Chapter 10: Correlation and Regression

Section 10-1: Correlation

1. a. r is a statistic that represents the value of the linear correlation coefficient computed from the paired sample data, and ρ is a parameter that represents the value of the linear correlation coefficient that would be computed by using all of the paired data in the population of all statistics students.

   b. The value of r is estimated to be 0, because it is likely that there is no correlation between body temperature and head circumference.

   c. The value of r does not change if the body temperatures are converted to Fahrenheit degrees.

2. No, with $r = 0,$ there is no *linear* correlation, but there might be some association with a scatterplot showing a distinct pattern that is not a straight-line pattern.

3. No, a correlation between two variables indicates that they are somehow associated, but that association does not necessarily imply that one of the variables has a direct effect on the other variable. Correlation does not imply causality.

4. a. -1
   b. 0.746
   c. 0.268
   d. 0.992
   e. 1

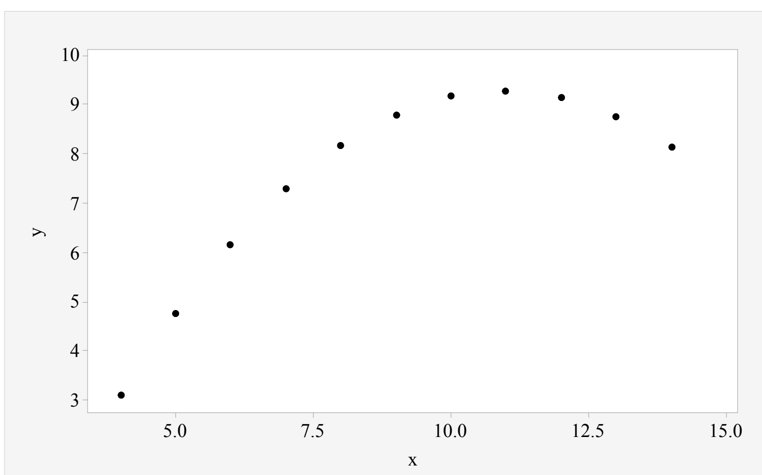
5. $r = 0.963;$ P-value $= 0.000;$ Critical values: $r = \pm 0.268$ (Table: $r \approx \pm 0.279$); Yes, there is sufficient evidence to support the claim that there is a linear correlation between the weights of bears and their chest sizes. It is easier to measure the chest size of a bear than the weight, which would require lifting the bear onto a scale. It does appear that chest size could be used to predict weight.

6. $r = 0.765;$ P-value < 0.01 (Using table A-6); Critical values: $r = \pm 0.497;$ Yes, there is sufficient evidence to support the claim that there is a linear correlation between calories and sugar in a gram of cereal.

7. $r = 0.552;$ P-value $< 0.0001;$ Critical values: $r \approx \pm 0.196;$ Yes, there is sufficient evidence to support the claim that there is a linear correlation between the heights of fathers and the heights of their sons.

8. $r = 0.354;$ P-value $< 0.0001;$ Critical values: $r \approx \pm 0.196;$ Yes, there is sufficient evidence to support the claim that there is a linear correlation between the heights of mothers and the heights of their daughters.

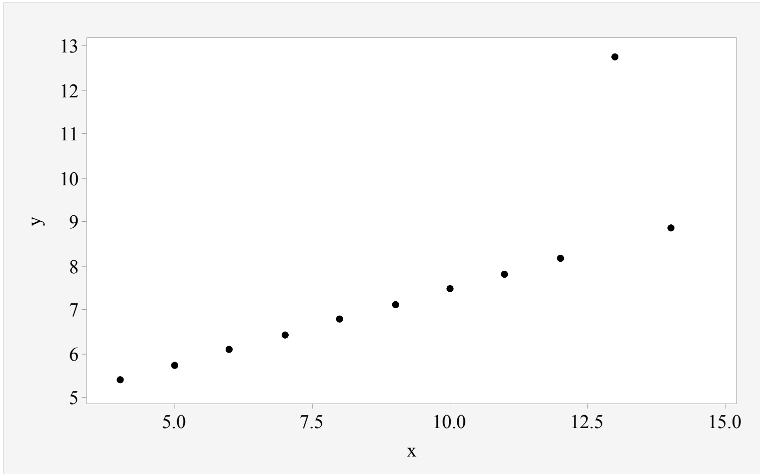
9. a.



   b. H_0: $\rho = 0$; H_1: $\rho \neq 0$; $r = 0.816;$ P-value $= 0.002$ (Table: P-value < 0.01);
   Critical values ($\alpha = 0.05$): $r = \pm 0.602;$ There is sufficient evidence to support the claim of a linear correlation between the two variables.

   c. The scatterplot reveals a distinct pattern that is not a straight line pattern.

