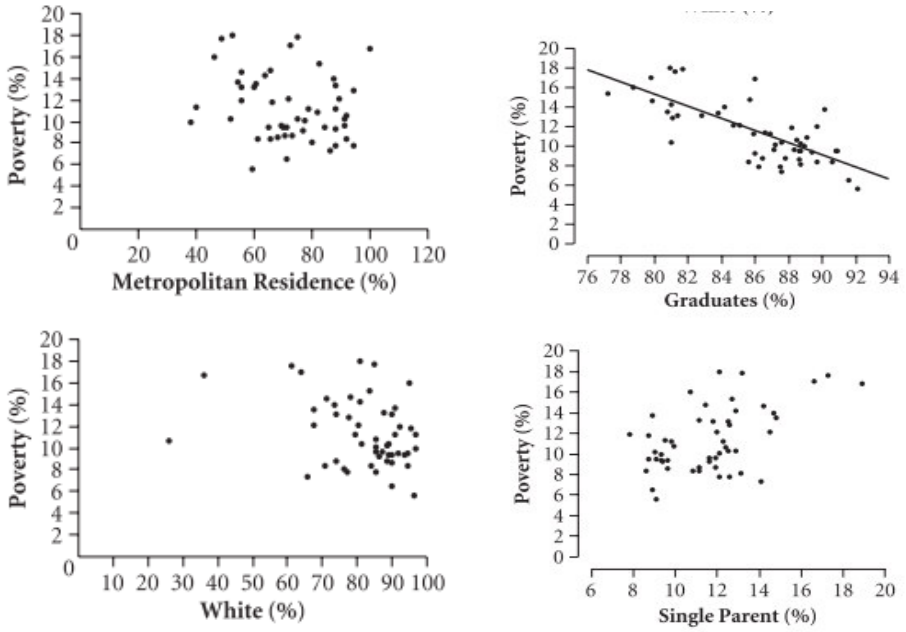
c. The plot is shown here. There appears to be no linear association between calories and cost.



d. In the analysis of the pizza data, *calories* has a moderately strong positive association with *fat*—the two tend to rise or fall together. This makes sense because fat has a lot of calories. *Fat* has a weak positive association with *cost.* There appears to be no association between *cost* and *calories.*

E30. The four scatterplots needed are shown next. There appears to be little or no association between the percentage living below the poverty line and the percentage living in metropolitan areas. Likewise, there appears to be little association between the poverty rate and the percentage of whites. The two outliers in this plot are regions with low percentages of whites, namely, Washington, D.C. (high poverty rate) and Hawaii (low poverty rate).

The poverty rate is negatively associated with the percentage of high school graduates; as the latter goes up, the percentage living in poverty generally goes down. Finally, poverty appears to be only weakly associated with percentage of families headed by a single parent (with Washington, D.C., again as an outlier). On the surface, it looks as though increasing graduation rates would have the largest effect on decreasing the poverty rate. Keep in mind, however, that the problem of poverty is much more complex than that, and many other variables are lurking in the background. Association is not the same as cause and effect. Simply increasing high school graduation rates, although it might be a good thing to do, will not automatically elevate all of those living below the poverty line to a better economic condition. It is even possible that a lower poverty rate may cause a higher high school graduation rate, or that a third 'lurking' variable is influencing both of these variables.



E31. The correlations of the scatterplots are

a. 0.66 **b.** 0.25 **c.** –0.06 **d.** 0.40 **e.** 0.85 **f.** 0.52
g. 0.90 **h.** 0.74

E32. **a.** about –0.37. An estimate between –0.50 and –0.25 is a good one.

b. about 0.65. An estimate between 0.50 and 0.80 is a good one.

c. about 0.53. An estimate between 0.40 and 0.65 is a good one.

E33. This problem isn't as much work as it looks like. Notice that for the first four tables, the means are all 0 and the standard deviations are all 1, so they have already been standardized. You can find the average product in your head. Table e has the same correlation as table a, table f has the same as table c, table g has the same as table b, and table h has the same as table c.

a. 1 **b.** 0.87 **c.** –0.5 **d.** –1 **e.** 1 **f.** –0.5
g. 0.5 **h.** –0.5

E34. Note that

$$\bar{x} = 150,\ \bar{y} = 13.5,\ s_x = \sqrt{6250} \approx 79.057,\ s_y = \sqrt{33} \approx 5.745 .$$

Next, consider the following table of values:

x	$z_x = \frac{x-150}{79.057}$	y	$z_y = \frac{y-13.5}{5.745}$	$z_x \cdot z_y$
50	-1.265	6.0	-1.305	1.651
100	-0.632	11.0	-0.435	0.275
150	0	12.5	-0.174	0
200	0.632	17.0	0.609	0.385
250	1.265	21.0	1.305	1.651

So, $r = \frac{1}{5-1}\sum z_x \cdot z_y = 0.9905$.

E35. a. About 0.95.

b. All but three of the points lie in quadrants I and III (based on the 'origin' $(\bar{x}, \bar{y})$.) In Quadrant I, z_x and z_y are positive and in Quadrant III, both , z_x and z_y are negative. In either of these quadrants the products $z_x \cdot z_y$ are positive. Thus, the correlation is positive.

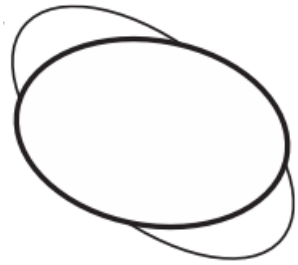
c. The point in the lower left corner of Quadrant III makes the largest contribution. It is the most extreme point for both the x and y values, giving it the largest (in absolute value) z–score for both variables.

d. The point just below $(\bar{x}, \bar{y})$ makes the smallest contribution. z_x and z_y are both near zero, so the product of these z–scores will be quite small.

E36. a. I. B II. C III. A

b.

i.



ii.

E37. a. No because the units will be different. For example, for the group that measures in chirps per second and uses temperature for *x,* the units of the slope will be chirps per second per degree temperature. For the group that measures in chirps per minute, the units will be chirps per minute per degree temperature. So the slope for the second group should be 60 times that of the first group. For a group that measures in chirps per minute and uses chirps for *x,* the units of the slope will be degrees temperature per chirps per minute. Even if they use the same units, groups that interchange *x* and *y* will get different slopes (chirps per minute per degrees Centigrade, or degrees Centigrade per chirps per minute).

b. Yes. The correlation is the same no matter what the units or what you use for *x* and for *y*. That is because *r* is equal to the average product of *z*–scores, which have no units and $z_x \cdot z_y = z_y \cdot z_x$.

E38. a. An estimate of the slope is

$$b_1 = r\frac{s_y}{s_x} = -0.5 \cdot \frac{0.083}{4.3} = -0.00965.$$

To find the *y*–intercept, use the fact that the point $(\overline{x}, \overline{y}) = (11.7, 0.827)$ is on the regression line:

$$y = \text{slope} \cdot x + y - \text{intercept}$$
$$0.827 = -0.00965(11.7) + y - \text{intercept}$$
$$y - \text{intercept} = 0.940$$

The equation is $\hat{y} = -0.00965x + 0.940$.

b. An estimate of the slope is

$$b_1 = r\frac{s_y}{s_x} = -0.5 \cdot \frac{4.3}{0.083} = -25.90$$

To find the *y*–intercept, use the fact that the point $(\overline{x}, \overline{y}) = (0.827, 11.7)$ is on the regression line:

$$y = \text{slope} \cdot x + y - \text{intercept}$$
$$11.7 = -25.90(0.827) + y - \text{intercept}$$
$$y - \text{intercept} = 33.12$$

The equation is $\hat{y} = -25.90x + 33.12$.

