Section 3.1 The Association Between Two Categorical Variables

3.1 Which is response/explanatory?

a) The explanatory variable is carat and the response variable is price.
b) The explanatory variable is dosage and the response variable is severity of adverse event.
c) The explanatory variable is construction type and the response variable is top speed.
d) The explanatory variable is type of college and the response variable is graduation rate.

3.2 Sales and advertising

a) The two variables are amount spent on advertising and monthly sales.
b) Both variables are quantitative.
c) The explanatory variable is amount spent on advertising and the response variable is monthly sales.

3.3 Does higher income make you happy?

a) The response variable is happiness and the explanatory variable is income.
b) Using 2010 data:

	Happiness				
Income	**Not Too Happy**	**Pretty Happy**	**Very Happy**	**Total**	***n***
Above Average	$\frac{21}{360} = 0.06$	$\frac{213}{360} = 0.59$	$\frac{126}{360} = 0.35$	**1.00**	**360**
Average	$\frac{96}{850} = 0.11$	$\frac{506}{850} = 0.60$	$\frac{248}{850} = 0.29$	**1.00**	**850**
Below Average	$\frac{143}{604} = 0.24$	$\frac{347}{604} = 0.57$	$\frac{114}{604} = 0.19$	**1.00**	**604**
Total	**260**	**1066**	**488**		**1814**

The proportion of people who are very happy is larger for those with above-average income (35%) compared to those with below-average income (19%), showing an association between these two variables. Also, the proportion of people who are not too happy is much larger (24%) for people with below-average income compared to people with average (11%) or above-average (6%) income.

c) Overall, the proportion of people who reported being very happy is $\frac{488}{1814} = 0.27$.

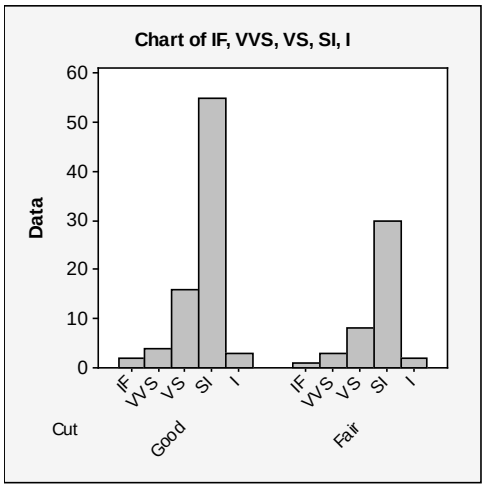
3.4 Diamonds

a)

	Clarity						
Cut	**IF**	**VVS**	**VS**	**SI**	**I**	**Total**	***n***
Good	$\frac{2}{80} = 0.025$	$\frac{4}{80} = 0.050$	$\frac{16}{80} = 0.200$	$\frac{55}{80} = 0.688$	$\frac{3}{80} = 0.038$	**1.00**	**80**
Fair	$\frac{1}{44} = 0.023$	$\frac{3}{44} = 0.068$	$\frac{8}{44} = 0.182$	$\frac{30}{44} = 0.682$	$\frac{2}{44} = 0.045$	**1.00**	**44**

b) The conditional proportions for the two cuts are very similar. For both cuts, the majority of diamonds are rated as slightly included (69% for good cuts, 68% for fair cuts) followed by very slightly included (20% for good cuts, 18% for fair cuts).

3.4 (continued)



c) The conditional proportions are very similar for the two cuts. Based on these data, there appears to be no meaningful association between the cut of a diamond and its clarity rating.

3.5 Alcohol and college students

a) The response variable is binge drinking and the explanatory variable is gender. We wonder if a person's binge drinking status can be explained in part by their gender. (We don't wonder if a person's gender can be explained by their binge-drinking!)

b) (i) There are 1908 male binge drinkers.
(ii) There are 2854 female binge drinkers.

c) The counts in (b) cannot be used to answer the question about differences in proportions of male and female students who binge drink. These are not proportions of male and female students; these are counts. There are far more females than males in this study, so it's not surprising that there are more female than male binge drinkers. This doesn't mean that the percentage of women who binge drink is higher than the percentage of men. If we used these numbers, we might erroneously conclude that women are more likely than are men to be binge drinkers.

d)

	Binge Drinking Status			
Gender	**Binge Drinker**	**Non-Binge Drinker**	**Total**	***n***
Male	$\frac{1908}{3925} = 0.49$	$\frac{2017}{3925} = 0.51$	1.00	3925
Female	$\frac{2854}{6979} = 0.41$	$\frac{4125}{6979} = 0.59$	1.00	6979

These data tell us that 49% of men are binge drinkers, whereas 51% are not. They also tell us that 41% of women are binge drinkers, whereas 59% are not.

e) It appears that men are more likely than are women to be binge drinkers.

3.6 Effectiveness of government at preventing terrorism

a) The response variable is the opinion about the ability of the government to prevent terrorism, categories: terrorists will always find a way, government can eventually prevent all major attacks. The explanatory variable is effectiveness with the categories, very effective, somewhat, not too or not at all effective.

3.6 (continued)

b)

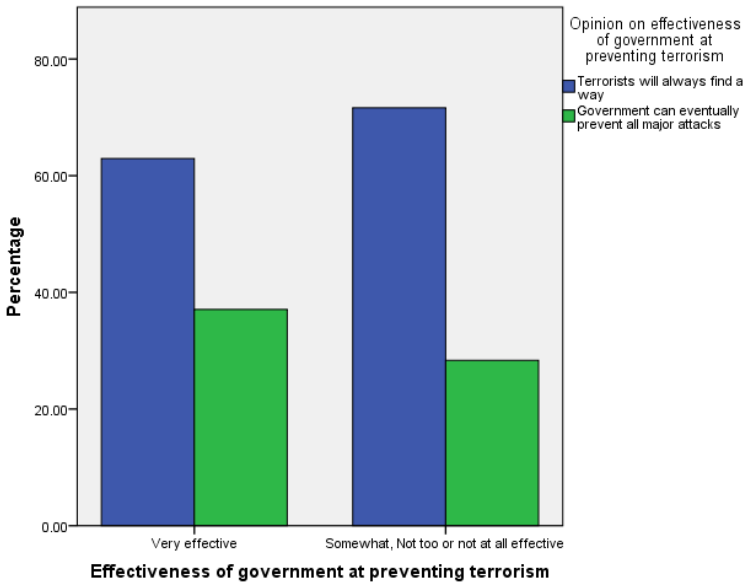
Effectiveness of government at preventing terrorism	Opinion on effectiveness of government at preventing terrorism: Terrorists will always find a way	Opinion on effectiveness of government at preventing terrorism: Government can eventually prevent all major attacks	Total
Very effective	321	189	510
Somewhat, not too or not at all effective	720	285	100

c)

Effectiveness of government at preventing terrorism	Opinion on effectiveness of government at preventing terrorism: Terrorists will always find a way	Opinion on effectiveness of government at preventing terrorism: Government can eventually prevent all major attacks	Total
Very effective	62.94	37.06	100
Somewhat, not too or not at all effective	71.64	28.36	100

d) Yes, the percentages reported in (c) are conditional percentages. We are reporting the percentages restricted to people who belong to a specific category according to their different opinions.

e)



f) We would need to have identical line's values of conditional percentages for each category in the table obtained in part (c). For example, if 37.06% of both individuals who consider government very effective or not very effective stated that it can prevent all major terrorism attack, we could conclude that there is no association.

g) We'd need to have identical line's values of conditional percentages for each category in the table obtained in part c.

3.7 In person or over the phone

a) Since either variable could be considered the outcome of interest, either variable could be taken as the response variable with the remaining being the explanatory variable.

b)

	Gender		
Interview type	**Female**	**Male**	**Total**
In person	1644	551	2195
Over the phone	320	17	337
Total	**1964**	**568**	**2532**

c)

	Gender	
Interview type	**Female**	**Male**
In person	0.8371	0.9701
Over the phone	0.1629	0.0299
Total	**1**	**1**

Overall, it appears that if a person was male, they were very likely interviewed in person (97.01%); however, if a person was female, the probability that they were interviewed in person decreases to (83.71%).

d)

	Gender		
Interview type	**Female**	**Male**	**Total**
In person	0.749	0.251	1
Over the phone	0.9496	0.0504	1

If a person was interviewed in person, they are much more likely to be female (74.9%) than male (25.1%). If a person was interviewed over the phone, they are almost certain to be female (94.96%).

e) (i) 1964/2532 = 77.57% of the respondents are female.

(ii) 2195/2532 = 86.69% of the respondents were interviewed in person.

3.8 Surviving the *Titanic*

a) The percentage of children and female adult passengers who survived is 373/(373 + 161) = 69.9%, the percentage of male adults is 338/(338 + 1329) = 20.3%.

b) The difference between children and female adult passengers and male adult passengers is 69.9 – 20.3 = 49.6. The proportion of children and female adult passengers surviving the sinking of the *Titanic* is 49.6 percentage points higher than the proportion for male adult passengers.

c) The ratio between children and female adult passengers and male adult passengers is 69.9/20.3 = 3.4. The proportion of children and female adult passengers surviving the sinking of the Titanic is 3.4 times larger than the proportion for male adult passengers.

3.9 Gender gap in party ID

a) The response variable is party identification, and the explanatory variable is gender.

b) (i) Male and Republican: 89/892 = 0.10

(ii) Female and Republican: 95/892 = 0.11

c) (i) Male: 355/892 = 0.40 (ii) Republican: 184/892 = 0.21

d) The proportions in (c) are marginal proportions because they are proportions referring to the entire sample.

e) These *conditional* proportions suggest that more females (about 43%) identify as Democrat than males (about 30%), whereas a higher proportion of males identifies as Independent or Republican than females.

3.10 Use the GSS

a)

	Happiness			
Gender	**Very happy**	**Pretty happy**	**Not too happy**	**Row total**
Male	7128	13,088	2840	**23,056**
Female	9402	16,164	3699	**29,265**
Column total	**16,530**	**29,252**	**6539**	**52,321**

b)

	Happiness			
Gender	**Very happy**	**Pretty happy**	**Not too happy**	**Row Total**
Male	$7128/23{,}056 = 0.309$	$13{,}088/23{,}056 = 0.568$	$2840/23{,}056 = 0.123$	1.00
Female	$9402/29{,}265 = 0.321$	$16{,}164/29{,}265 = 0.552$	$3699/29{,}265 = 0.126$	1.00
Column total	$\mathbf{16{,}530/52{,}321 = 0.316}$	$\mathbf{29{,}252/52{,}321 = 0.559}$	$\mathbf{6539/52{,}321 = 0.125}$	**1.00**

c) The conditional proportions for males and females are similar across all three types of happiness. For those reporting being not too happy, the difference is 12.3 – 12.6 = –0.3. The proportion of males being not too happy is 0.3 percentage points smaller than the proportion for females. The ratio is 12.3/12.6 = 0.98. The proportion of males being not too happy is 2 percent smaller than the proportion for females.

Section 3.2 The Association Between Two Quantitative Variables

3.11 Used cars and direction of association

a) We would expect a positive association because as cars age, they tend to have covered more miles. Higher numbers on one variable tend to associate with high numbers on the other variable (and low with low).

b) We would expect a negative association because as cars age, they tend to be worth less. High numbers on one variable tend to associate with low numbers on the other.

c) We would expect a positive association because older cars tend to have needed more repairs.

d) We would expect a negative association. Heavier cars tend to travel fewer miles on a gallon of gas.

e) We would expect a positive association; the heavier the car, the more fuel it will burn to move forward.

3.12 Broadband and GDP

a) For countries with GDP less than $5000 billion, there is a clear trend in that countries with larger GDP have a larger number of broadband subscribers. Three countries stand out in terms of both their GDP and number of subscribers. Although Japan, the third-largest country in terms of GDP, seems to follow the trend, China (the second-largest country in terms of GDP) has by far the most broadband subscribers, whereas the United States (the country with the largest GDP by far) has fewer broadband subscribers than China.

b) The country is China, with approximately $x = 8000$ billion GDP and $y = 160$ million broadband subscribers.

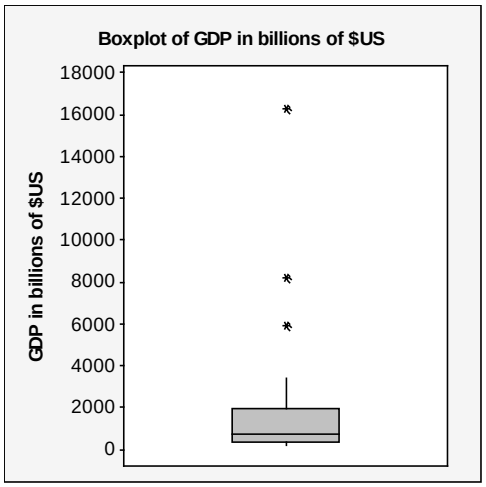
c) $r = 0.77$; the positive sign indicates a positive association between GDP and number of broadband subscribers; as GDP increases, the number of broadband subscribers tends to increase as well.

d) The United States has fewer than expected; China has more than expected.

e) The correlation coefficient would not change because it does not depend on the units of measurement.

3.13 Economic development based on GDP

a)



b) The distribution is skewed to the right with two clear outliers (China and United States).

c) Nations with small GDP can have both large and small population sizes, resulting in both small and large per capita GDP and revealing no overall trend.

d) These two variables are not measuring the same thing. If the GDP were divided by the same value for all nations (such as in standardizing when dividing by the standard deviation of GDP), the correlation between GDP and standardized GDP would be 1. Here, each nation's GDP value is divided by a different value (the nation's population size).