10. a.



b. H_0: $\rho = 0$; H_1: $\rho \neq 0$; $r = 0.816$; P-value $= 0.002$ (Table: P-value < 0.01);
Critical values ($\alpha = 0.05$): $r = \pm 0.602$; There is sufficient evidence to support the claim of a linear correlation between the two variables.

c. The scatterplot reveals a perfect straight-line pattern, except for the presence of one outlier.

11. a. Answer will vary, but because there appears to be an upward pattern, it is reasonable to think that there is a linear correlation.

b. H_0: $\rho = 0$; H_1: $\rho \neq 0$; $r = 0.906$; Critical values ($\alpha = 0.05$): $r = \pm 0.632$; P-value $= 0.000$ (Table: P-value < 0.01); There is sufficient evidence to support the claim of a linear correlation.

c. H_0: $\rho = 0$; H_1: $\rho \neq 0$; $r = 0$; Critical values ($\alpha = 0.05$): $r = \pm 0.666$; P-value $= 1.000$ (Table: P-value > 0.05); There is not sufficient evidence to support the claim of a linear correlation.

d. The effect from a single pair of values can be very substantial, and it can change the conclusion.

12. a. There does not appear to be a linear correlation.

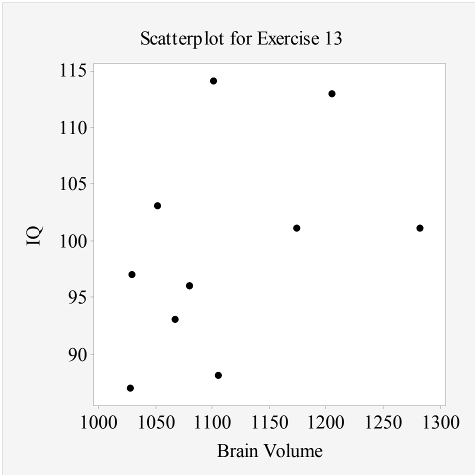
b. There does not appear to be a linear correlation.

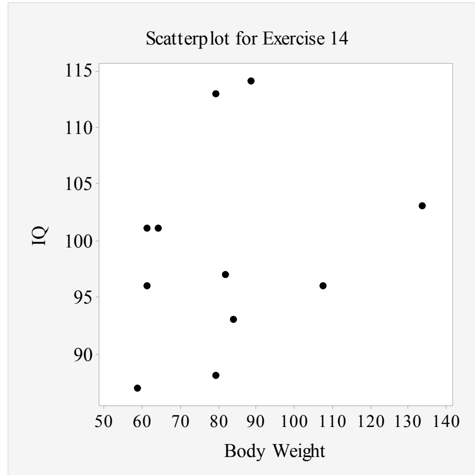
c. H_0: $\rho = 0$; H_1: $\rho \neq 0$; $r = 0$; Critical values ($\alpha = 0.05$): $r = \pm 0.950$; P-value $= 1.000$ (Table: P-value > 0.05); There does not appear to be a linear correlation. The same results are obtained with the four points in the upper right corner.

d. H_0: $\rho = 0$; H_1: $\rho \neq 0$; $r = 0.985$; Critical values ($\alpha = 0.05$): $r = \pm 0.707$; P-value $= 0.000$ (Table: P-value < 0.01); There is sufficient evidence to support the claim of a linear correlation.

e. There are two different populations that should be considered separately. It is misleading to use the combined data from women and men and conclude that there is a relationship between x and y.