E39. a. True. This is a direct result of the formula

$$b_1 = r\frac{s_y}{s_x}$$

If $s_x < s_y$, the factor on the right, s_y/s_x, will be greater than 1, which makes b_1 greater in absolute value than *r*.

b. The formula

$$1.6 = 0.8 \cdot \frac{s_y}{s_x}$$

can be true only if $s_x = 25$ and $s_y = 50$.

c. $b_1 = r\dfrac{s_{estimated}}{s_{actual}} \Rightarrow 0.36 = r \cdot \dfrac{4.12}{0.93} \Rightarrow r = 0.081$

d. $b_1 = r\dfrac{s_{actual}}{s_{estimated}} \Rightarrow b_1 = 0.081 \cdot \dfrac{0.93}{4.12} \Rightarrow b_1 = 0.0183$

E40. An estimate of the slope of the regression line is

$$b_1 = r \cdot \frac{s_y}{s_x} = 0.8 \cdot \frac{8}{30} \approx 0.21$$

To find the *y*–intercept, use the fact that $(\overline{x}, \overline{y}) = (280, 75)$ is on the regression line.

$$75 = b_0 + 0.21 \cdot 280$$

$$b_0 = 16.2$$

The regression equation is $\hat{y} = 16.2 + 0.21 \cdot x$. So, Julie's final exam score is predicted to be $16.2 + 0.21 \cdot 300 = 79.2$.

E41. a. A large brain helps animals live smarter and therefore longer. The lurking variable is overall size of the animal. Larger animals tend to develop more slowly, from gestation to "childhood" through old age, and larger animals tend to live longer than smaller ones.

b. If we kept the price of cheeseburgers down, college would be more affordable. The lurking variable is inflation over the years—all costs have gone up over the years.

c. The Internet is good for business. The lurking variable is years. Stock prices generally go up due to inflation over the years. The Internet is new technology, and so the number of Internet sites also is increasing over the years.

E42. a.

(*calories, fat*) 0.95
(*calories, saturated fat*) 0.95
(*calories, sodium*) –0.5
(*fat, saturated fat*) 0.95
(*fat, sodium*) –0.5
(*saturated fat, sodium*) –0.5

b. If you use more salt in your food, it will reduce the calories and the fat.

c. The lurking variable is whether the cheese is low–fat. Fat makes cheese taste good (and adds calories). In order to make up for this loss of flavor in low–fat cheese, manufacturers may add more salt.

E43. a.

$$r^2 = \frac{\text{SST} - \text{SSE}}{\text{SST}} = \frac{0.088740 - 0.016304}{0.088740} \approx 0.8163$$

So r =0.903. The value of r^2 above is equal to the "R–sq" in the regression analysis.

b. The largest residual occurs for the point (55, 0.318). Its value is

$$y - \hat{y} = 0.318 - [0.202 + 0.00306(55)] = 0.318 - 0.3703 = -0.0523.$$

c. Yes. Hot weather causes people to want to eat something cold. But the correlation alone does not tell us that.

d. Degrees Fahrenheit, pints per person, pints per person per degree Fahrenheit, and no units.

e. "MS" was computed by dividing "SS" by "DF."

E44. a. From the table, SST = 71.833857 and SSE = 23. 335476. So,

$$r^2 = \frac{\text{SST} - \text{SSE}}{\text{SST}} = \frac{71.833857 - 23.335476}{71.833857} \approx 0.675.$$

This value is somewhat different from the value of 0.7548 given in the table.

b. The value of the residual that is largest in absolute value occurs for Chicago, at O_3 = 0.032. This value is:

$$\left|8.3 - \left(-2.452415 + 277.02991 \times 0.032\right)\right| = 1.887.$$

c. Lurking variables are almost always present. Other environmental factors can be at play here, possibly obfuscating the true relationship.

d. The quantities x, y, and b_1 have units ppm, while r has no units.

e. Root mean square error is $\sqrt{\sum \frac{(y-\hat{y})^2}{n}}$. This is an estimate of the average of the residuals.

f. They seem to be clustered above and below the line with very few points near the actual line. This, coupled with the rather weak correlation coefficient, suggests that a linear relationship is perhaps not the best fit.

E45. The center of this cloud of points should not be modeled by a straight line. This can be considered as a plot with curvature, or as a plot with two very influential points in the upper right corner. In either case, some adjustments should be made to the data before attempting to fit a line to it. One possibility is to transform the data by techniques you will learn in Section 3.5. Because r^2 is a measure of how closely the points cluster about the regression line, it would not make sense in this context.

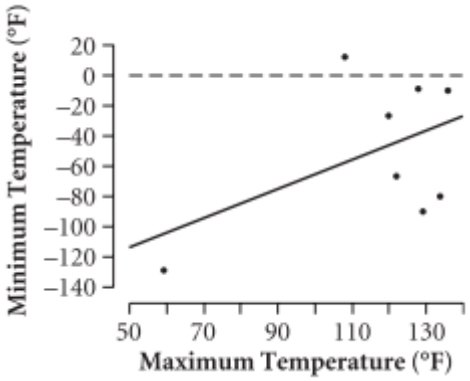
E46. It has a positive relationship, curved like the left portion of an upside down U. The correlation coefficient would be reasonably large, say around 0.80, but the fit to a curve rather than a straight line would probably provide a better description of the data.

E47. Scoring exceptionally well, for example, on a test involves more than just studying the material. There is a certain amount of randomness involved, too—the instructor asked questions about what the student knew, the student was feeling well that day, the student was not distracted, and so on. It's unlikely that this combination of knowing the material and good luck will happen again on the next test for this same student. The student probably will get a lower, but still high, score on the next test even if he or she doesn't slack off. However, it would appear as if doing well the first time and getting praised prompted the student to relax and study less. On the second test, the student's place at the top of the class may be taken by another student who knew just as much for the first test but was also effected by randomness on the first test—perhaps unlucky in the questions the instructor chose or unlucky in another way at the time.

At the other end, a student who scores exceptionally poorly on the first test also has a bit of randomness involved—bad luck this time. Whether or not he or she is praised, the student scoring exceptionally poorly probably won't have all of the random factors go against him or her on the next test and the student's score will tend to be higher.

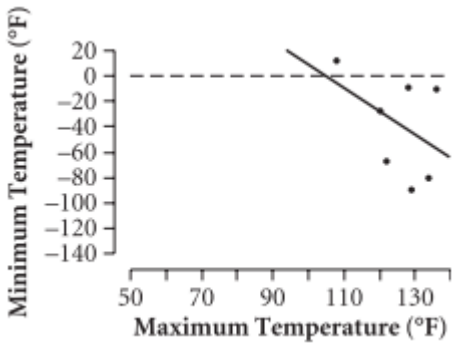
E48. This is an example of the regression effect. The explanation is similar to that of E47. The same students are not likely to score at the top of their class two times in a row.

E49. a. The scatterplot appears here (with the regression line). A line is not a good model because the cloud of points is not elliptical with one extremely influential point.



b. The regression equation is $\hat{y} = -161.90 + 0.954x$, and $r = 0.49$.

c. With Antarctica removed, the slope of the regression line changes from positive (0.954) to negative (–1.869) and the correlation becomes negative, $r = -0.45$. The plot appears as shown here. (Notice that a new potential influential point has appeared: Oceania.) Without a plot of the data, you might come to the following incorrect conclusion: In general, continents tend to be "warm" or "cold"; that is, continents with higher maximums also tend to have higher minimums. In fact, there is little relationship.



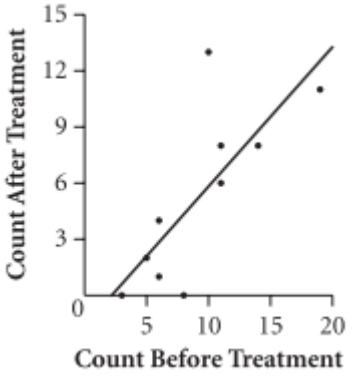
E50. a. The point (0.137, 2.252) appears to be the most influential because it is an outlier in both variables. It appears that if this point were eliminated the right end of the line would drop, decreasing the slope. The correlation would probably also decrease.

Eliminating (0.137, 2.252) does indeed decrease the slope of the regression line from 13.0 to 7.46. The correlation also decreases from 0.896 to 0.634.

b. A point near the regression line and near $(\overline{x}, \overline{y})$ is likely to have little effect on the slope or correlation. One such point would be (0.0194, 0.517). Eliminating this point actually leaves the regression equation almost the same and very slightly increases the correlation from 0.896 to 0.897.

c. Removing the point with the largest residual in absolute value (0.0764, 0.433) would cause a slight increase in slope but a large increase in correlation. The actual result is to cause the slope to increase from 13.0 to 15.4 and the correlation to increase from 0.896 to 0.969.