3.14 Email use and number of children

a) The association is weak. We know this because it is close to zero and zero indicates no relation.

b) Internet hours per week has a stronger association with e-mail hours per week than does ideal number of children because the magnitude of the correlation is larger.

3.15 Internet use correlations

a) Internet users and broadband subscribers have the strongest linear relationship.

b) Facebook users and population have the weakest linear relationship.

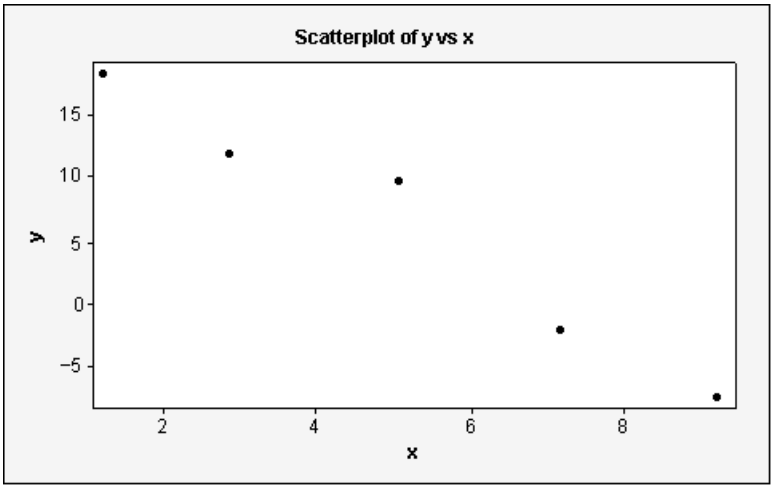
c) The correlation between Internet users and Facebook users does not take population size into account whereas the correlation between Internet use and Facebook use does.

3.16 Match the scatterplot with *r*

1) (c); strong negative association
2) (a); moderate negative association
3) (d); no linear association
4) (b); moderate positive association

3.17 What makes $r = -1$?

a)



b) It is the pair: (5, 10).

c) The value of 10 would have to be changed to 5.

3.18 Gender and Chocolate Preference

The correlation coefficient is only valid to measure the association between two quantitative variables, not between two categorical variables.

3.19 $r = 0$

Consider this data:

x | 0 1 2 4 2

y | 1 0.7 0.5 1 2

The correlation obtained for this data is about 0.10, but when the data point (1, 0.7) is removed, it is just about 0.

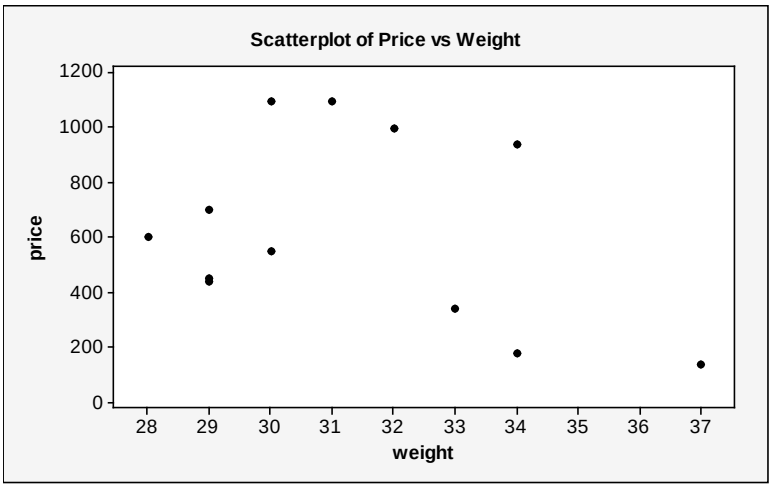
3.20 Correlation inappropriate

It is only appropriate to use this measure of correlation when an association is linear. If an association is curvilinear (e.g., U-shaped), we should not use this statistic. For example, in some situations people perform poorly at very low and very high levels of anxiety. They perform best with a moderate amount of performance-enhancing anxiety. This would form an upside-down U-curve, and it would not be appropriate to measure this association using the correlation you learned in this chapter.

3.21 Which mountain bike to buy?

a) (i) The explanatory variable would be weight.
(ii) The response variable would be price.

b)

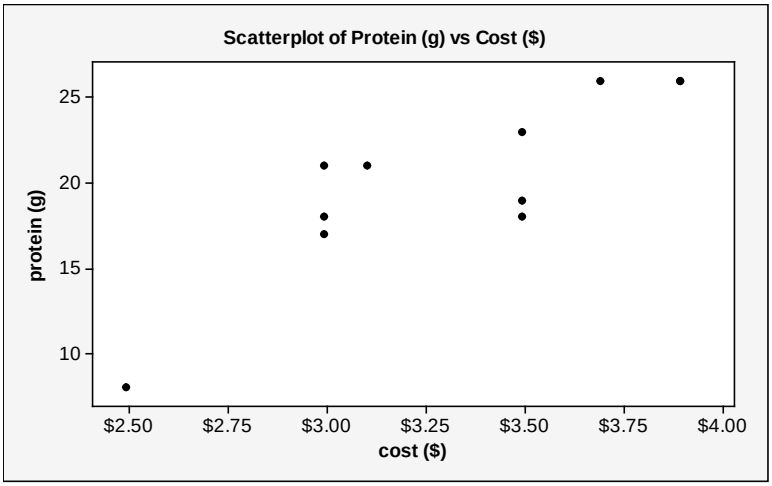


The relation deviates from linearity in that the bikes with weights in the middle tend to cost the most, with those weighing less and more tending to cost less.

c) The correlation is negative and fairly small. This indicates some relation between variables, such that as weight increases, price tends to decrease. Because, however, these variables deviate from linearity in their relation, this correlation coefficient is not an entirely accurate measure of the relation.

3.22 Prices and protein revisited

a)



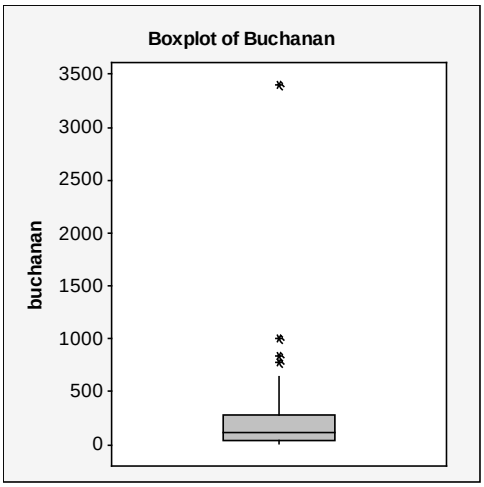
The association seems to be positive; however, the association is probably heavily influenced by the unusual data point representing the sandwich that costs $2.49 and has a protein content of 8 grams.

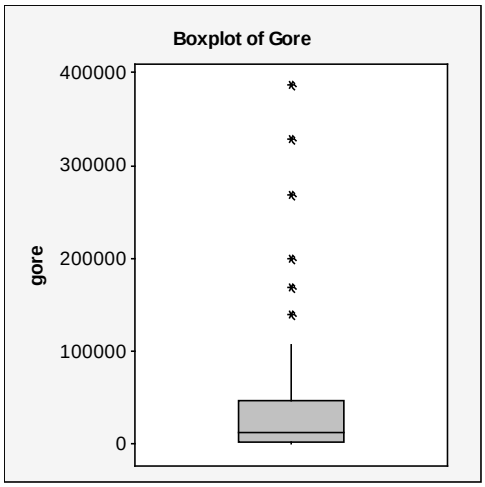
b) The unusual data point is the only vegetarian sandwich on the list which reasonably explains why it has a lower protein content as well as a lower cost than the other sandwiches.

c) The correlation is 0.864. This is a strong positive correlation. It suggests that as the cost of the sandwich increases, the protein content of the sandwich increases as well.

3.23 Buchanan vote

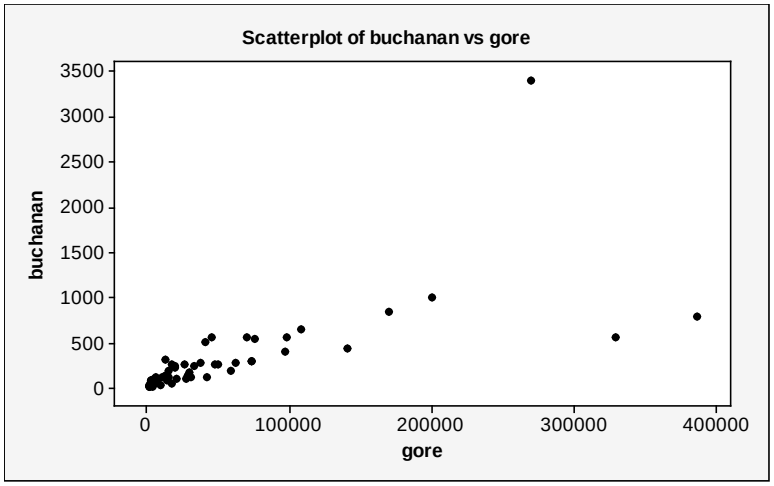
a)





Both box plots indicate that the counts are skewed to the right with few counties in the high ranges of vote counts.

b)

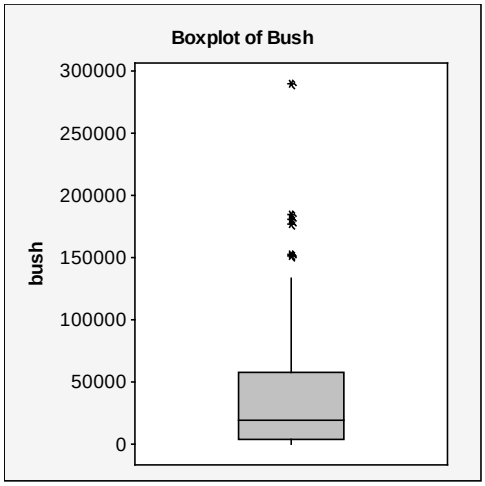


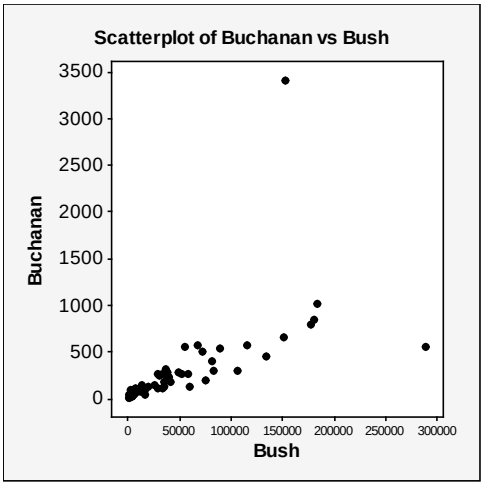
The point close to 3500 on the variable "Buchanan" is a regression outlier; we were unable to make this comparison from the box plots because there were two separate depictions, one for each candidate.

c) We would have expected Buchanan to get around 1000 votes.

d) The box plot for Buchanan would be the same as in (a).

3.23 (continued)





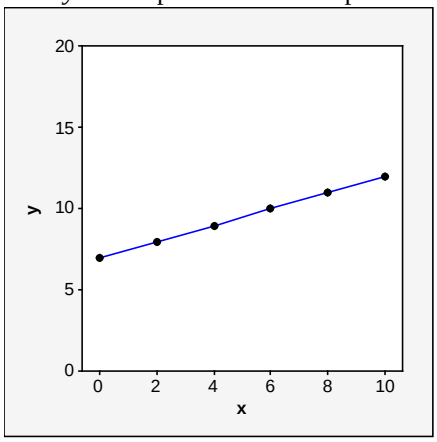
As with the scatterplot with the data for Gore, the point close to 3500 on the variable "Buchanan" is an outlier.

Section 3.3 Predicting the Outcome of a Variable

3.24 Sketch plots of lines

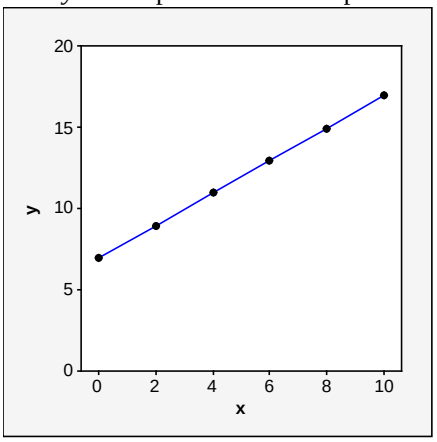
a) $\hat{y} = 7 + 0.5x$

The *y*-intercept is 7 and the slope is 0.5.



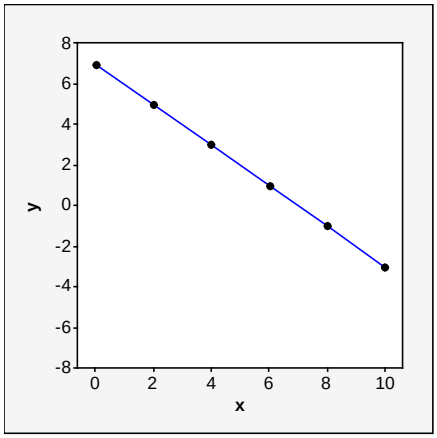
b) $\hat{y} = 7 + x$

The *y*-intercept is 7 and the slope is 1.