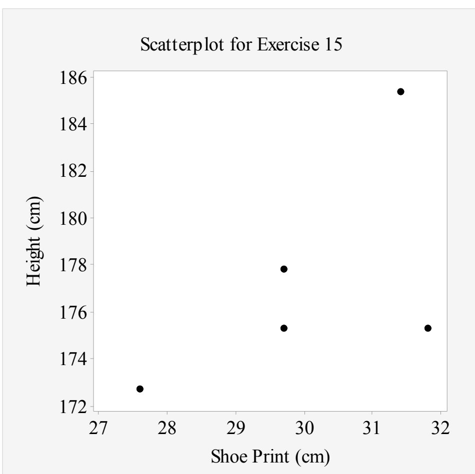
13. H_0: $\rho = 0$; H_1: $\rho \neq 0$; $r = 0.441$; P-value = 0.174 (Table: P-value > 0.05); Critical values: $r = \pm 0.602$; Fail to reject H_0. There is not sufficient evidence to support the claim that there is a linear correlation between IQ scores and brain volumes. It does not appear that people with larger brains have higher IQ scores.

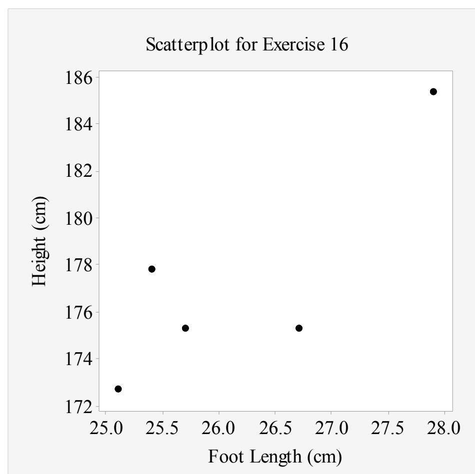




14. H_0: $\rho = 0$; H_1: $\rho \neq 0$; $r = 0.240$; P-value = 0.478 (Table: P-value > 0.05); Critical values: $r = \pm 0.602$; Fail to reject H_0. There is not sufficient evidence to support the claim that there is a linear correlation between IQ scores and body weights. It does not appear that heavier people have higher IQ scores.

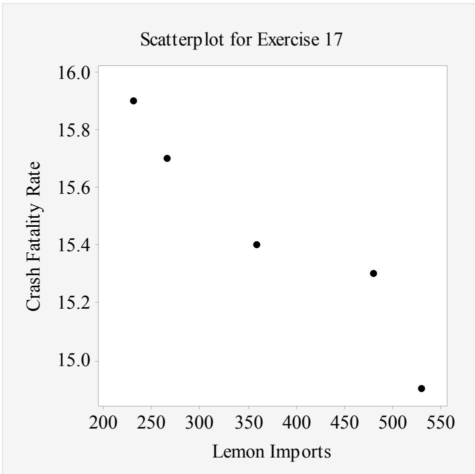
15. H_0: $\rho = 0$; H_1: $\rho \neq 0$; $r = 0.591$; P-value = 0.294 (Table: P-value > 0.05); Critical values: $r = \pm 0.878$; Fail to reject H_0. There is not sufficient evidence to support the claim that there is a linear correlation between shoe print lengths and heights of males. The given results do not suggest that police can use a shoe print length to estimate the height of a male.

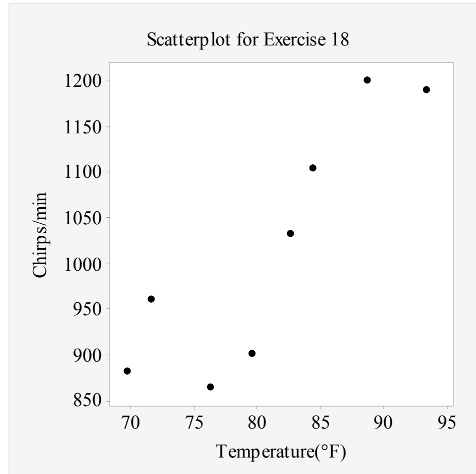




16. H_0: $\rho = 0$; H_1: $\rho \neq 0$; $r = 0.827$; P-value = 0.084 (Table: P-value > 0.05); Critical values: $r = \pm 0.878$; Fail to reject H_0. There is not sufficient evidence to support the claim that there is a linear correlation between foot lengths and heights of males. The given results do not suggest that police can use a foot length to estimate the height of a male.