E51. a. The plot with the regression line and the regression summary are shown here. The equation of the line is $\hat{y} = -1.63 + 0.745x$.

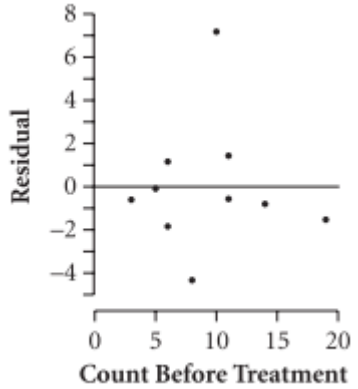


```
Dependent variable is:   y
No Selector
13 total cases of which 3 are missing
R squared = 58.4%   R squared (adjusted) = 53.2%
s = 3.177 with 10 − 2 = 8 degrees of freedom

                Sum of
Source          Squares   df   Mean Square   F-ratio
Regression      113.348   1    113.348       11.2
Residual        80.7516   8    10.0939

                                  s.e. of
Variable   Coefficient            Coeff     t-ratio   prob
Constant   −1.63057               2.299     −0.709    0.4984
x          0.745223               0.2224    3.35      0.0101
```

Here is the completed table:

x	*y*	Predicted Values	Residual
11	6	6.567	−0.567
8	0	4.331	−4.331
5	2	2.096	−0.096
14	8	8.803	−0.803
19	11	12.529	−1.529
6	4	2.841	1.159
10	13	5.822	7.178
6	1	2.841	−1.841
11	8	6.567	1.433
3	0	0.605	−0.605

b. The residual plot shown here is a little unusual in that it shows more variability in the middle than at either end. But this is partly because there are more cases in the middle.



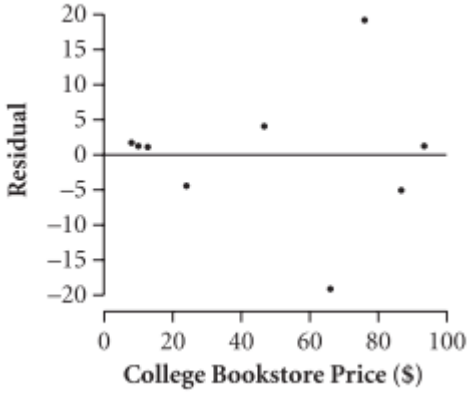
c. The disinfectant appears to be unusually effective for the person with the large negative residual, the point (8, 0) on the original scatterplot. It is seemingly ineffective for the person with the large positive residual, the point (10, 13) on the scatterplot.

E52. a. The slope of the line is very close to 1, and the *y*–intercept is –3.57. Because the slope is about 1, the *y*–intercept means that textbooks bought online tend to cost about $3.57 less than those bought at the college bookstore. (In fact, the mean cost of the college textbooks is $47.04, and the mean cost of the online textbooks is $45.03. Their difference is not $3.57 because the slope is not exactly equal to 1.) The slope is 1.03 which means that for every one dollar increase in price for a book sold through the college bookstore, there tends to be a $1.03 increase, on average, for the same book bought online.

b. The table for computing the residuals is shown here.

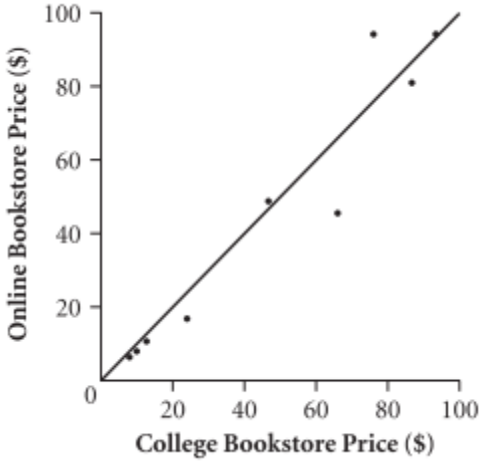
College	Online	Fitted	Residual
93.40	94.18	92.9281	1.2519
9.95	7.96	6.7092	1.2508
46.70	48.75	44.6786	4.0714
76.00	94.15	74.9508	19.1992
86.70	80.95	86.0058	−5.0558
7.95	6.36	4.6428	1.7171
24.00	16.80	21.2254	−4.4254
12.70	10.66	9.5505	1.1095
66.00	45.50	64.619	−19.119

The next residual plot shows that a line is a reasonable model for these data. The points are scattered randomly above and below 0, except that the points fan out to the right. This pattern indicates that the points lie farther from the regression line as the prices increase. It is easy to spot the two textbooks whose prices vary the most between the two types of bookstores. The calculus textbook is the one with residual about 19, meaning the online price is about $19 more than the price in the college bookstore. The art history textbook is the one with residual about –19, meaning the online price is about $19 less than the price in the college bookstore.



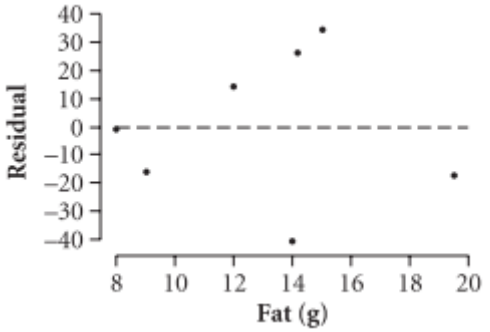
c. The plot with the line $y = x$ is shown next. A point above the line represents a textbook that costs more at the online bookstore. A point below the line represents a point that costs less at the online bookstore. A point on the line represents a book that costs the

same at both stores.

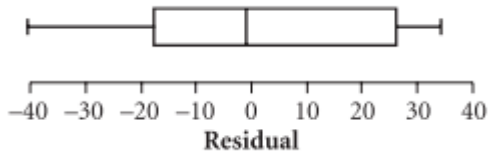


d. The boxplot is centered slightly above 0, indicating that most college bookstore prices tend to be slightly higher than the online prices. However, most of the differences are close to 0, so there is little difference in price. The median difference is about $2 more for the college bookstore. There are two outliers shown, which means that for two textbooks, the prices vary greatly, in one case being less expensive at the college bookstore and in the other case being less expensive at the online bookstore. The overall lesson is that with more expensive textbooks, it pays to shop around.

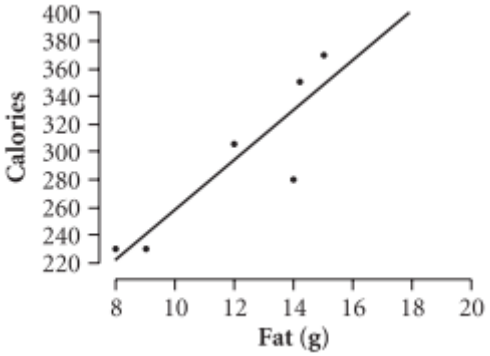
E53. a. Here is what the residual plot for these data actually looks like.



b. The largest positive residual belongs to Pizza Hut's Stuffed Crust, which has more calories than would be predicted from a simple linear model. The largest negative residual belongs to Pizza Hut's Pan Pizza. None appear so far away that it would be called an exception, or outlier. In fact, you can check this by making a boxplot of the residuals, as shown here.



c. The complete data set shows a moderately strong positive trend with a slope of about 14.9 calories per gram of fat and correlation of 0.908. The most influential data point would be the one farthest away from the main cloud of points on the *x*–axis (Domino's Deep Dish). Removing Domino's Deep Dish yields a slope of around 18 calories per gram of fat and correlation of 0.893. None of the other points have nearly as much influence on the slope.



E54. a. The pattern of the scatterplot is basically linear, so the slope is constant across the numbers of seats.

b. The spread in the flight lengths increases as the number of seats increases. The points fan out to the right.

c. This is a good bet only in the first case. It would be easier to predict flight length for planes with fewer numbers of seats because there is less variation in flight lengths for the smaller planes than for the larger. When the number of seats is between 50 and 150, the values for flight length vary between about 175 to about 1065 miles whereas when the number of seats is between 200 and 300, the values for flight length vary between 947 and 3559.

d. The residual plot for this scatterplot also fans out (spread out more) as the number of seats increases. In fact, the fan shape may be seen better in the residual plot.