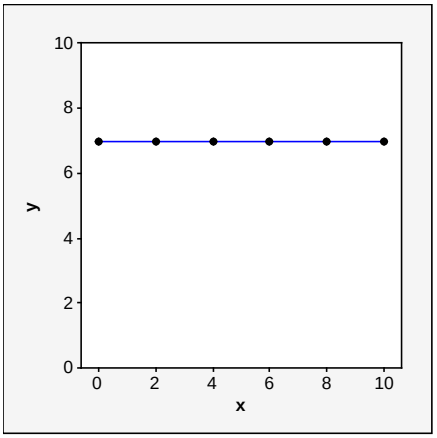
c) $\hat{y} = 7 - x$

The *y*-intercept is 7 and the slope is –1.



d) $\hat{y} = 7$

The *y*-intercept is 7 and the slope is 0.



3.25 Sit-ups and the 40-yard dash

a) (i) $\hat{y} = 6.71 - 0.024x = 6.71 - 0.024(10) = 6.47$

(ii) $\hat{y} = 6.71 - 0.024x = 6.71 - 0.024(40) = 5.75$

The regression line would the line that connects the points (10, 6.47) and (40, 5.75).

b) The *y*-intercept indicates that when a person cannot do any sit-ups, she/he would be predicted to run the 40-yard dash in 6.71 seconds. The slope indicates that every increase of one sit-up leads to a decrease in predicted running time of 0.024 seconds.

c) The slope indicates a negative correlation. The slope and the correlation based on the same data set always have the same sign.

3.26 Wage bill of Premier League Clubs

a) (i) $\hat{y} = -1.537 + 1.056(100) = 104.063$; we would, therefore, predict a Premier League Club whose wage bill of £100 million in 2013 would have a wage bill of £104.063 million in 2014.

(ii) $\hat{y} = -1.537 + 1.056(200) = 209.663$; we would, therefore, predict a premier league club whose wage bill of £200 million in 2013 would have a wage bill of £209.663 million in 2014.

b) For every increase of one unit (a million £ wage bill in 2013), wage bills in 2014 are predicted to increase by £1.056 million. We see this from the value of the slope in the equation, and from the results in part (a) by subtracting the predicted value for two units of area from three units of area and dividing by 100 $\left(\frac{209.663 - 104.063}{100} = 1.056\right)$.

c) The correlation between these variables is positive. There is a positive slope. In addition, by putting in different values for *x*, we can see that as wage bill in 2013 increases, so do wage bill in 2014.

d) The predicted value is £104.063 (see part a), and the actual value is £105 m. The formula for the residual is $-\hat{y}$. In this case, 105 – 104.063 = 0.937. The residual is a measure of error; thus, error for this data point is 0.937; the wage bill in 2014 was £937,000 higher than what would be predicted by this equation.

3.27 Rating restaurants

a) (i) The predicted cost of a dinner in a restaurant that gets the lowest food quality rating of 21 is $\hat{y} = -70 + 4.9(21) = \$32.90$.

(ii) The predicted cost of a dinner in a restaurant that gets the highest food quality rating of 28 is $\hat{y} = -70 + 4.9(28) = \$67.20$.

b) For every 1 point increase in food quality rating, the predicted price of the dinner increases by \$4.90.

c) The correlation between the cost of a dinner and the food quality rating is 0.68, which is a moderate positive correlation. This indicates that higher costs are associated with restaurants receiving higher food quality ratings.

d) The slope can be calculated using the formula: $b = r\left(s_y / s_x\right) = 0.68\left(14.92/2.08\right) = 4.9$.

3.28 Predicting cost of meal from rating

Since the service rating has the highest absolute correlation with the cost of a dinner (0.69), it can be used to make the best predictions of the cost of a dinner.

3.29 Internet and email use

a) Although slope and correlation usually have different values, they always have the same sign. This is due to the relationship $b = r\left(s_y / s_x\right)$ where s_x and s_y are always positive.

b) $\hat{y} = 3.54 + 0.25(60) = 18.54$. Your friend's predicted email use is 18.54 hours per week.

c) The predicted value is 18.54 hours per week (see part b), and the actual value is 10. The formula for the residual is $y - \hat{y}$. In this case, 10 – 18.54 = –8.54. The residual is a measure of error; thus, error for this data point is –8.54; it is 8.54 hours lower than what would be predicted by this equation.

3.30 Government debt and population

a) The slope is –13.495, a negative number; hence, this is a negative association. For every unit increase in population size, the predicted amount of government debt per person decreases by $13.495.

b) (i) $\hat{y} = 19560.405 - 13.495(4) = 19506.425$; at the minimum population size of 4 million, we would predict there to be $19506.425 government debt per person.

(ii) $\hat{y} = 19560.405 - 13.495\ (1367.5) = 1105.993$; at the maximum population size of 1367.5 million, we would predict there to be $1105.993 government debt per person.

c) $\hat{y} = 19560.405 - 13.495\ (319.3) = 15251.452$; the predicted value is $15251.452 and the actual value is $58,604. The formula for the residual is $y - \hat{y}$. In this case, 58,604 – 15,251.452 = $43,352.548. The residual is a measure of error; thus, error for this data point is 43,352.548; it is 43,352.548 higher than what would be predicted by this equation. The U.S. has far more government debt per person than would be predicted from this equation.

3.31 Diamond weight and price

a) $\hat{y} = 109.618 + 0.043(34.65) = 111.108$. Princie's predicted price is $111.108 million.

b) The formula for the residual is $y - \hat{y}$. In this case, 39.3 – 111.108 = –71.808. The residual is a measure of error; thus, error for this data point is –76.458; it is $76.458 million less than what would be predicted by this equation. Thus, Princie's sale price was $76.458 million lower than would be predicted using this equation.

c) $r^2 = 0.003$ indicates that the prediction error using the regression line to predict y is 0.3% smaller than the prediction error using the mean of y to predict y. Therefore, a diamond's weight appears not to be a reliable predictor of its sold price.

3.32 How much do seat belts help?

a) The slope is the amount that y is predicted to change for every unit increase for x. As seat belt usage (x) increases by 1 percentage point, the predicted number of deaths per 100,000 population (y) decreases by 24.45. Thus, the slope is –24.45.

b) (i) $\hat{y} = 32.42 - 24.45(0) = 32.42$. If no one wears seat belts, the predicted number of deaths per 100,000 people in a state in 2013 is 32.42.

(ii) $\hat{y} = 32.42 - 24.45(0.74) = 14.33$. If 73% of people wear seat belts, the predicted number of deaths per 100,000 people in a state in 2013 is 14.33.

(iii) $\hat{y} = 32.42 - 24.45(1) = 7.97$. If everyone wears seat belts, the predicted number of deaths per 100,000 people in a state in 2013 is 7.97.

3.33 Regression between cereal sodium and sugar

a) The software calculates the line for which the sum of squares of the residuals is a minimum.

b) No, any other line would have a larger sum of squares of the residuals.

c) The two rightmost bars represent two cereals whose predicted sodium contents were much less than their actual sodium contents. The cereals are Rice Krispies and Raisin Bran.

d) The amount of sugar is not a reliable predictor for the amount of sodium since r^2 is close to zero (0.2%).

3.34 Expected time for weight loss

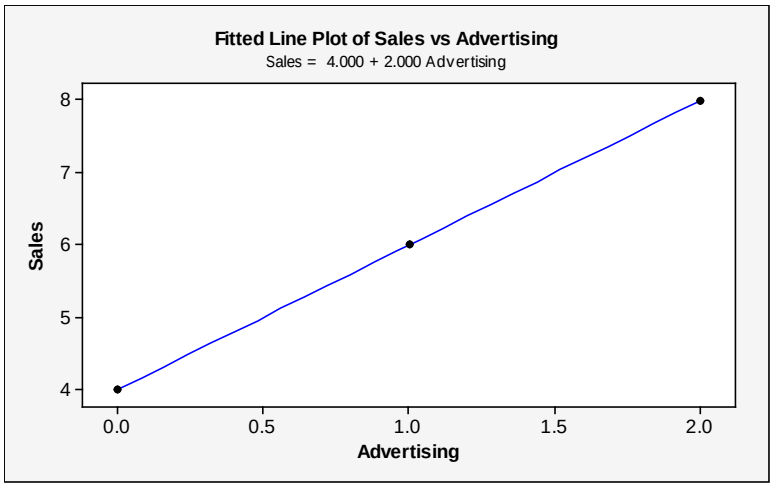
a) r^2 would be relatively large because with good predictors we expect a reasonable gain by using the regression equation with weight to predict number of weeks.

b) Almost 37% ($r^2 = 0.37$) of the variability in number of weeks can be explained by desired weight loss.

c) The slope can be calculated using the formula: $b = r\left(s_y / s_x\right) = 0.607\left(14.393/20.005\right) = 0.437$.

3.35 Advertising and sales

a)



b) The correlation is 1.0. The equation for the regression line is $\hat{y} = 4 + 2x$. See (a).

c) Advertising:

$$\bar{x} = \frac{\sum x}{n} = \frac{0+1+2}{3} = \frac{3}{3} = 1;$$

$$s^2 = \frac{\sum (x - \bar{x})^2}{n-1} = \frac{(0-1)^2 + (1-1)^2 + (2-1)^2}{3-1} = \frac{2}{2} = 1$$

$$s = \sqrt{s^2} = \sqrt{1} = 1$$

Sales:

$$\bar{y} = \frac{\sum y}{n} = \frac{4+6+8}{3} = \frac{18}{3} = 6;$$

$$s^2 = \frac{\sum (y - \bar{y})^2}{n-1} = \frac{(4-6)^2 + (6-6)^2 + (8-6)^2}{3-1} = \frac{8}{2} = 4$$

$$s = \sqrt{s^2} = \sqrt{4} = 2$$

d) $b = r\left(s_y / s_x\right) = 1(2/1) = 2$; $a = \bar{y} - b\bar{x} = 6 - 2(1) = 4$; $\hat{y} = 4 + 2x$

The y-intercept of 4 indicates that when there is no advertising, it is predicted that sales will be about \$4000. The slope of 2 indicates that for each increase of \$1000 in advertising, predicted sales increase by \$2000.

3.36 Midterm–final correlation

a) (i) $\hat{y} = 30+0.6x = 30+0.6(100) = 90$

(ii) $\hat{y} = 30+0.6x = 30+0.6(50) = 60$

In both cases, the prediction for final exam score is closer to the mean of 75 than to the original midterm score.

b) $b = r\left(s_y / s_x\right) \Rightarrow 0.6 = r(10/10) \Rightarrow 0.6 = r(1) \Rightarrow r = 0.6$

(Note that when the spread for two variables is the same, the correlation equals the slope. In this instance, both variables (midterm and final) have a standard deviation of 10; thus, the correlation would be the same as the slope, 0.6.) This indicates that these two variables have a positive correlation. Those who scored higher on the midterm also tended to score higher on the final.

3.37 Predict final exam from midterm

a) $b = r\left(s_y / s_x\right) = 0.70(10/10) = 0.70$; $a = \bar{y} - b\bar{x} = 80 - (0.70)(80) = 24$

The regression equation is $\hat{y} = 24 + 0.70x$.

b) The predicted final exam score for a student with an 80 on the midterm is 24 + 0.70(80) = 80.
The predicted final exam score for a student with an 90 on the midterm is 24 + 0.70(90) = 87.

3.38 NL baseball

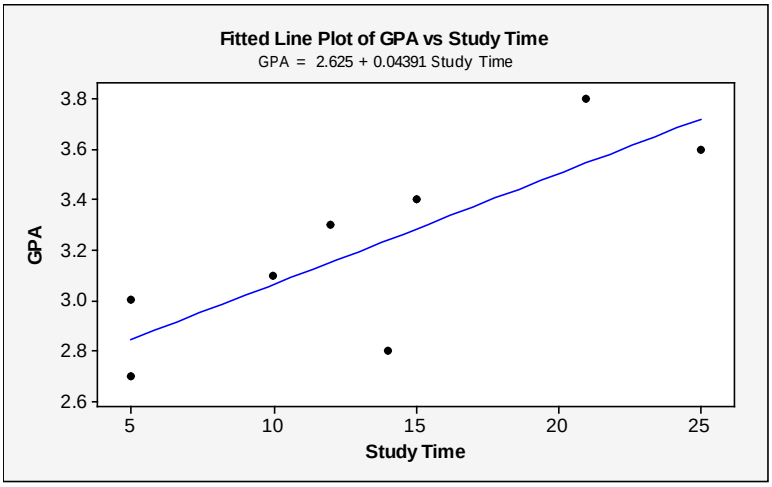
a) The slope indicates that for an increase in team batting average of 0.010, the predicted team scoring increases by 0.415. (Note that there is never a difference of 1 between any two teams given the range of the team batting averages, so it is not relevant to consider an increase of 1. For an increase of 0.01 in team batting average, the predicted increase is 0.42 runs per game.)

b) $b = r\left(s_y / s_x\right)$; b = 0.900(0.3604/0.00782) = 41.5

c) An r^2 of 0.81 indicates 81% of the variability in runs per game is accounted for by variability in team batting averages.

3.39 Study time and college GPA

a)



The linear correlation between GPA and study time appears to be positive and fairly strong since the data points follow a positive linear trend.

b) The correlation is 0.81. This indicates that the association between GPA and study time is strong and positive; longer study times are associated with higher GPAs.