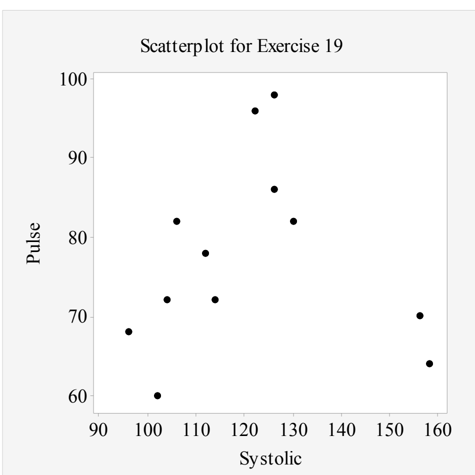
17. H_0: $\rho = 0$; H_1: $\rho \neq 0$; $r = -0.959$; P-value $= 0.010$; Critical values: $r = \pm 0.878$; Reject H_0. There is sufficient evidence to support the claim that there is a linear correlation between weights of lemon imports from Mexico and U.S. car fatality rates. The results do not suggest any cause-effect relationship between the two variables.

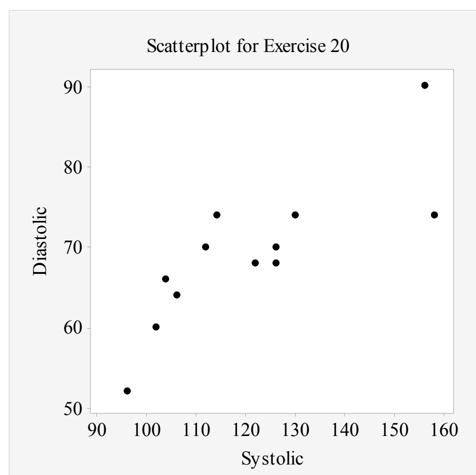




18. H_0: $\rho = 0$; H_1: $\rho \neq 0$; $r = 0.874$; P-value $= 0.005$ (Table: P-value < 0.01); Critical values: $r = \pm 0.707$; Reject H_0. There is sufficient evidence to support the claim of a linear correlation between the number of cricket chirps and the temperature.

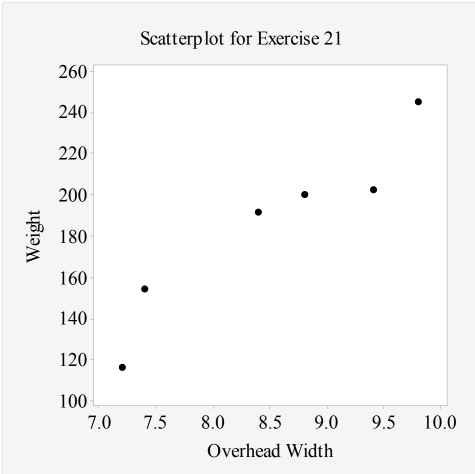
19. H_0: $\rho = 0$; H_1: $\rho \neq 0$; $r = 0.028$; P-value $= 0.932$ (Table: P-value > 0.05); Critical values: $r = \pm 0.576$; Fail to reject H_0. There is not sufficient evidence to support the claim that there is a linear correlation between pulse rates and systolic blood pressures of adult females.

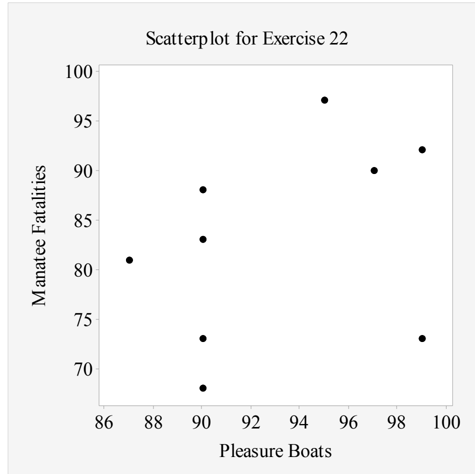




20. H_0: $\rho = 0$; H_1: $\rho \neq 0$; $r = 0.819$; P-value $= 0.001$ (Table: P-value < 0.01); Critical values: $r = \pm 0.576$; Reject H_0. There is sufficient evidence to support the claim that there is a linear correlation between diastolic and systolic blood pressures of adult females.