E55. A. I **B.** IV **C.** III **D.** II A linear model would be appropriate for C and D. Both C and D show a random scatter of points around the residual, but the slope of the regression line is almost zero for plot C, and there appears to be no correlation. Plot B does not have a random scatter around the line; the pattern appears to be cyclical. This is typical of a situation in which something changes approximately linearly over time. What happens next usually depends on what just happened, causing this up–and–down pattern in the residuals.

E56. No, for the plot on the left. The pattern is impossible for residuals because the major linear trend shown here would already have been removed by fitting the regression line. The residuals show what is left over after any linear trend is removed from the data.

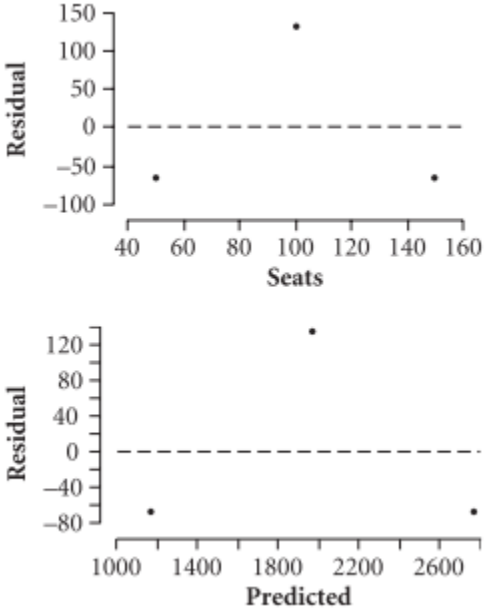
No, for the plot on the right. These residuals do not have a mean of zero. Residual plots

may, however, show a curved pattern that could come from a scatterplot that wrapped around a line. The plot shown in the student text is an altered residual plot with the vertical scale slightly changed.

E57. a. The regression equation is $\hat{y} = 366.67 + 16x$. The completed table is shown here.

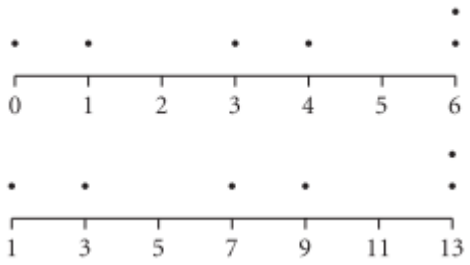
Aircraft	Seats	Cost	Predicted	Residual
ERJ–145	50	1100	1166.67	−66.67
DC9	100	2100	1966.67	133.33
MD–90	150	2700	2766.67	−66.67

b. The plot of the residuals versus x (seats) is given first followed by the plot of the residuals versus $\hat{y}$. The only difference between the two plots is the scaling on the horizontal axis.



E58. The fitted value $\hat{y}$ is a linear transformation of x; that is, $\hat{y} = a + bx$. Thus, using $\hat{y}$ rather than x on the horizontal axis does not change the relative distance of the values from each other—it just translates and stretches the horizontal axis.

Consider this example for a regression line $\hat{y} = 1 + 2x$ with values of $x = 0, 1, 3, 4, 6, 6$. Then the values of $\hat{y}$ are 1, 3, 7, 9, 13, 13. These two sets of points are plotted on horizontal scales in the next plot. Note that the relative spacing of the points is exactly the same on each scale.

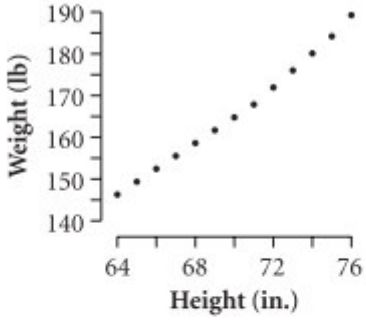


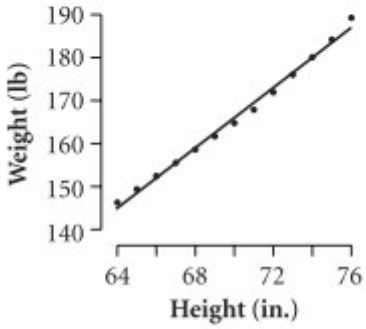
If the regression line has a negative slope, then the larger y's correspond to the small x's and one residual plot will be the mirror image of the other.

E59. Scatterplots of the original data, without and with the regression line, follow the commentary.

To estimate the recommended weight for a person whose height is 64 inches, add the fitted weight (given to be 145 pounds) to the residual of about 1.2 to get 146.2. The slope of the regression line must be about (187 – 145)/(76 – 64) = 3.5. For the second height, 65 inches, the fitted weight would be 145 + 3.5(1) = 148.5. The residual is about 0.9 pounds. Thus, the recommended weight is about 148.5 + 0.9 = 149.4 pounds. You could continue point by point to get the next plot, but a rough sketch can be obtained by making use of the linear patterns in the residuals (and hence in the original scatterplot). The points on the scatterplot must form a straight line up to a height of 71, where the weight must be about 145 + 7(3.5) – 1.5 = 168.0.

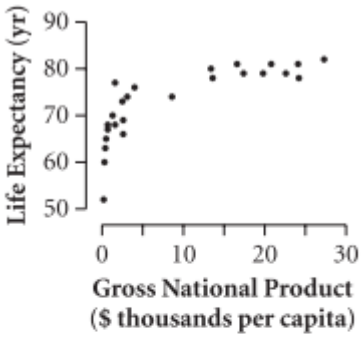
The remainder of the points must form (approximately) another straight line up to a height of 76, where the weight must be about 145 + 12(3.5) + 2.2 = 189.2.

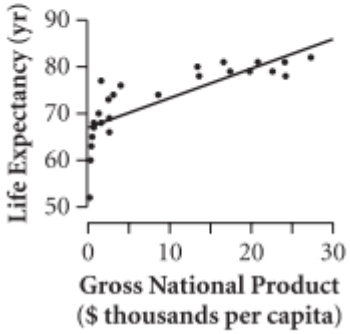




It is difficult to see the strong V shape of the residuals in a scatterplot drawn on the actual scale of the data.

E60. Scatterplots of the data, without and with the regression line, are shown here. Begin by drawing in the regression line. Then use the residual plot to determine how far above or below the regression line each point should be placed. Note that a linear model is not a good one for predicting life expectancy from GNP.





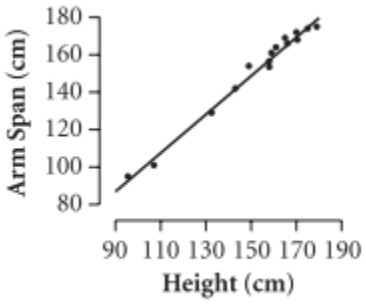
E61. a. If Leonardo is correct, the data should lie near the lines:

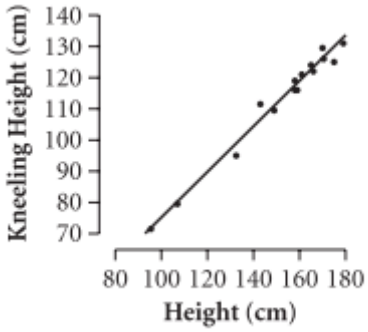
$$arm\ span = height$$

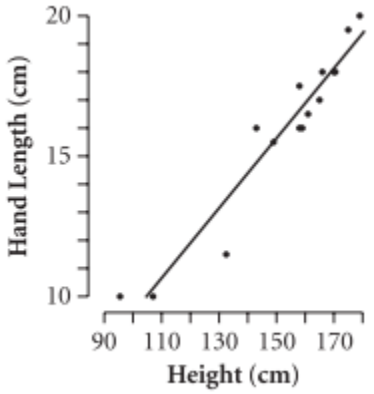
$$kneeling = \frac{3}{4}\ height$$

$$hand = \frac{1}{9} height$$

Looking at the next plots, these rules appear to be approximately correct.







The least squares regression equation for predicting the arm span from the height is:

$$\hat{y} = -5.81 + 1.03x.$$

The least squares regression equation for predicting the kneeling height from the height is:

$$\hat{y} = 2.19 + 0.73x.$$

The least squares regression equation for predicting the hand length from the height is:

$$\hat{y} = -2.97 + 0.12x.$$

b. For the first plot, the slope is 1.03. This means for every 1 cm increase in *height,* there tends to be a 1.03 cm increase in *arm span.* Leonardo predicted a 1 cm increase.