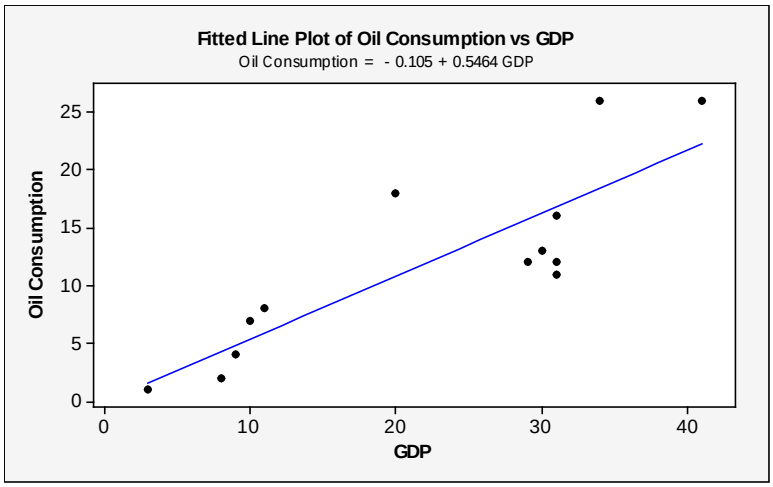
c) See (a); The prediction equation is $\hat{y} = 2.625 + 0.0439x$.

(i) A student who studies 5 hours per week is predicted to have a GPA of $2.625 + 0.0439(5) = 2.84$.

(ii) A student who studies 25 hours per week is predicted to have a GPA of $2.625 + 0.0439(5) = 3.72$.

3.40 Oil and GDP

a)



The linear correlation between oil consumption and GDP appears to be positive and fairly strong since the data points follow a positive linear trend.

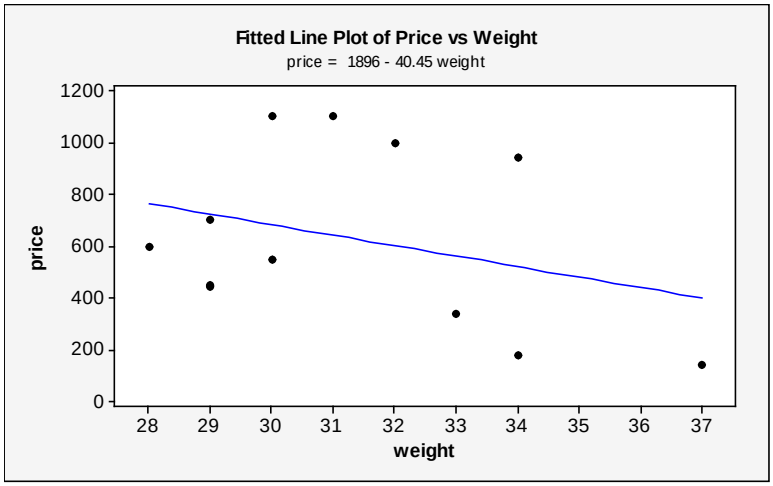
b) See (a); The prediction equation is given by $\hat{y} = -0.105 + 0.5464x$.

c) The correlation is 0.85. This indicates that the association between GDP and oil consumption is strong and positive, higher gross domestic products are associated with higher annual oil consumptions per person.

d) Canada has a GDP value of 34; thus, the predicted annual oil consumption per person for Canada is –0.105 + 0.5464(34) = 18.5. This gives a residual value of 26 – 18.5 = 7.5.

3.41 Mountain bikes revisited

a)



b) See (a); The regression equation is price = 1896 – 40.5weight.
For every 1 unit increase in weight, the predicted price decreases by $40.50. Because it's impossible for a bike to have 0 weight, the *y*-intercept has no contextual meaning here.

c) $\hat{y} = 1896 - 40.5x = 1896 - 40.5(30) = 681$; the predicted price is $681.

3.42 Mountain bike and suspension type

a) The relationship between weight and price seems to be linear among bikes with front end suspension, and it also seems to be linear among bikes with full suspension, but when all bikes are included, the relationship is not completely linear. Thus, the simple regression line is not the best way to fit the data. It's better to calculate separate regression lines for each type of suspension.

3.42 (continued)

b) Front end: The regression equation is price_FE = 2136 – 55.0weight_FE.
Full: The regression equation is price_FU = 2432 – 44.0weight_FU.
Compared to the slope calculated in Exercise 3.41 (which was –40.5), these slopes are larger in magnitude, an indication of a stronger relationship between variables when types of bike are looked at separately.

c) If the correlations for full and front end suspension bikes are found separately, I would expect that the correlations would be higher for each type of bike. For front end suspension bikes, the correlation is –0.888, whereas for full suspension bikes, the correlation is –0.952. Thus, the correlations are still negative (i.e., heavier bikes cost less), but the magnitude of the relationships are far stronger when correlations are examined only among bikes of a certain suspension type.

d) You may justify numerous ways such as: plotting the point on the scatterplot and seeing which cluster of points it most likely belongs to; plugging into least squares regression equations, then taking predicted values and computing the residuals to see which suspension type has the smallest residual (*see below for the solution using this method*); recalculating correlation coefficients for each suspension type with the new point to see if the original correlation coefficients change.
For front end: $\hat{y} = 2136 - 55.0x = 2136 - 55.0(28.5) = \$568.50$; Residual: $y - \hat{y} = 700 - 568.5 = 131.5$
For full: $\hat{y} = 2432 - 44.0x = 2432 - 44.0(28.5) = \$1178.00$; Residual: $y - \hat{y} = 700 - 1178 = -478$
The error is smaller when the prediction is made with the formula for front end suspension bikes than with the formula for full suspension bikes; I would predict that the bike has a front end suspension.

3.43 Fuel Consumption

a) The scatterplot reveals a nonlinear (curved) pattern. The correlation coefficient measuring the strength of a linear relationship is meaningless for nonlinear relationships.

b) Due to the nonlinear relationship, the regression equation is not appropriate to model the relationship between driving speed and mpg and cannot be used for making predictions.

c) For the range from about 40 mph to 85 mph (or from 5 mph to 40 mph). Over each of these ranges, the relationship is approximately linear.

Section 3.4 Cautions in Analyzing Associations

3.44 Extrapolating murder

a) The *x*-value was approximately 14 for Utah and 30 for Mississippi.

b) $\hat{y} = -8.25 + 0.56x = -8.25 + 0.56(0) = -8.25$; This prediction makes no sense because we cannot have a murder rate that is less than 0! This occurs because we are extrapolating beyond our data, a statistically dangerous practice.

3.45 Men's Olympic long jumps

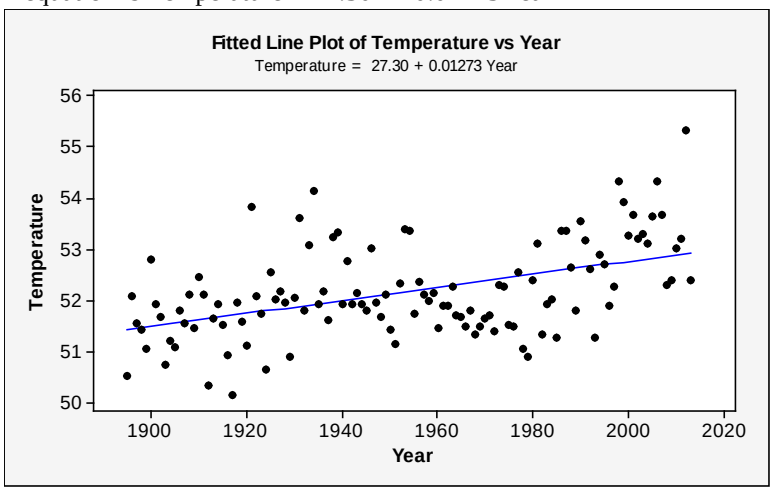
a) The observation in the lower left of the scatterplot (1896) may influence the fit of the regression line. This observation was identified because it well below the general trend of the data.

b) The prediction from the regression line would be more reasonable than would the prediction based on the mean because of the strong positive linear trend exhibited by the data.

c) No. Extrapolating predictions well beyond the range of the observed *x* values is unreliable. No one knows whether the linear trend continues so many years out.

3.46 U.S. average annual temperatures

a) The regression equation is Temperature = 27.304 + 0.01273Year



The slope of 0.01273 indicates a predicted increase of 0.01273 degree for each increase of one year.

b) (i) $\hat{y} = 27.304 + 0.01273(2016) = 52.97$

(ii) $\hat{y} = 27.304 + 0.01273(2500) = 59.13$

c) I have more faith in the prediction made for 2016. It is dangerous to extrapolate to a year that is so far off – like 2500. The temperature trends might have changed drastically by that point (although they might have even changed by 2016).

3.47 Murder and education

a) $x = 15\%$; $\hat{y} = -3.1 + 0.33(15) = 1.85$

$x = 40\%$; $\hat{y} = -3.1 + 0.33(40) = 10.1$

b) $x = 15\%$; $\hat{y} = 8.0 - 0.14(15) = 5.9$

$x = 40\%$; $\hat{y} = 8.0 - 0.14(40) = 2.4$

c) D.C. is a regression outlier because it is well removed from the trend that the rest of the data follow.

d) Because D.C. is so high on both variables, it pulls the line upwards on the right and suggests a positive correlation, when the rest of the data (without D.C.) are negatively correlated. The relationship is best summarized after removing D.C.

3.48 Murder and poverty

a) Yes; D.C. has a large influence on this regression analysis. When it is included, the intercept decreases by 4.5, and the slope, although still positive, becomes over twice as large.

b) Based on this information, the poverty values of D.C. would be relatively large. A high poverty value, coupled with the high D.C. murder rate, would lead this to be a regression outlier that would pull the right hand side of the regression line upwards.

3.49 TV watching and the birth rate

a) The U.S. is an outlier on (i) *x*, (ii) *y*, and (iii) relative to the regression line for the other six observations.

b) The two conditions under which a single point can have such a dramatic effect on the slope: (1) the *x* value is relatively low or high compared to the rest of the data; (2) the observation is a regression outlier, falling quite far from the trend that the rest of the data follow. In this case, the observation for the U.S. is very high on *x* compared to the rest of the data. In addition, the observation for the U.S. is a regression outlier, falling far from the trend of the rest of the data. Specifically, TV watching in the U.S. is very high despite the very low birth rate.

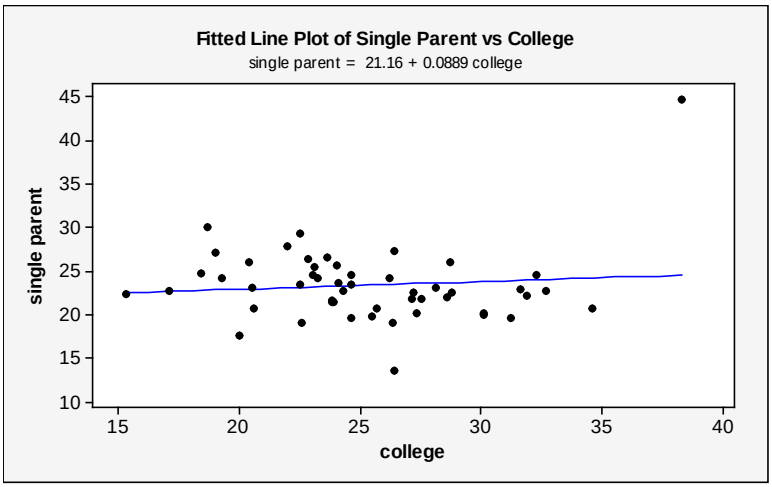
3.49 (continued)

c) The association between birth rate and number of televisions is (i) very weak without the U.S. point because the six countries, although they vary in birth rates, all have very few televisions and these amounts don't seem to relate to birth rate. The association is (ii) very strong with the U.S. point because the U.S. is so much higher in number of televisions and so much lower on birth rate that it makes the two variables seem related. A very high number of televisions does coincide with a very low birth rate in the U.S., whereas all the Asian countries are relatively high in birth rates and low in numbers of televisions.

d) The U.S. residual for the line fitted using that point is very small because that point has a large effect on pulling the line downward. There are no other data points near that line, and all other data points are in the far corner, so the line runs almost directly through the U.S. point.

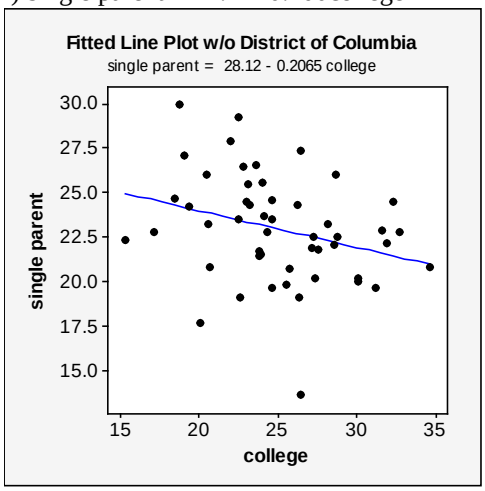
3.50 Looking for outliers

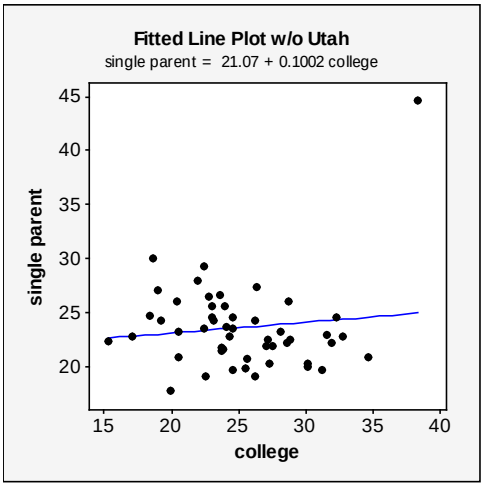
a)



The point at x (college) equals approximately 38 (District of Columbia) is quite a bit different from other observations and at y (single parent) equals approximately 13 (Utah) is a little different from other observations.

b) (i) single parent = 21.2 + 0.089college; see (a).
(ii) single parent = 28.1 – 0.206college
(iii) single parent = 21.1 + 0.100college





c) The observation for the District of Columbia, with x is about 38, has a pretty strong influence on the slope. When it is deleted, the previously positive slope is now negative.