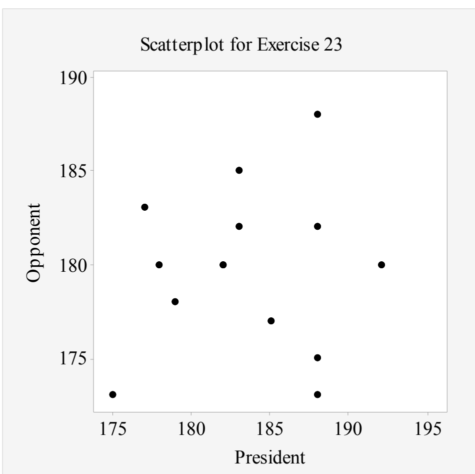
21. H_0: $\rho = 0$; H_1: $\rho \neq 0$; $r = 0.948$; P-value = 0.004 (Table: P-value < 0.01); Critical values: $r = \pm 0.811$; Reject H_0. There is sufficient evidence to support the claim of a linear correlation between the overhead width of a seal in a photograph and the weight of a seal.

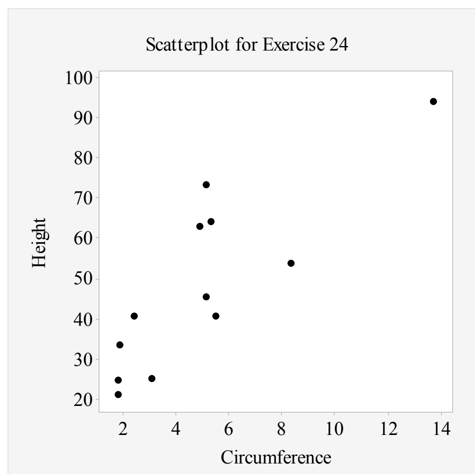




22. H_0: $\rho = 0$; H_1: $\rho \neq 0$; $r = 0.341$; P-value = 0.369 (Table: P-value > 0.05); Critical values: $r = \pm 0.666$; Fail to reject H_0. There is not sufficient evidence to support the claim that there is a significant linear correlation between numbers of registered pleasure boats and numbers of manatee boat fatalities.

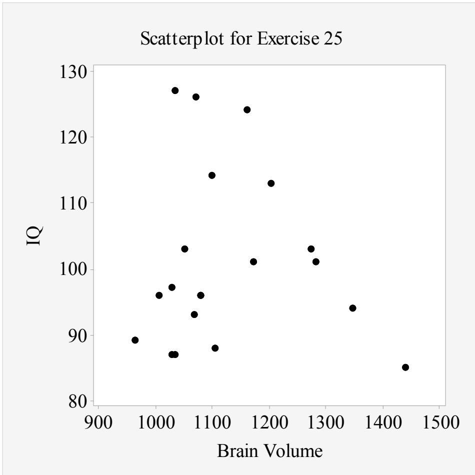
23. H_0: $\rho = 0$; H_1: $\rho \neq 0$; $r = 0.113$; P-value = 0.700 (Table: P-value > 0.05); Critical values: $r = \pm 0.532$; Fail to reject H_0. There is not sufficient evidence to support the claim that there is a linear correlation between heights of winning presidential candidates and heights of their main opponents. In an ideal world, voters would focus on important issues and not height or physical appearance of candidates, so there should not be a correlation.

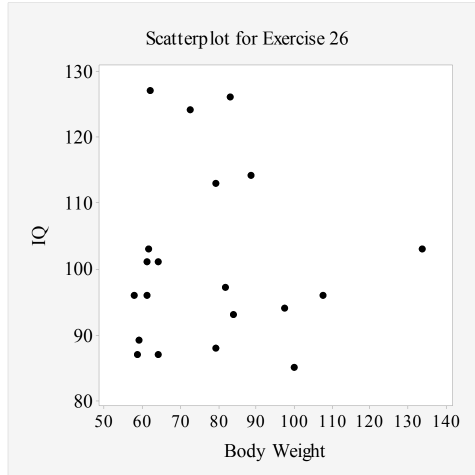




24. H_0: $\rho = 0$; H_1: $\rho \neq 0$; $r = 0.831$; P-value = 0.001 (Table: P-value < 0.01); Critical values: $r = \pm 0.576$; Reject H_0. There is sufficient evidence to support the claim of a linear correlation between circumferences and heights of trees. As trees grow, the become taller and wider, so there should be a correlation.