For the second plot, the slope is 0.73, which means that for every 1 cm increase in *height,* there tends to be a 0.73 cm increase in *kneeling height.* Leonardo predicted a 0.75 cm increase.

For the third plot, the slope is 0.12, which means that for every 1 cm increase in *height,* there tends to be a 0.12 cm increase in *hand length.* Leonardo predicted a $\frac{1}{9} \approx 0.11$ cm increase.

Leonardo's claims hold reasonably well. The slopes are about what he predicted, and the y-intercepts are close to 0 in each case.

c. In each case, the points are packed tightly about the regression line and so there is a very strong correlation. The correlations are

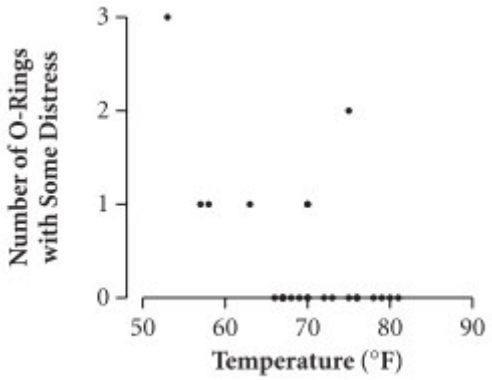
arm span and *height:* 0.992 (strongest)
kneeling height and *height:* 0.989
hand length and *height*: 0.961 (weakest)

E62. a. This scatterplot shows no obvious association between temperature and number of distressed O-rings. Although the highest number of distressed O-rings was at the lowest temperature, the second highest number was at the highest temperature.

b. It is difficult to look at the scatterplot of the complete set of data, shown next, and not see that any risk is almost entirely at lower temperatures. The correlations are

$r = -0.263$ for the incomplete set of data but $r = -0.567$ after all points are included, which is only moderately strong at any rate.

This is a tragic example of scientists and engineers not asking the right question: Do I have all of the data?



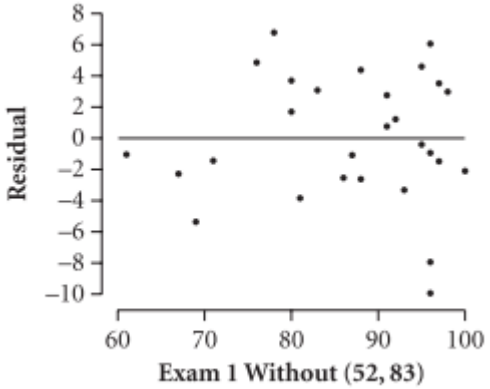
Background information: After each launch, the two rocket motors on the sides of the *Challenger* were recovered and inspected. Each rocket motor was made in four pieces, which were fit together with O-rings to seal the small spaces between them. The O-rings were 37.5 feet in diameter and 0.28 inch thick.

The Rogers Commission, which was appointed by President Reagan to find the cause of the accident, noted that the flights with zero incidents were left off the plot because it was felt that these flights did not contribute any information about the temperature effect. The Commission concluded that "A careful analysis of the flight history of O-ring performance would have revealed the correlation of O-ring damage in low temperature" [*Report of the Presidential Commission on the Space Shuttle Challenger Accident* (Washington, D.C. 1986), page 148.].

E63. a. Yes, the student who scored 52 on the first exam and 83 on the second lies away from the general pattern. This student scored much higher on the second exam than would have been expected. This point will be influential because the value of x is extreme on the low side and the point lies away from the regression line. The point sticks out on the residual plot. There is a pattern in the rest of the points; they have a positive correlation.

b. The slope should increase, and the correlation should increase. In fact, the slope increases from 0.430 to 0.540, and the correlation increases from 0.756 to 0.814.

c. The residual plot appears next. The residuals now appear scattered, without any obvious pattern, so a linear model fits the points well when point (52, 83) is removed.



d. Yes, there is regression to the mean in any elliptical cloud of points whenever the correlation is not perfect. For example, the student who scored lowest on Exam 1 did much better on Exam 2. The highest scorer on Exam 1 was not the highest scorer on Exam 2. A line fit through this cloud of points would have slope less than 1.

E64. a. Use the formula

$$b_1 = r\frac{s_y}{s_x}$$

to obtain

$$r = 0.51\left(\frac{11.6}{7.0}\right) \approx 0.845.$$

b. Using the fact that $(\overline{x}, \overline{y})$ is on the regression line,

$$\hat{y} = \overline{y} + b_1(x - \overline{x}) = 87.8 + 0.51(x - 82.3) = 45.83 + 0.51x.$$

E65. a. Each value should be matched with itself.

b. Match each value with itself, except match 0.5 (or –0.5) with 0: (–1.5, –1.5) (–0.5, –0.5) (0, 0), (0, 0.5), (0.5, 0), (1.5, 1.5) for a correlation of .950.

c. (–1.5, 0) (–0.5, 0.5) (0, 1.5), (0, –1.5), (0.5, –0.5), (1.5, 0) has a correlation of –.1.

d. Match the biggest with the smallest, the next biggest with the next smallest, etc.: (–1.5, 1.5) (–0.5, 0.5) (0, 0), (0, 0), (0.5, –0.5), (1.5, –1.5).

E66. Do not actually compute the correlations to answer these questions.

a. Zero correlations occur between A and B, B and E, B and F, C and D, and E and F. The correlation between D and E is .02. The pairs with zero correlation all have scatterplots that are symmetric around a center vertical line or center horizontal line or both.

b. A and E, D and F, and A and C

c. A and D, A and F, B and C, and C and F

d. Common responses may include the following: The scatterplot of A and B shows that there can be a pattern in the points (a shape) even though the correlation is zero. The fact that the scatterplots of A and D and of A and F have about the same correlation again shows that the correlation does not tell anything about their quite different shapes.

For your information, the complete correlation matrix is shown here.

	A	B	C	D	E
B	0				
C	0.447	0.258			
D	0.224	0.129	0		
E	0.875	0	0.091	0.018	
F	0.258	0	0.289	0.577	0

E67. a. True. Both measure how closely the points cluster about the "center" of the data. For univariate data, that center is the mean; for bivariate data, the center is the regression line.

b. True. Refer to E66 for examples.

c. False. For example, picture an elliptical cloud of points with major axis along the *y*-axis. The correlation will be zero, but there will be a wide variation in the values of *y* for any given *x*.

d. True. Intuitively, a positive slope means that as *x* increases, *y* tends to increase. This is equivalent to a positive correlation. Similar statements can be made for a negative slope and a zero slope. Alternatively, the fact that this statement is true can be seen from the relationship

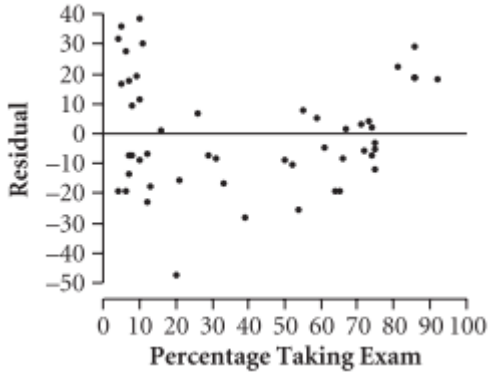
$$b_1 = r\frac{s_y}{s_x}$$

Because the standard deviations are always positive, b_1 and r must have the same sign.

E68. a. The correlation is quite high, about –0.84.

b. There are two clusters of points—one of states with a small percentage of students taking the SAT and one of states with more than 50% taking the SAT. The second cluster has almost no correlation and would have a relatively flat regression line, whereas the first cluster has a strong negative relationship. Combining these two clusters results in summary statistics that do not adequately describe either one of them.

c. A residual plot would be U-shaped with points above zero on the left, below zero in the middle, and above zero again on the right. The actual residual plot is shown below.



E69. a. Public universities that have the highest in-state tuition also tend to be the universities with the highest out-of-state tuition, and public universities that have the lowest in-state tuition also tend to be the universities with the lowest out-of-state tuition. This relationship is quite strong.

b. No, the correlation does not change with a linear transformation of one or both variables. However, if you were to take logarithms of the tuition costs, the correlation would change.

c. The slope would not change with the first transformation. Consider the formula for the slope:

$$b_1 = r\frac{s_y}{s_x}$$

The correlation remains unchanged with the change of units. The standard deviations would each be $\frac{1}{1000}$ as large as previously, but the factor of $\frac{1}{1000}$ would be in both the numerator and denominator and so would cancel out. But if you were to take logarithms of the tuition costs, the slope would change because the proportion s_y/s_x would be different.