3.50 (continued)

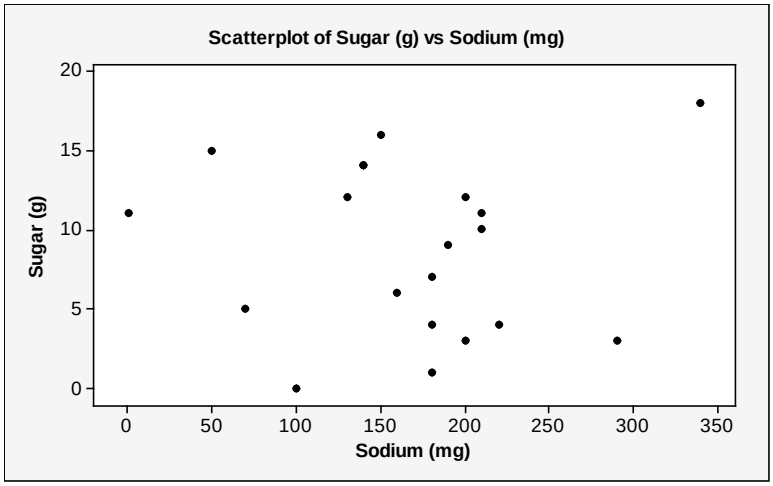
d) $\hat{y} = 21.2 + 0.089(38.3) = 24.61$ (rounds to 24.6)

$\hat{y} = 28.1 - 0.206(38.3) = 20.21$ (rounds to 20.2)

Yes, the predicted value for D.C. is about four points different depending on which equation is used.

3.51 Regression between cereal sodium and sugar

a)



The point (18, 340), which represents Raisin Bran, meets the two criteria in that it has an *x* value far from all the others and falls far from the trend that the rest of the data follow.

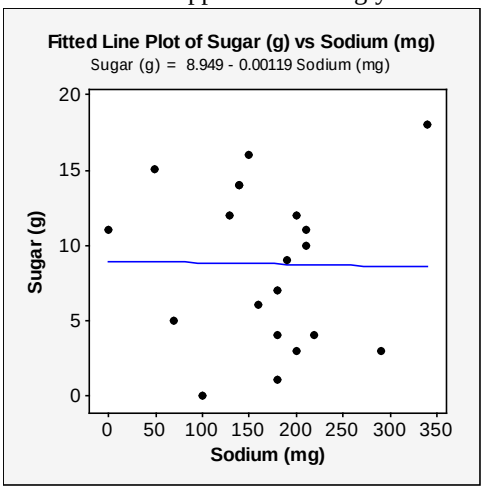
b) All data points

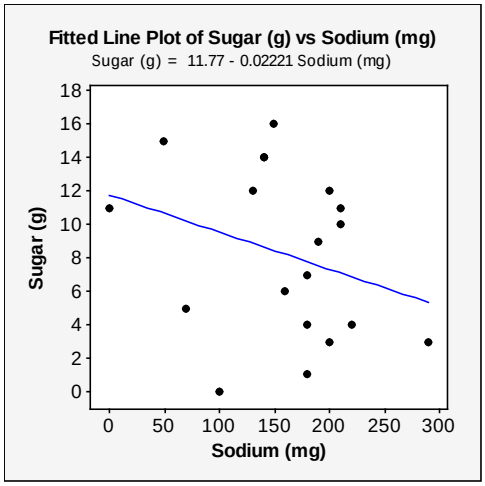
Regression line: SUGAR(g) = 8.949 – 0.00119SODIUM(mg); Correlation: –0.017

All data points except Raisin Bran

Regression line: SUGAR(g) = 11.77 – 0.02221SODIUM(mg)

Correlation: –0.30; Raisin Bran lowers the intercept and makes the slope less steep; overall, the two variables appear less strongly associated when Raisin Bran is included.





3.52 Gestational period and life expectancy

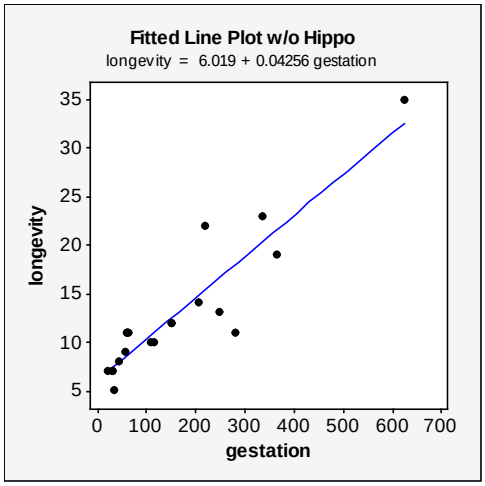
a) The scatterplot shows a strong positive correlation between the length of the gestational period and longevity. The hippo has a higher longevity than what would be expected based on the data from other animals. Although the elephant has by far the longest gestational period, its longevity is also the longest and follows the general trend.

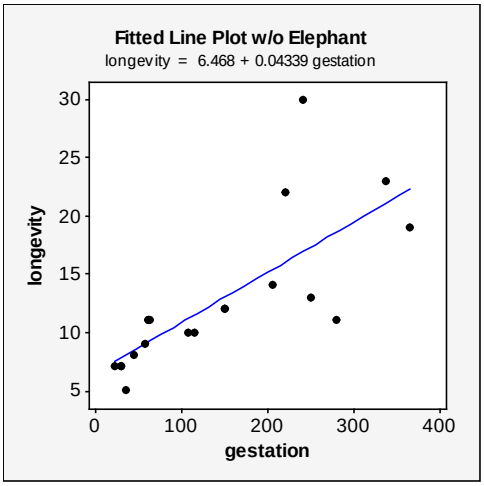
b) Statistical software will verify the results.

3.52 (continued)

c) The elephant and the hippo are outliers. The elephant has a gestational period that is far longer than that of the other animals but is not a regression outlier (because it follows the general trend). The hippo is unusual in its combination of average gestational length with an above-average longevity and is a regression outlier because it does not fit with the general trend. Neither seems too influential in that removing either will not change the slope much. See (d).

d) Removing the elephant: $\hat{y} = 6.47 + 0.043x$ and $r = 0.75$. The slope is almost identical, but the correlation is much weaker, dropping from 0.86 to 0.75. Removing the hippo: $\hat{y} = 6.02 + 0.043x$ and $r = 0.92$. The slope is almost identical, but the correlation is much stronger, increasing from 0.86 to 0.92.



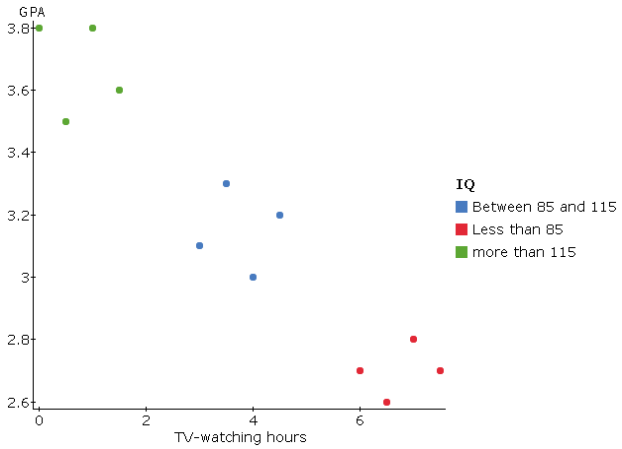


3.53 GPA and hours spent watching TV

a) There is not likely a causal relationship between GPA and the number of hours spent watching TV. Rather, it is more likely that GPA increases and number of hours spent watching TV decrease with IQ score.

b) More academically able students would certainly tend to have a higher GPA, and it may be that they are also less likely to watch TV (because they would rather engage in more intellectual pastimes). At each IQ range, there should be no overall trend (i.e., some lower IQs would be high on both, some would be low on both, and some would be low on one and high on the other).
IQ score plays a role in the association because it could predict both GPA and TV-watching hours. GPA and TV-watching are related because they have a common cause.

c) In the below scatterplot, we see an overall negative correlation between TV-watching hours and GPA (assessed on a scale of 0–4) if we ignore IQ. However, if we look within each IQ range, we see roughly a horizontal trend and no particular association. It is IQ that predicts both TV-watching and GPA.



3.54 Hospital size and length of stay

a) No. Having more beds in a hospital is likely an indicator that the hospital is more likely to have the facilities to treat some serious problems. It is the seriousness of the medical case that leads both to the increased number of beds in a hospital and to the increased stay length. The two variables in this correlation (number of beds in a hospital and stay length) have the common cause of severity of the medical case.

b) As mentioned in part (a), the severity of the medical case is a possible third variable that could consider a common cause of *x* and *y*. Each student's hypothetical scatterplot will be different, but points that are high in number of beds in a hospital and stay length will be high in medical case severity, and points that are low in number of beds in hospital and stay length will be low in medical case severity.

3.55 Does ice cream prevent flu?

a) Although there are several possible responses to this exercise, one possible lurking variable could be temperature/weather. People tend to eat ice cream in the warmer months that is when people are less vulnerable to getting the flu. The flu season is generally winters when it is cold and most people do not prefer to eat ice cream.

b) Catching flu might be caused by temperature/weather, but almost might be caused by other variables, some of which could be associated with temperature/weather. Including vulnerable immunity conditions, other such causal variables could include influenza viruses and hygiene negligence.

3.56 What's wrong with regression?

a) It's dangerous to extrapolate far beyond one's data.

b) Correlation is not causation. For example, there could be a third variable (e.g., income) that is related to both of these variables.

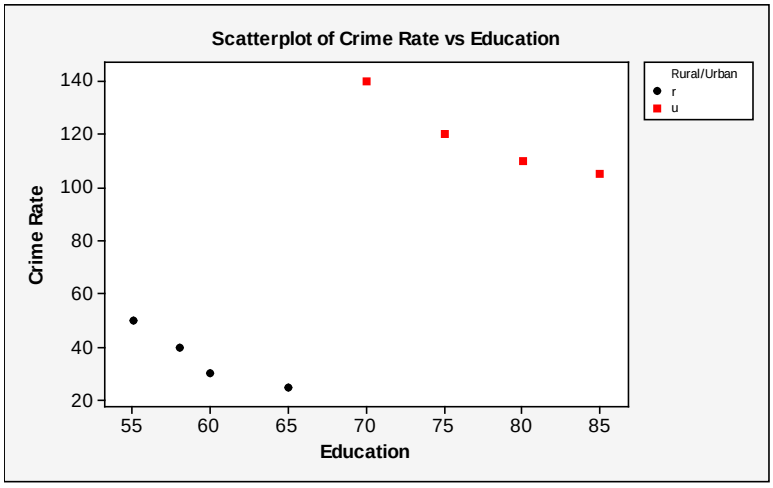
c) This point would be a regression outlier, and could greatly affect the regression equation. We should report the results without this regression outlier.

3.57 Education causes crime?

a) In Minitab, your columns should look similar to the following:

Education	Crime Rate	Rural/Urban
70	140	u
75	120	u
80	110	u
85	105	u
55	50	r
58	40	r
60	30	r
65	25	r

b)



c) The correlation for all 8 data points is 0.73. This indicates a strong, positive linear correlation.

3.57 (continued)

d) (i) The correlation for the urban counties is –0.96, which is a very strong, negative linear correlation.

(ii) The correlation for the rural counties is –0.95, which is also a very strong, negative linear correlation. Note that for each subset of data, a higher education rate is associated with a lower crime rate; however, because both the education and crime rates are so much higher for urban counties than for rural, the correlation appears positive when all of the data is considered together. This is a good example of why it is always important to look at a graphical display of your data to determine if a measure of linear correlation is appropriate.

3.58 Death penalty and race

a)

	Death penalty	**No death penalty**
White defendant	$\frac{53}{467} = 0.11$	$\frac{414}{467} = 0.89$
Black defendant	$\frac{11}{48} = 0.23$	$\frac{37}{48} = 0.77$

Black defendants were more likely than were white defendants to get the death penalty when the victim was white.

b)

	Death penalty	**No death penalty**
White defendant	$\frac{0}{16} = 0.00$	$\frac{16}{16} = 1.00$
Black defendant	$\frac{4}{143} = 0.03$	$\frac{139}{143} = 0.97$

Black defendants were more likely than were white defendants to get the death penalty when the victim was black.

c)

	Death Penalty		
Defendant's Race	**Yes**	**No**	**Total**
White	53	430	483
Black	15	176	191

	Death penalty	**No death penalty**
White defendant	$\frac{53}{483} = 0.11$	$\frac{430}{483} = 0.89$
Black defendant	$\frac{15}{191} = 0.08$	$\frac{176}{191} = 0.92$

These data indicate that white defendants were more likely than were black defendants to get the death penalty.

d) These data satisfy Simpson's paradox which occurs when the association between two variables changes after a third variable is included and the data are analyzed at separate levels of that variable. In this case, the race of the victim played a role. There were so many more white victims, and most people were killed by a member of their own race. Thus, there were more white killers to be put to death. Yet, the few blacks who killed white people were more likely to be put to death than were the many white people who killed white people, and the few white people who killed black people were never put to death.

e) We would call victim's race a confounding variable. Victim's race and defendant's race both predict death penalty status. Confounding occurs when two explanatory variables are associated with a response variable, but also with each other. In such cases, it is difficult to determine whether either of them truly causes the response, because the variable's effect could be at least partly due to its association with the other variable.

3.59 NAEP scores

a) The response variable is eighth grade math scores, and the explanatory variable is state.

b) The third variable is number of pages read in school and for homework. Connecticut has the overall higher mean because the number of pages read is quite different from that in Maryland. There is a higher percentage of people who read more and a lower percentage of people who read less in Connecticut than in Maryland, and overall, people who read more tended to have higher math scores than people who read less.

3.60 Diabetes and breast cancer

a) The researchers wondered whether race/ethnicity was responsible for the association between diabetes and breast cancer. It is possible that race/ethnicity, rather than diabetes, was responsible for the diagnosis of breast cancer. It is possible that both diabetes and breast cancer are correlated to race/ethnicity.

b) If race/ethnicity was not actually measured, it would be a possible lurking variable. A lurking variable is one that might be present, but is not measured in the study. We do not know if it is a confounding variable, but it might be. If it were included in the study, and were found to be linked with both the explanatory and response variables, then it would become a confounding variable.

Chapter Problems: Practicing the Basics

3.61 Choose explanatory and response variables

a) The response variable is the weight of an infant at birth, and the explanatory variable is number of weeks of gestation.

b) The response variable is the preferred smartphone operating system, and the explanatory variable is gender.

c) The response variable is average number of airline trips taken, and the explanatory variable is annual income.

d) The response variable is weekly grocery budget, and the explanatory variable is marital status.

3.62 Graphing data

a) (a) Both variables are quantitative.
(b) Both variables are categorical.
(c) Both variables are quantitative.
(d) Pounds lost is quantitative, and diet is categorical.

b) (a) These data could be graphed with a scatterplot (or histogram for individual variables).
(b) These data could be graphed with a bar graph with side-by-side bars for the two genders for each political party (or we could use two separate pie charts, one for each gender).
(c) These data could be graphed with a scatterplot (or histogram for individual variables).
(d) These data could be graphed with side-by-side box plots or histograms (one for each diet).

3.63 Life after death for males and females

a)

	Opinion about life after death			
Gender	**Yes**	**No**	**Total**	***n***
Male	$\frac{621}{808} = 0.769$	$\frac{187}{808} = 0.231$	1.00	808
Female	$\frac{834}{979} = 0.852$	$\frac{145}{979} = 0.148$	1.00	979

b) 76.9% of males believe in life after death, as opposed to 85.2% of females. The difference in the proportions between females and males is 0.852 – 0.769 = 0.083. The proportion of females believing in life after death is about 8 percentage points higher than the one for males. The ratio of proportions is 0.852/0.769 = 1.108. The proportion of females believing in life after death is about 11% higher (or 1.1 times higher) than the one for males.

3.64 God and happiness

a)

		Happy			
		1: VERY HAPPY	2: PRETTY HAPPY	3: NOT TOO HAPPY	**Row Total**
GOD	1: DONT BELIEVE	13	60	14	**87**
	2: NO WAY TO FIND OUT	36	88	19	**143**
	3: SOME HIGHER POWER	77	190	49	**316**
	4: BELIEVE SOMETIMES	28	49	13	**90**
	5: BELIEVE BUT DOUBTS	108	252	55	**415**
	6: KNOW GOD EXISTS	520	752	190	**1462**
	Column Total	**782**	**1391**	**340**	**2513**

b)

		Happy			
		1: VERY HAPPY	2: PRETTY HAPPY	3: NOT TOO HAPPY	**Row Total**
GOD	1: DONT BELIEVE	$\frac{13}{87} = 0.149$	$\frac{60}{87} = 0.690$	$\frac{14}{87} = 0.161$	**1.00**
	2: NO WAY TO FIND OUT	$\frac{36}{143} = 0.252$	$\frac{88}{143} = 0.615$	$\frac{19}{143} = 0.133$	**1.00**
	3: SOME HIGHER POWER	$\frac{77}{316} = 0.244$	$\frac{190}{316} = 0.601$	$\frac{49}{316} = 0.155$	**1.00**
	4: BELIEVE SOMETIMES	$\frac{28}{90} = 0.311$	$\frac{49}{90} = 0.544$	$\frac{13}{90} = 0.144$	**1.00**
	5: BELIEVE BUT DOUBTS	$\frac{108}{415} = 0.260$	$\frac{252}{415} = 0.607$	$\frac{55}{415} = 0.133$	**1.00**
	6: KNOW GOD EXISTS	$\frac{520}{1462} = 0.356$	$\frac{752}{1462} = 0.514$	$\frac{190}{1462} = 0.130$	**1.00**
	Column Total	**0.311**	**0.554**	**0.135**	**1.00**

Those who respond that they "know God exists" are most likely to report being "very happy."

c) It is more informative to view the proportions in (b) because they allow us to see the proportion of people with a certain level of reported happiness given that they have a certain belief about God. The table in (a) doesn't allow us to make these kinds of comparisons.

3.65 Jobs and income

a) The response variable is annual income. It is quantitative.

b) The explanatory variable is job. It is categorical.

c) A bar graph could have a separate bar for each job type. The height of each bar would correspond to annual salary level for a given category.

3.66 Bacteria in ground turkey

a) The difference in proportion of positive tests between ground turkey with no claim of antibiotic use and a claim of antibiotic use was 26/46 – 23/28 = 0.565 – 0.821 = –0.26; the percent of packages that tested positive for Enterococcus is 26 percentage points lower for conventional packages than for those claiming no use of antibiotics.

b) The ratio of the proportion of positive tests between ground turkey with no claim of antibiotic use and a claim of antibiotic use was 0.565/0.821 = 0.69; the proportion of packages testing positive for Enterococcus is 31% lower for conventional packages compared to those claiming no use of antibiotics.

3.67 Women managers in the work force

a) The response variable is gender, and the explanatory variable is type of occupation.

b)

Percent of Total in Executive, Administrative, and Managerial Positions

Year	Female	Male	Total
1972	0.197	0.803	1.00
2002	0.459	0.541	1.00

c) Based on (b), it does seem that there is an association between these variables. Women made up a larger proportion of the executive work force in 2002 than in 1972.

d) The two explanatory variables shown are year and type of occupation.

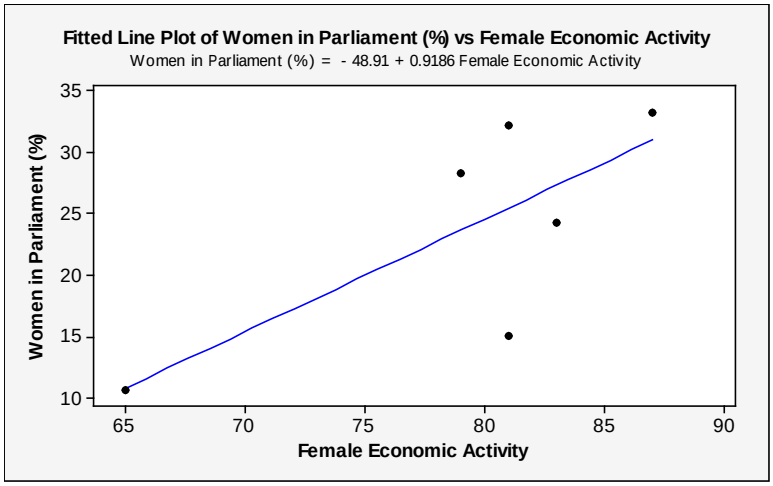
3.68 RateMyProfessor.com

a) The easier a professor grades, the more likely he or she is to receive a higher quality rating.

b) We would expect the correlation to be closer to 0 if there was no association between quality rating and easiness of grading.

3.69 Women in government and economic life

a)



The correlation between women in parliament and female economic activity is 0.745. This correlation is supported by the positive linear trend evident in the scatterplot, but note this is largely driven by the point (for Japan) having female economic activity very low (65).

b) See (a): The regression equation is given by $\hat{y} = -48.91 + 0.9186x$. Since the y-intercept would correspond to an x-value of 0, the y-intercept is not meaningful in this case (Female economic activity = 0 is outside of the range of observed data).

c) The predicted value for the U.S. is –48.91 + 0.9186(81) = 25.5 with 15.0 – 25.5 = –10.5 as the corresponding residual. The regression equation underestimates the percentage of women in parliament by 10.5% for the U.S.

d) $b = 0.56(9.8/7.7) = 0.7127$ and $a = 26.5 - 0.7127(76.8) = -28.24$. Thus, the prediction equation is given by $\hat{y} = -28.24 + 0.7127x$.

3.70 African droughts and dust

a) (i) B, (ii) C, (iii) A

b) Dust and rainfall amounts are negatively related. As one increases, the other decreases.

3.71 Crime rate and urbanization

a) An increase of 100 is an increase of 100 times the slope = 0.56(100) = 56. As the urban nature of a county goes from 0 to 100, the predicted crime rate increases by 56%.

b) The correlation indicates a relatively strong, positive relationship.

c) The slope and correlation are related by the formula $b = r\left(s_y / s_x\right)$; $0.56 = 0.67\left(28.3/34.0\right)$.

3.72 Gestational period and life expectancy revisited

a) Animals with a gestational period that is 100 days longer are predicted to live 0.045(100) = 4.5 years longer.

b) Leopards are predicted to live 6.28 + 0.045(98) = 10.7 years.

c) 73% of the variability in the longevity of animals can be explained by the linear relationship between longevity and gestational period.

d) 40 weeks is 280 days. The regression equation would predict an average longevity of 6.28 + 0.045(280) = 18.9 years for humans.

3.73 Gas consumption and temperature

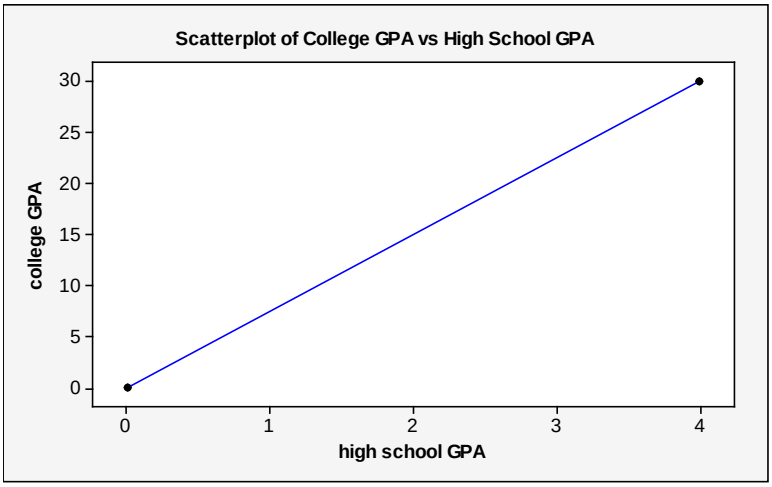
a) The response variable is monthly residential gas consumption, and the explanatory variable is average monthly temperature.

b) The slope of the regression equation is –641.79 when average monthly temperature is measured in Celsius degrees and residential gas consumption in MMCF. An increase of one degree predicts a decrease in gas consumptions of 641.79 MMCF.

c) 3209 is 5 times the slope of the regression line, an increase of 5 degrees in temperature leads to a decrease of 3209 MMCF in gas consumption.

3.74 Predicting college GPA

a)



This equation is not realistic because it predicts an increase of seven in college GPA for an increase of one in high school GPA when GPA ends at 4.0! This would predict a college GPA of 28.5 when high school GPA is 4.0.

b) $\hat{y}$ = 0.5 + 0.7(3.0) = 2.6

$\hat{y}$ = 0.5 + 0.7(4.0) = 3.3

As high school GPA goes up by 1.0 (from 3.0 to 4.0), predicted college GPA goes up by exactly the amount of the slope, 0.7 (from 2.6 to 3.3).

3.75 College GPA = high school GPA

The *y*-intercept would be zero (the line would cross the *y*-axis at zero when *x* was zero), and the slope would be one (an increase of one on *x* would mean an increase of one on *y*). In this case, the line matches up with the exact points on *x* and *y* (0.0 with 0.0, 3.5 with 3.5, etc.). This means that your predicted college GPA equals your high school GPA.

3.76 Salary and employee satisfaction

a) The slope is the amount that *y* is predicted to increase when *x* increases by one unit. In this case, employee satisfaction (*y*) increases by 1 point when annual salary (*x*) increases by $4,000. It would follow that an increase of one-dollar per year would be 1/4000 of the $4,000 yearly increase: 1/4000 = 0.00025.

b) A one-dollar increase to the monthly salary would be an increase of $12 on a yearly basis. Thus, the slope would be 12 times the slope in (a) (i.e., 0.003). We also can calculate it in the same way we did in part (a), but first dividing the yearly increase of $4,000 by 12.