E70. a. Quadrant I: +, +, +; Quadrant II: – ,+, –;
Quadrant III: –, –, +: Quadrant IV: +, –, –

b. The points near the origin or near one of the new axes make the smallest contributions. The contributions are small because either z_x or z_y would be close to 0.

E71. You can compute the values of r using the formula

$$b_1 = r\frac{s_y}{s_x}$$

Solving for r, the formula becomes

$$r = b_1\frac{s_x}{s_y}$$

These correlations are A: 0.5; B: 0.3; C: 0.25. So from weakest to strongest, they are ordered C, B, A.

E72. **a.** No. It is possible for one bookstore to be a lot more expensive than the other. For example, suppose your local bookstore sold each book for $10 less than the online price. The correlation would be 1.

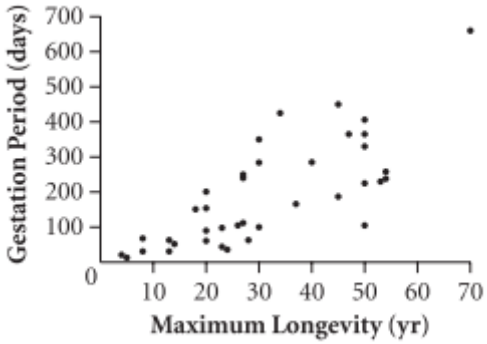
b. The main reason for the high correlation is that the bookstores pay approximately the same wholesale cost for a book. They then add on an amount to cover their overhead costs and to give them a profit. To have a cause-and-effect relationship, a change in one variable should trigger a change in the other variable. That is not necessarily the case with these prices; however, if the online bookstore lowers its prices, it might force local bookstores to do the same.

E73. For example, stocks that do the best in one quarter may not be the ones that do the best in the next quarter. As another example, the best 20 and worst 20 hitters this year in major League Baseball are not likely to repeat this kind of performance again next year.

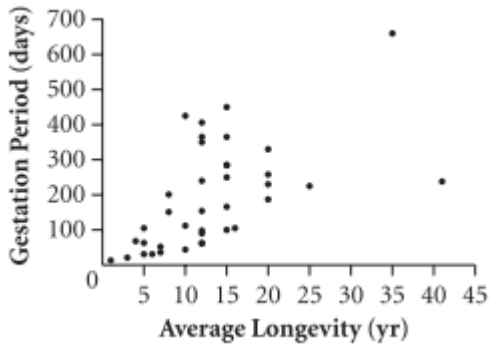
E74. This matrix gives the correlations between all pairs of variables in this exercise.

	Ave Long	Max Long	Gestation
Max Long	0.769		
Gestation	0.577	0.761	
Speed	−0.215	−0.237	0.018

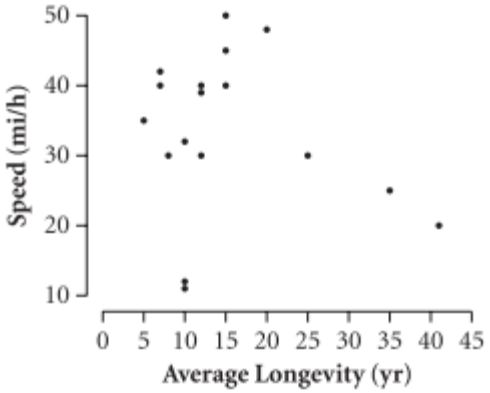
a. In general, animals with longer maximum longevity have longer gestation periods. The trend is reasonably linear, with a moderate correlation. The plot shows heteroscedasticity, with the points fanning out as maximum longevity increases. The elephant is an outlier, although it follows the general linear trend.



b. The pattern here is similar: The animals with longer average longevity have longer gestation periods. However, the relationship is not as strong as in part a, so maximum longevity is the better predictor of gestation period. There are two outliers, the elephant and the hippopotamus.



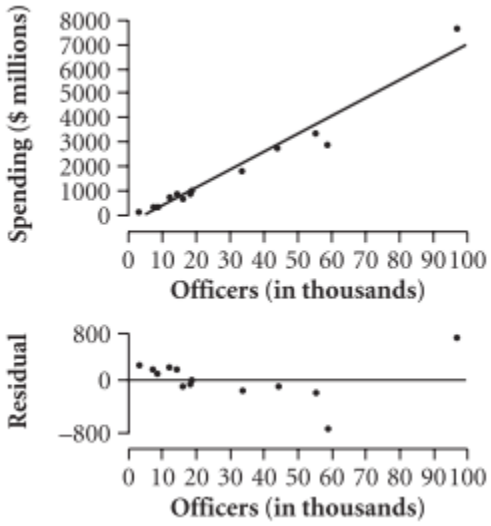
c. The three relatively slow animals (elephant, hippopotamus, and grizzly bear) in the lower-right corner give these data a slight negative correlation of –0.215. However, the rest of the animals show a positive trend, with longer average longevity associated with greater speed. The lurking variable is the size of the animal. Larger animals tend to live longer and be faster (unless they get very big, like an elephant).



E75. a. The next scatterplot shows a very strong linear relationship between the expenditures for police officers and the number of police officers per state. This makes sense. There is one outlier and influential point, California, which has by far the largest population of any state listed. The correlation is 0.976, and the equation of the regression line is

$$expPolice = -403 + 73.7(number\ of\ police)$$

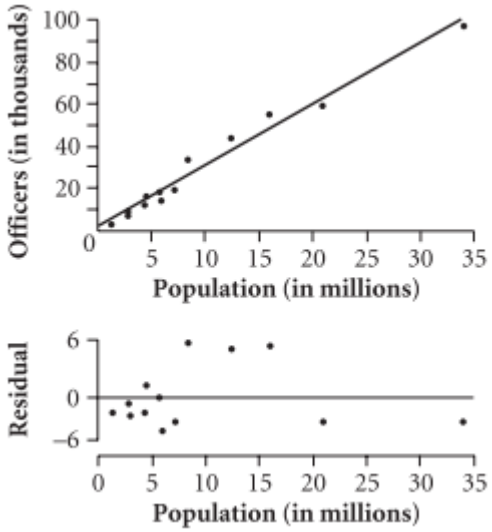
So, for every additional thousand police officers, costs tend to go up by $73,700,000, or $73,700 each. If the influential point of California is removed, the slope of the line decreases but the correlation increases.



b. The scatterplot again shows a very strong linear relationship, with California as an outlier and influential point. The larger the population of the state, the more police officers. This time the correlation is 0.987, and the equation of the regression line is

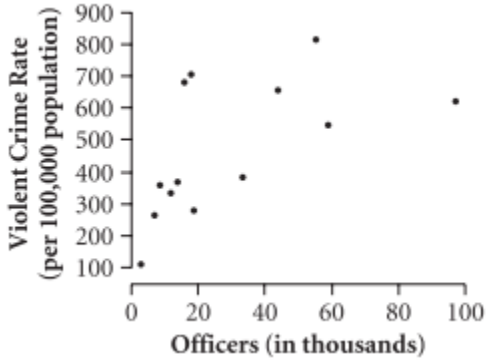
$$\textit{number of police} = 1.5 + 2.91\ \textit{population}$$

That is, for every increase of 1 million in the population, the number of police officers tends to go up by 2910. If the influential point of California is removed, the slope of the line increases a little and the correlation decreases a little.