3.77 Car weight and gas hogs

a) The slope indicates the change in *y* predicted for an increase of one in *x*. Thus, a 1000 increase in *x* would mean a predicted change in *y* of 1000 times the slope: (–0.0052)(1000) = –5.2 (poorer mileage).

b) $\hat{y}$ = 47.32 – 0.0052(6400) = 14.04. The actual mileage is 17; thus, the residual is 17 – 14.04 = 2.96.

The Hummer gets 2.96 more miles to the gallon than one would predict from this regression equation.

3.78 Predicting Internet use from cell phone use

a) (i) The response variable is Internet use, and the explanatory variable is cell-phone use.
(ii) The scatterplot shows a positive association.
(iii) There is little variability of internet use for cellular use below 30%; for cellular use above 30%, internet use is generally higher but also has higher variability.

b) One nation that has less Internet use than one would expect, given its level of cell-phone use is the point with approximate *x*- and *y*-coordinates of 75 and 13, respectively, or (75, 13).

c) As *x* increases from 0 to 90, predicted *y* increases from 1.3 to 44.0. This represents a positive association.

$\hat{y}$ = 1.27 + 0.475(0) = 1.3

$\hat{y}$ = 1.27 + 0.475(90) = 44.0

d) $\hat{y}$ = 1.27 + 0.475(45.1) = 22.7; the predicted Internet use for the U.S. is 22.7%.

The residual (27.5) is the difference between the actual value of 50.15 and the predicted value of 22.7; 50.15 – 22.7 = 27.5. The large positive residual indicates that the U.S. has a much higher Internet use percentage than one would predict from this regression equation.

3.79 Income depends on education?

a) For each increase of one percentage in *x*, we would expect an increase in the predicted value on *y* by 0.42. Thus, an increase in 10 would be 10 times the slope: 0.42(10) = 4.2 (or $4200).

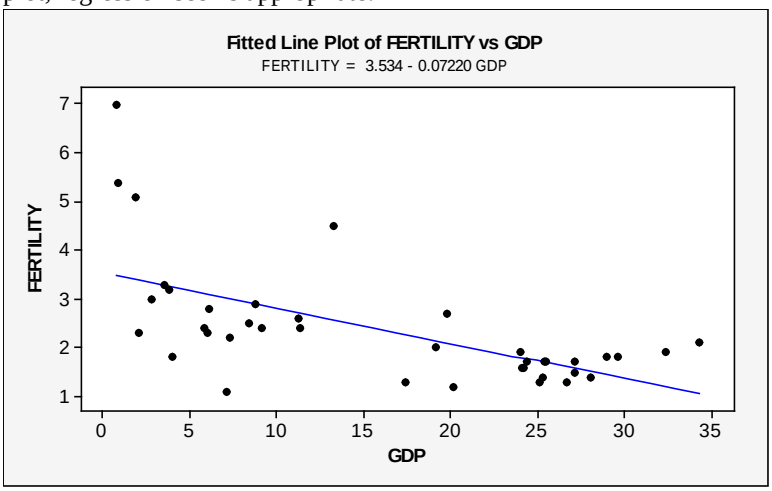
b) The slope can be calculated using the formula $b = r\left(s_y / s_x\right)$.

Thus, $0.42 = r\left(4.69/8.86\right) \Rightarrow r = 0.42\left(8.86/4.69\right) = 0.79$.

(i) The positive sign indicates a positive relationship; as one variable goes up, the other goes up. As one goes down, the other goes down.
(ii) A correlation of 0.79 indicates a strong relationship.

3.80 Fertility and GDP

a) Based on the plot, regression seems appropriate.



b) The correlation is –0.615. The regression equation is: FERTILITY = 3.53 – 0.072GDP.

c) We cannot compare slopes to determine which one indicates a bigger association because slopes are dependent on the measures used. We must use correlation which does not depend on the measure; correlation is based on standardized variables.

d) From (c). the correlation between GDP and fertility is –0.615. Thus, contraception has a stronger association with fertility than does GDP.

3.81 Women working and birth rate

a) $\hat{y} = 36.3 - 0.30(0) = 36.3$

$\hat{y} = 36.3 - 0.30(100) = 6.3$

When women's economic activity is 0%, predicted birth rate (36.3) is much higher than the predicted birthrate (6.3) when women's economic activity is 100%.

b) The correlation between birth rate and women's economic activity is bigger in magnitude than the correlation between crude birth rate and nation's GNP, indicating a larger association between birth rate and women's economic activity.

3.82 Education and income

a) $b = r\left(s_y / s_x\right) = 0.50\left(16{,}000/2\right) = 4000$, so the slope is 4000.

b) The correlation will not change because it is not dependent on which variable is considered the explanatory and which is considered the response. The slope will change in value as shown in the equation below.

$b = r\left(s_y / s_x\right) = 0.50\left(2/16{,}000\right) = 0.000625$

3.83 Income in euros

a) The intercept is –20,000 in dollars; thus, the intercept in euros is $-\$20{,}000\left(\frac{1\text{ euro}}{\$1.25}\right) = -16{,}000$ euro.

b) The slope of the regression equation is 4000 in dollars; thus, the slope in euros is $\$4000\left(\frac{1\text{ euro}}{\$1.25}\right) =$ 3200 euro.

c) The correlation remains the same when income is measured in euros because correlation is not dependent on the units used – whether dollars or euros. It is still 0.50.

3.84 Changing units for cereal data

a) SODIUM(mg) = 169 – 0.00025SUGAR(mg)
The slope changes from grams to milligrams. For every 1 milligram increase in sugar, we expect the sodium content to decrease by 0.00025 milligrams. The *y*-intercept does not change, it remains in milligrams.

b) SODIUM(mg) = 169 – 0.009SUGAR(oz)
If we change the unit of measurement for sugar from grams to ounces, we would divide the slope by the appropriate constant. The new slope can be calculated using this relationship as –0.25/28.35 = –0.009.

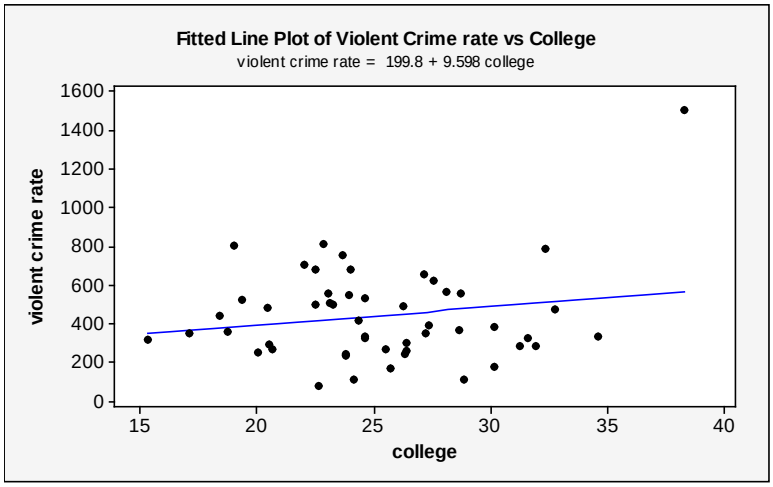
3.85 Murder and single-parent families

a) The District of Columbia is the outlier to the far, upper right. This would have an effect on the regression analysis because it is a regression outlier; that is, it is an outlier on *x* and also is somewhat out of line with the trend of the rest of the data.

b) When the District of Columbia is included, the *y*-intercept decreases and the slope increases. The District of Columbia point pulls the regression line upwards on the right side.

3.86 Violent crime and college education

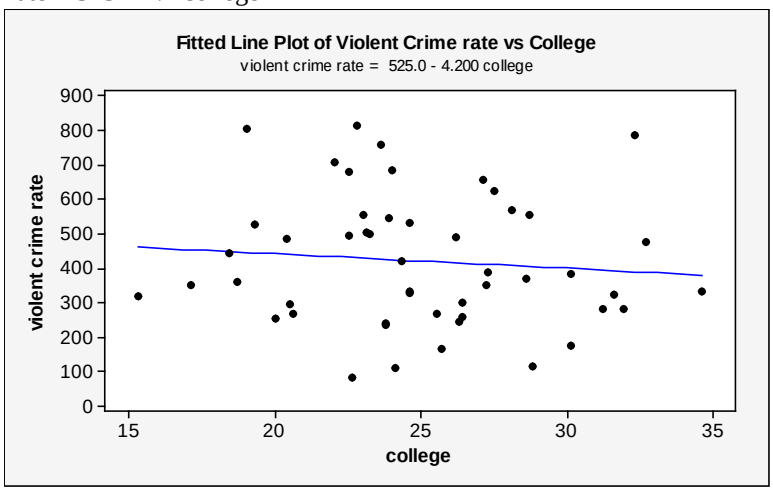
a)



The point at approximately $x = 38$ might be influential in a regression analysis.

b) See (a), violent crime rate = 200 + 9.6college
This slope suggests that for every 1% increase in college educated people, there is a predicted increase of 9.6 in the violent crime rate.

c) violent crime rate = 525 – 4.2 college

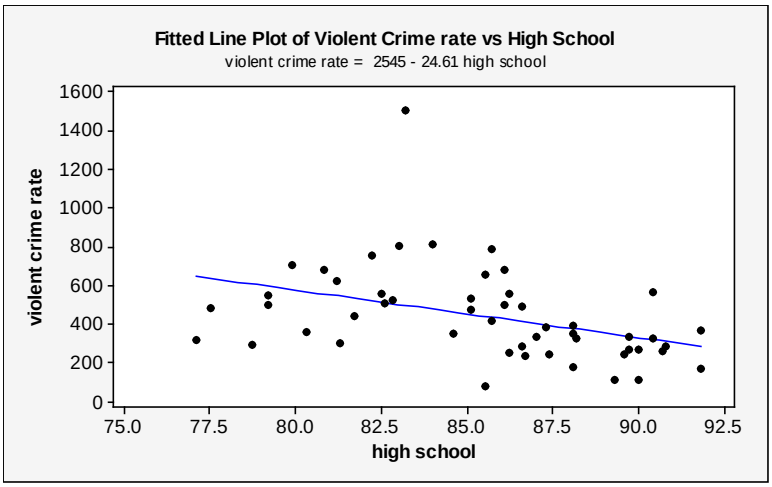


3.86 (continued)

This slope indicates that for every increase of 1% in college educated people, there is a predicted decrease of 4.2 in the violent crime rate. The deletion of one point has changed the association from a positive one to a negative one.

3.87 Violent crime and high school education

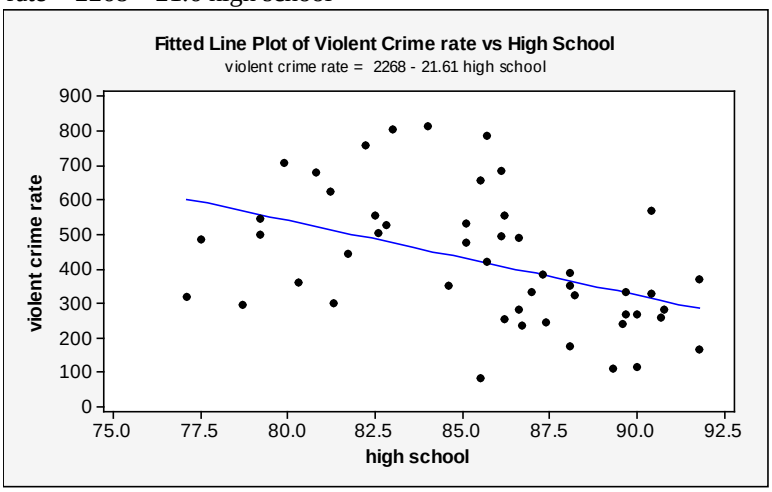
a)



The point with a *y* value of around 1500 is furthest from other data points.

b) See (a), violent crime rate = 2545 – 24.6 high school
The slope indicates that for each increase of one percent of people with a high school education, the predicted violent crime rate decreases by 24.6.

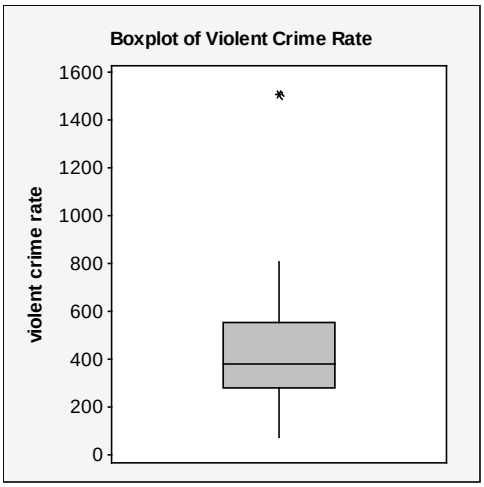
c) violent crime rate = 2268 – 21.6 high school



The slope indicates that for each increase of one percent of people with a high school education, the predicted violent crime rate decreases by 21.6. This is similar to the slope in (b).

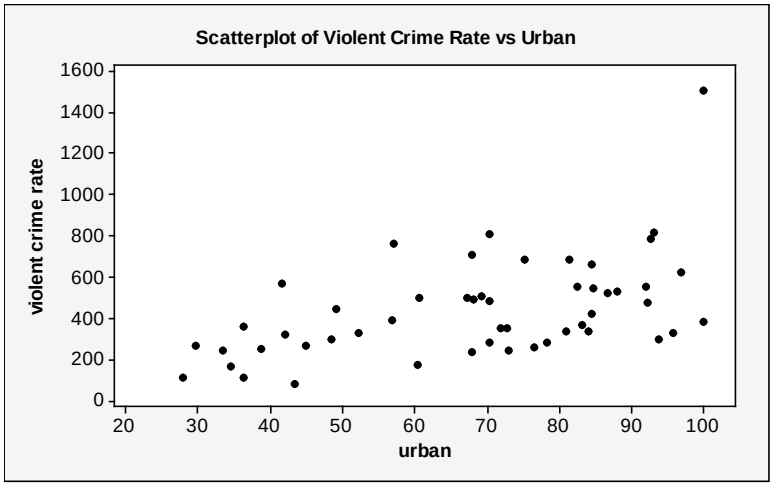
3.88 Crime and urbanization

a)



Along with the box plot, the mean of 441.6 and standard deviation of 241.4 suggest that there might be some skew; the standard deviation is fairly large compared to the mean. Because the lowest possible value is 0, scores can only be as much as 1.8 standard deviations below the mean. Thus, we expect the distribution to be skewed to the right.

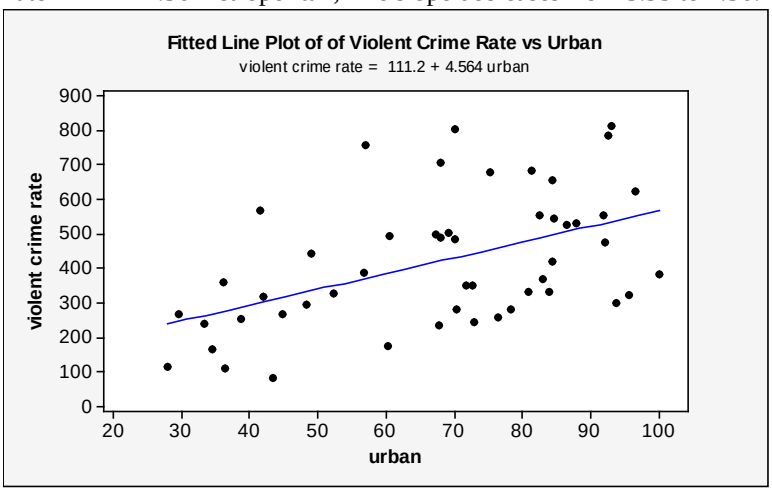
b)



The observation at around $y = 1500$ appears to be a potentially influential observation, although it is only a regression outlier on one of the two criteria; it is outside of the trend of the rest of the data, but is not far from the rest of the data on x. Without this point, we might expect that the slope would be somewhat smaller. This point is very high on both x and y and so we would expect that it would pull the regression line upwards.

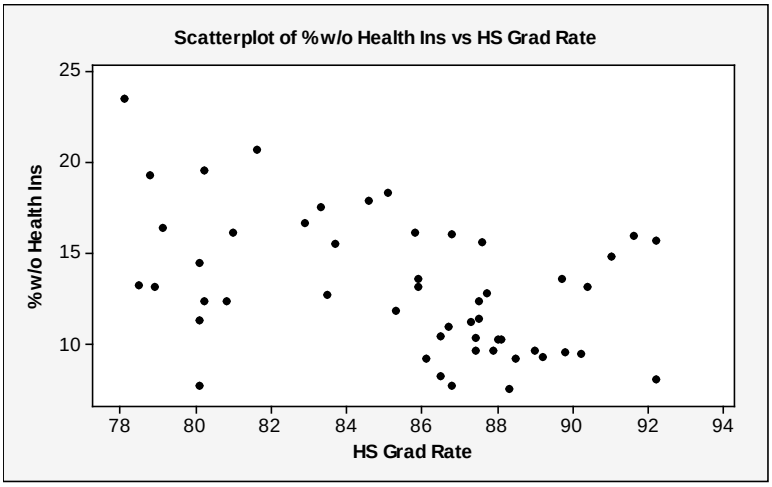
3.88 (continued)

c) violent crime rate = 111 + 4.56 metropolitan; The slope decreases from 5.93 to 4.56.



3.89 High school graduation rates and health insurance

a)



The scatterplot suggests a negative relationship.

b) The correlation is –0.45. As does the scatterplot, this correlation indicates a negative association.

c) The regression equation is: Health Insurance = 49.2 – 0.42HS_Grad_Rate. The slope of –0.42 indicates that for each increase of one in the percentage who are high school graduates, the predicted percentage of individuals without health insurance goes down by 0.42. This summarizes the negative relationship between the variables.

3.90 Women's Olympic high jumps

a) Women_Meters = –10.94 + 0.0065Year_Women

(i) For 2016: $\text{Women\_Meters} = -10.94 + 0.0065(2016) = 2.16$ meters, or 7.1 feet.

(ii) For 3000: $\text{Women\_Meters} = -10.94 + 0.0065(3000) = 8.56$ meters, or 28.1 feet.

b) Although 2016 is outside of the range of data, it is the next data point in the time sequence and the regression equation should be able to predict its value fairly well assuming there are no major changes to the sport; however, the year 3000 is too far beyond the range of the data to extrapolate.

3.91 IQ and shoe size

a) As age increases, students tend to have higher IQs and bigger foot sizes. Age could be the common cause of both of these variables.

b) If age had actually been measured, it would be a confounding variable. When measured, a lurking variable becomes a confounding variable.

3.92 More TV watching goes with fewer babies?

a) Correlation does not indicate causation. There are lots of other possible ways these two variables could be related.

b) There are several possible responses to this exercise. One possible lurking variable is GDP, because nations with higher GDP tend to have lower birth rates and higher television ownership.

3.93 More coffee protects from cancer?

a) Coffee drinkers usually drink their coffee indoors, thus limiting their exposure to the sun helps reduce skin cancer. Avoiding sun could be the lurking variable that influences both.

b) Avoiding sun might be the common cause of both of the variables reported in this study. It might actually cause people to drink more coffee indoors and cause them to reduce their exposure to skin cancer.

3.94 Ask Marilyn

a)

White Collar	Hired		
Gender	Yes	No	Total
Male	30	170	200
Female	40	160	200

Blue Collar	Hired		
Gender	Yes	No	Total
Male	300	100	400
Female	85	15	100

b) Male: 270/600 = 45%; Female: 175/300 = 58%

	Hired		
Gender	Yes	No	Total
Male	330	270	600
Female	125	175	300

c) This is an example of Simpson's paradox, the fact that the direction of an association between two variables can change after we include a third variable and analyze the data at separate levels of that variable. When the third variable of type of job (white vs. blue collar) is included in the analysis, women fare better than do men, whereas when this variable is not included, women fare worse than do men.

Chapter Problems: Concepts and Investigations

3.95 NL baseball team ERA and number of wins

The responses will be different for each student depending on the methods used.

3.96 Time studying and GPA

The responses will be different for each student depending on the methods used.

3.97 Warming in Newnan, GA

The regression equation is: Temp = 119 – 0.029Year

The regression line indicates a very slight decrease over time. The Central Park data from Example 12 indicate the opposite, a very slight increase over time.

3.98 Regression for dummies

Answers will vary. Your answer should say something like: Regression allows us to use information we currently know to develop a way to predict in the future (as long as we don't predict too far into the future because the trends might change!). For example, we know how many catalogs we mail in a given month, and we know our total sales for the next month. If we can look over the data for a whole year, we can get an idea of how well the numbers of catalogs mailed predicted total sales. The technique of regression allows us to develop an equation that we can use to do this kind of prediction. For example, we might find that the more catalogs we send out, the more sales we have. Then we can make predictions for future months so that we can have an idea of our sales.

3.99 Fluoride and AIDS

San Francisco could be higher than other cities on lots of variables, but that does not mean those variables cause AIDS, as association does not imply correlation. Alternative explanations are that San Francisco has a relatively high gay population or relatively high intravenous drug use, and AIDS is more common among gays and IV drug users.

3.100 Fish fights Alzheimer's

a) A lurking variable is a possible third variable that might affect the relationship between two other variables. In this example, those who eat fish had a lower risk of Alzheimer's disease. There might be another variable, however, that's related to both of these. For example, those who eat healthy foods like fish might also exercise; it might be exercise, rather than the fish, that leads to the lower rate of Alzheimer's.

b) There can be multiple causes for any particular response variable. In the example in (a) of this exercise, exercise and fish might both cause lower rates of Alzheimer's. A lurking variable (a third variable as described above) and the explanatory variable (the variable that we think is causing something else) might both cause the outcome of interest, in this case Alzheimer's.

c) People should be skeptical when they read new research results such as in this story because the researcher might not have considered all possible explanations for the correlation.

3.101 Dogs make you healthier

Stress level, physical activity, wealth, and social contacts are all possible lurking variables. Any one of these variables may contribute to one's physiological and psychological human health as well as be associated with whether or not a person owns a dog. For example, it may be that people who are more active are more likely to own a dog as well as being physically healthier. Thus, it is possible that one of these lurking variables is responsible for the perceived association between health and dog ownership and if they had been controlled in the study, the association would not be present.

3.102 Multiple choice: Correlate GPA and GRE

The best answer is (d).

3.103 Multiple choice: Properties of *r*

The best answer is (b).

3.104 Multiple choice: Interpreting *r*

The best answer is (a).

3.105 Multiple choice: Correct statement about *r*

The best answer is (d).

3.106 Multiple choice: Describing association between categorical variables

The best answer is (b).

3.107 Multiple choice: Slope and correlation

The best answer is (c).

3.108 Multiple choice: Interpretation of r^2

The best answer is (d).

3.109 True or false

a) False, the weakest correlation is between y and x_1.
b) True, the slope and the correlation would have the same sign.
c) True, the slope tells us that an increase of one year leads to a predicted increase of 0.4 thousands of dollars, which translates into \$400.
d) True, ten times the slope equals four. Income is in thousands of dollars and thus, this predicted increase is \$4000.

♦♦3.110 Correlation does not depend on units

a) If we convert income from British pounds to dollars, then each pound is now worth \$2. In other words, we multiply each score by two. Thus, each y-value doubles, the mean of y is now doubled, and the distance of each score from the mean doubles. If this is all so, then the variability, as represented by the standard deviation, has now doubled. As an example, imagine two scores in pounds, 5 pounds and 10 pounds. If both are converted to dollars, they are now \$10 and \$20. The mean of the two was 7.5 pounds, and it is now \$15. The distance of each of the two scores in pounds from the mean was 2.5, but now the distance of each of the two scores in dollars from the mean is \$5.
b) The correlation would not change in value, however, because the correlation is based on standardized versions of the measures, and is not affected by the measure used. The formula for the correlation uses z-scores, rather than raw scores. Thus, both pounds and dollars would be converted to z-scores and would lead to the same correlation.

♦♦3.111 When correlation = slope

If the two standard deviations are equal, $s_y / s_x = 1$, so the formula for slope is $b = \left(s_y / s_x\right) r = 1 \times r = r$.

Thus, in cases where the standard deviations are equal, mathematically the slope must equal the correlation.

♦♦3.112 Center of the data

a) Algebraically, we can manipulate the formula $a = \bar{y} - b\bar{x}$ to become $\bar{y} = a + b\bar{x}$ by isolating $\bar{y}$. The latter formula is very similar to the regression equation, except the generic predicted y, $\hat{y}$, is replaced by the mean of y, $\bar{y}$. Similarly, the generic x is replaced by $\bar{x}$. Thus, a score on any x that is at the mean will predict the mean for y.
b) Here are the algebraic steps to go from one formula to the other.
Step 1: Because we know that $a = \bar{y} - b\bar{x}$, we can replace the a in the regression equation with this formula. This yields $\hat{y} = \bar{y} - b\bar{x} + bx$.
Step 2: We can now subtract $\bar{y}$ from both sides. It cancels itself out on the right, and is now subtracted from the left. This yields $\hat{y} - \bar{y} = -b\bar{x} + bx$, or $\hat{y} - \bar{y} = bx - b\bar{x}$, (if we switch the two parts of the right hand side of the equation)
Step 3: Finally, we can take the b on the right and put it outside parentheses to denote that it is to be multiplied by both variables within the parenthesis. This yields $\hat{y} - \bar{y} = b\left(x - \bar{x}\right)$.
This formula tells us that if we figure out how far from the mean our x is, we can multiple that deviation by the slope to figure out how far from the mean the predicted y is.

♦♦3.113 Final exam "regress toward mean" of midterm

a) As we saw in (b) of Exercise 3.112, $\hat{y} - \bar{y} = b\left(x - \bar{x}\right)$ is mathematically equivalent to the usual equation we use for regression, $\hat{y} = a + bx$. We also know from Exercise 3.111 that when standard deviations are equal, the slope equals the correlation. If we fill in 0.70 for b, we obtain $\hat{y} - \bar{y} = 0.70\left(x - \bar{x}\right)$.

3.113 (continued)

b) This means that the predicted difference between one's final exam grade and the mean for the class is 70% of the difference between your midterm exam score and the mean for the class. For example, if the class mean were 80 and your score were 90, you'd be 10 points above the mean. If you multiplied that by 0.70, you'd predict that you'd deviate from the mean on *y* by seven points. If the mean on *y* also were 80, your predicted score would be 87. Thus, your predicted score is closer to the mean; it regresses, or comes back, toward the mean.

Chapter Problems: Student Activities

3.114 Analyze your data

The responses to this exercise will vary for each class depending on the data files that each class constructed.

3.115 Activity: Effect of moving a point

The responses to this exercise will vary depending on the data points provided by the instructor.

3.116 Activity: Guess the correlation and regression

The responses to this exercise will vary depending on the randomly generated data points.