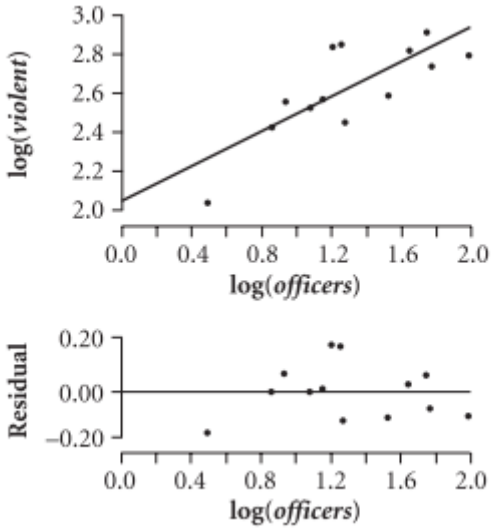


c. The scatterplot shows that there is a moderate positive but possibly curved relationship. For this scatterplot, it is not appropriate to compute the correlation or equation of the regression line. But, in general, the larger the number of police, the higher the rate of violent crime. (Note that this is the rate per 100,000 people in the state, not the number of violent crimes.) Almost equivalently, the larger the population of a state, the higher the violent crime rate. It is not at all obvious why this should be the case. Why

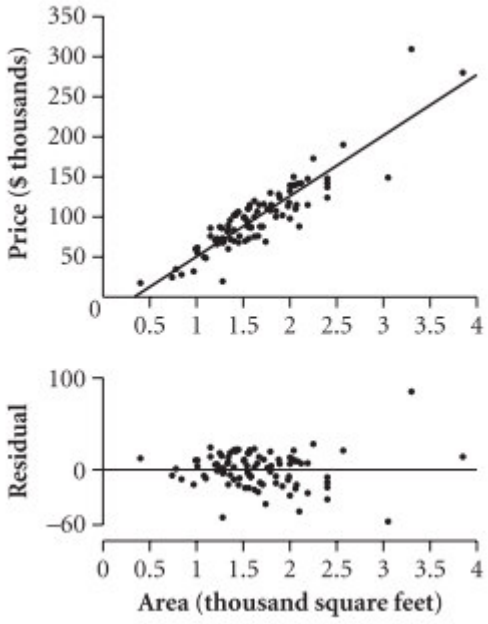
would larger states (more police = more population) tend to have higher *rates* of violent crime? (Because of the strong relationship between the number of police officers per state and the population, a scatterplot of the crime rate versus the number of police officers per 100,000 population looks about the same.)



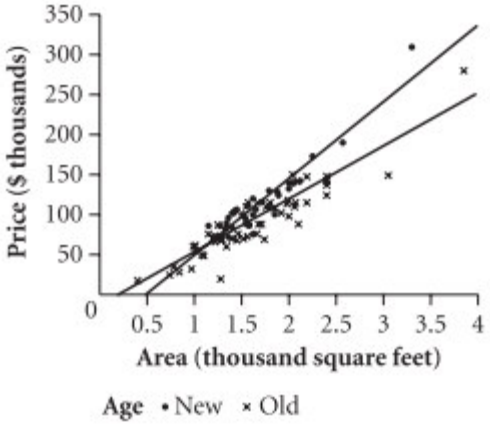
A log-log transformation of the number of police straightens these data quite well, as shown here in the scatterplot and residual plot.



E76. a. A linear model works fairly well, as shown here in the scatterplot and residual plot. There is some heteroscedasticity, however, so the residuals for houses with a large number of square feet are larger. The equation is *price* = –25.2 + 75.6 *area,* and the correlation is 0.899.



Separating the houses into new houses and old houses, the corresponding regression equations are *price* = –48.4 + 96.0 *area* and *price* = –16.6 + 66.6 *area.* These equations are very different, so you should not use one equation for both groupings. New houses cost quite a bit more per square foot. You can see the two relationships in this plot.

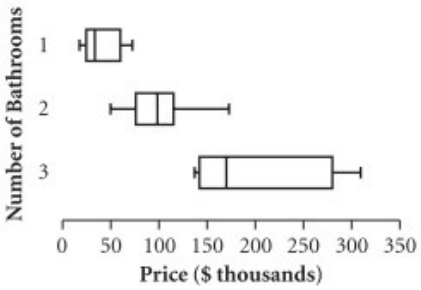


b. The two largest houses clearly are influential points, as could be the third largest house and the smallest house in the lower-left corner. If the two largest are removed from the data set, the correlation drops to .867 and the regression line changes to *price* = –10.3 + 65.7 *area.* This is quite a change in the model—the price is now increasing $10,000 less per increase of 1,000 square feet.

c. Using the equation from part a, the price for an old house of 1,000 square feet would be *price* = –16.6 + 66.6 *area* = –16.6 + 66.6(1) = 50, or $50,000. The price for a house of 2,000 square feet would be *price* = –16.6 + 66.6 *area* = –16.6 + 66.6(2) = 116.6, or $116,600. You should have more confidence in the first prediction because the spread in

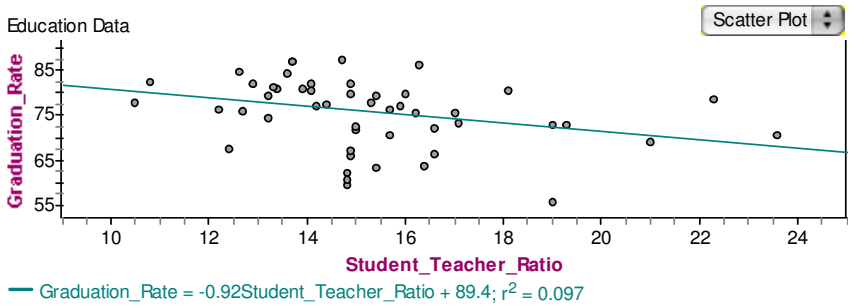
the prices is less for the smaller houses than for the larger houses.

d. As seen in the boxplots shown here, the number of bathrooms is strongly related to the selling price. A lurking variable here is the number of square feet in the house, which is very strongly related to both the price and the number of bathrooms. A regression line is not appropriate here mostly because of the skewness in the prices for houses with three bathrooms and because you can do something better. You can compute the mean (or perhaps the median) price of a house with one bathroom, with two bathrooms, and with three bathrooms: $40,320; $99,290; and $201,400. This process is equivalent to regression but does not require a linear relationship.



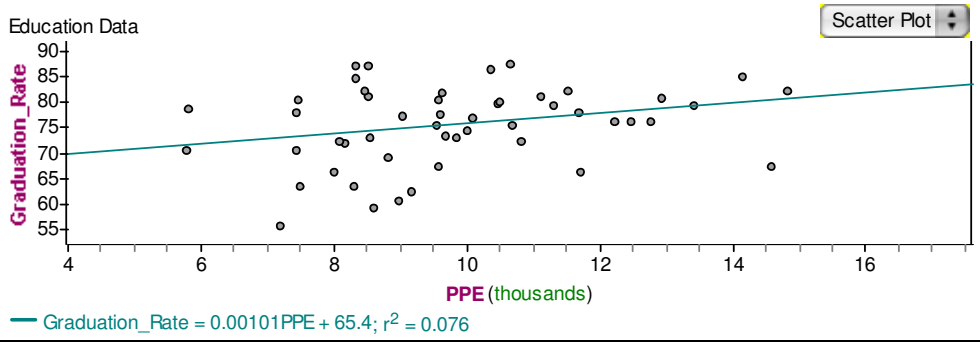
E77. a. The scatterplot shows that *graduation rate* and *student-teacher ratio* are not correlated: $r = 0.311$. The equation of the regression line is

Graduation Rate = -0.92 *Student-Teacher Ratio* + 89.4



b. The scatterplot shows that *graduation rate* and *PPE* are not correlated: $r = 0.275$. The equation of the regression line is

Graduation Rate =0.00101 PPE + 65.4



Concept Review Solutions

C1. A. There aren't any points in the upper left-hand corner because the oldest child has to be older (or the same age, in the case of twins) than the youngest child. Thus, all points must lie on or below the line $y = x$.

C2. C. The predicted birthrate is $(-0.38)(60) + 53.5 = 30.7$, so the residual is the actual birthrate of 47 minus this prediction is 16.3.

C3. C. Curvature in the residual plot of a linear regression is a sign of curvature in the original plot, so statement I is true. When points in the residual plot lie below the line $y = 0$, the points in the original scatterplot lie below the regression line and so the prediction is too large. Thus, statement II is true. Statement III is false because, for example, the pattern could be exponential with a high correlation.

C4. A. Outliers should not be removed permanently from a data set simply because they are outliers. Further investigation is needed, as described in B, C, and D.

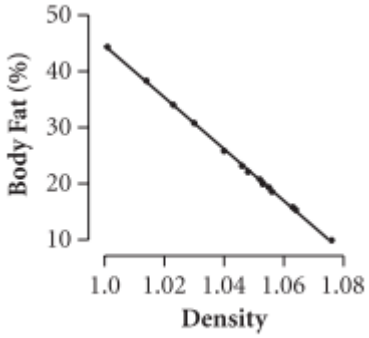
C5. A. B is incorrect because the slope of the regression equation is positive, so the correlation is 0.228. C is incorrect because the value of r^2 does not give any information about linearity versus curvature. E is incorrect because it implies that each person's satisfaction tends to increase over their stay in the hospital. Instead, there may be a lurking variable of age: older people have to stay longer and they also tend to be more satisfied with their care. Or, the lurking variable might be severity of the problem. The more seriously ill a patient is, the longer he or she tends to have to stay, and the more grateful they are for the care they were given.

C6. D

C7. E. A is a poor choice because each point represents a different Barbarian, and so does not establish the trends in a particular Barbarian. B is closer to an interpretation of the intercept than to the slope. C might be close to correct if the y-intercept was near zero, but here it is far from zero. For D, you would have to know the scores on the two sections had equal standard deviations before making this interpretation.

C8. D. Note that the correct statement E is equivalent to saying that 81% of the variation in the number of raids among Barbarians is explained by personal cleanliness.

C9. a. i. From the plot, this line looks to be a good fit:



ii. The regression equation is $\hat{y} = 505.254 - 460.678x$; the analysis is as follows:

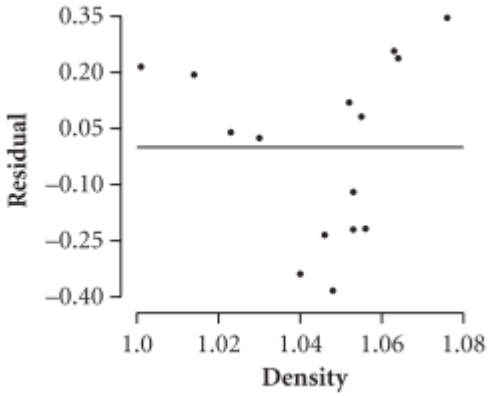
Dependent variable is: % Fat
No Selector
R squared = 99.9% R squared (adjusted) = 99.9%
s = 0.2443 with 15 - 2 = 13 degrees of freedom

Source	Sum of Squares	df	Mean Square	F-ratio
Regression	1206.48	1	1206.48	20223
Residual	0.775560	13	0.059658	

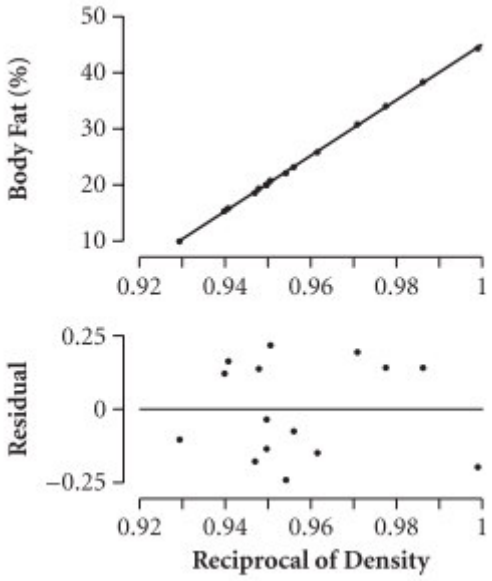
Variable	Coefficient	s.e. of Coeff	t-ratio	prob
Constant	505.254	3.386	149	≤0.0001
Density	-460.678	3.239	-142	≤0.0001

iii. The r^2 value of 0.999 seems to confirm that this model is a good fit.

iv. The residual plot uncovers some of the problems.



b. i. As the percentage of fat increases, the body density decreases. Perhaps the positive association between the *reciprocal of density* and the *percentage of fat* would be easier to model. The pertinent plots and the regression analysis are as follows:



Dependent variable is: % Fat
No Selector

R squared = 100.0% R squared (adjusted) = 100.0%
s = 0.1690 with 15 - 2 = 13 degrees of freedom

Source	Sum of Squares	df	Mean Square	F-ratio
Regression	1206.88	1	1206.88	42246
Residual	0.371389	13	0.028568	

Variable	Coefficient	s.e. of Coeff	t-ratio	prob
Constant	-450.632	2.309	-195	≤0.0001
1/Density	495.654	-0.412	206	≤0.0001

The residuals show less pattern; the plot is more like one of random scatter, suggesting that this is a better model.

The equation of the regression line is

$$\%\ \text{body fat} = -450.63 + 495.65\left(\frac{1}{\text{density}}\right).$$

This is very close to the Siri equation.

ii. The correlation is close to 1 for both models, but the second proves to be a better fitting model than the first. As such, one must never choose a model based on correlation alone.

iii. *Percent body fat* as a function of $\log(density)$ works almost as well as Siri's model. The residual plot, however, has a hint of a pattern.

C10. a. For women, the regression equation was

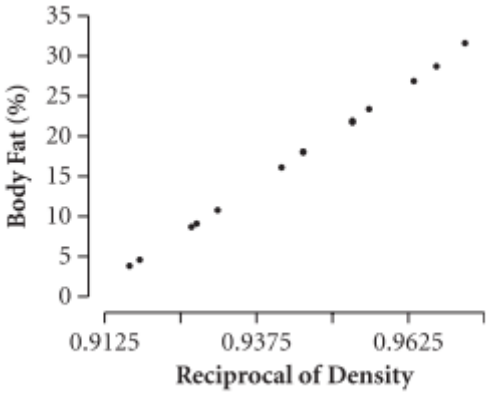
$$\%\ \text{body fat} = -450.63 + 495.65\left(\frac{1}{\text{density}}\right),$$

which is nearly identical to Siri's model (see C9). The relationship is very strong and linear, with correlation nearly equal to 1 and no pattern in the residual plot.

Using the same variables for men, the relationship is again extraordinarily linear (as shown in the next scatterplot) with a correlation near 1. But, this time the regression equation is

$$\%\ \text{body fat} = -453.7 + 498.97\left(\frac{1}{\text{density}}\right),$$

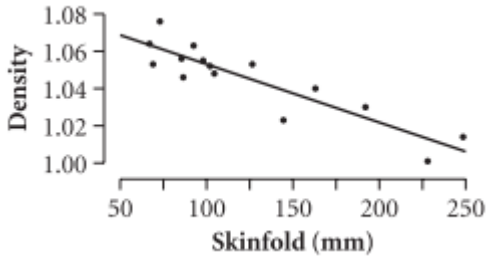
which is similar to Siri's model, but not nearly as close as the equation for women. Thus, the model fits better for women than for men.



b. For women, this scatterplot of *density* against *skinfold* is quite linear and does not require re-expression. Thus, a reasonable model is the regression equation

$$density = 1.084 - 0.000311(skinfold).$$

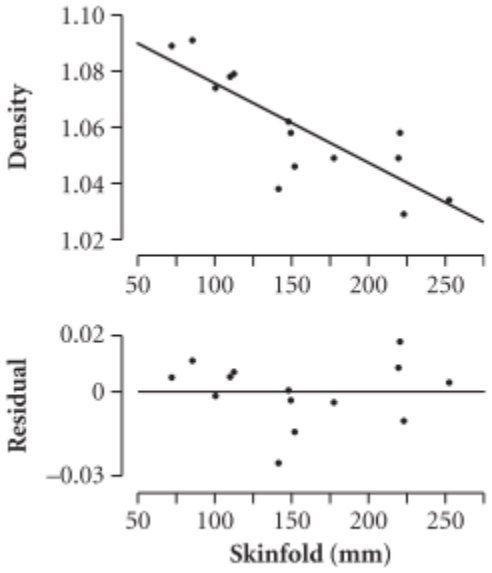
The correlation is -0.897.



For men, the relationship is less strong and has some curvature (see the scatterplot); however, the linear model is an adequate one. The equation is

$$density = 1.105 - 0.000295(skinfold).$$

The residual plot shows some heteroscedasticity.


Density
1.10
1.08
1.06
1.04
1.02
50
100
150
200
250
Residual
0.02
0
–0.03
50
100
150
200
250
Skinfold (mm)