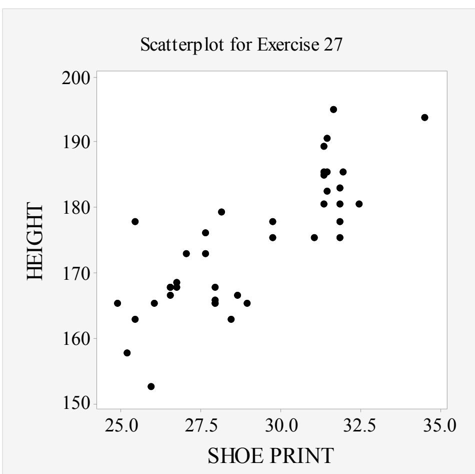
25. H_0: $\rho = 0$; H_1: $\rho \neq 0$; $r = -0.063$; P-value $= 0.791$ (Table: P-value > 0.05); Critical values: $r = \pm 0.444$; Fail to reject H_0. There is not sufficient evidence to support the claim that there is a linear correlation between IQ scores and brain volumes.

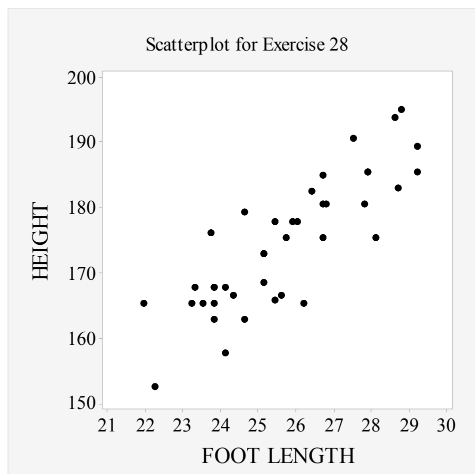




26. H_0: $\rho = 0$; H_1: $\rho \neq 0$; $r = -0.003$; P-value $= 0.991$ (Table: P-value > 0.05); Critical values: $r = \pm 0.444$; Fail to reject H_0. There is not sufficient evidence to support the claim that there is a linear correlation between IQ scores and body weight.

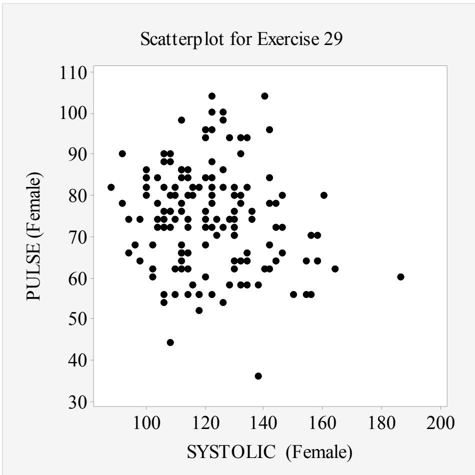
27. H_0: $\rho = 0$; H_1: $\rho \neq 0$; $r = 0.594$; P-value $= 0.007$ (Table: P-value < 0.01); Critical values: $r \approx \pm 0.456$; Reject H_0. There is sufficient evidence to support the claim that there is a linear correlation between shoe print lengths and heights of males. The given results do suggest that police can use a shoe print length to estimate the height of a male.

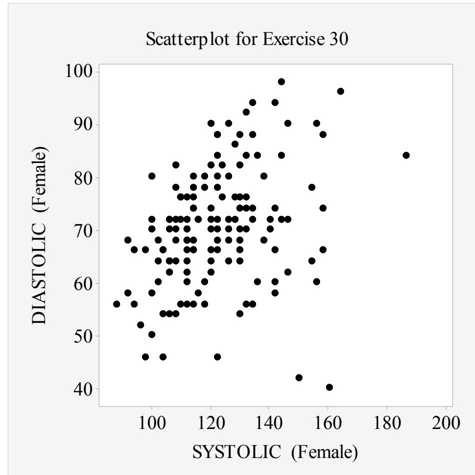




28. H_0: $\rho = 0$; H_1: $\rho \neq 0$; $r = 0.706$; P-value $= 0.001$ (Table: P-value < 0.01); Critical values: $r \approx \pm 0.456$; Reject H_0. There is sufficient evidence to support the claim that there is a linear correlation between foot lengths and heights of males. The given results do suggest that police can use a foot length to estimate the height of a male.

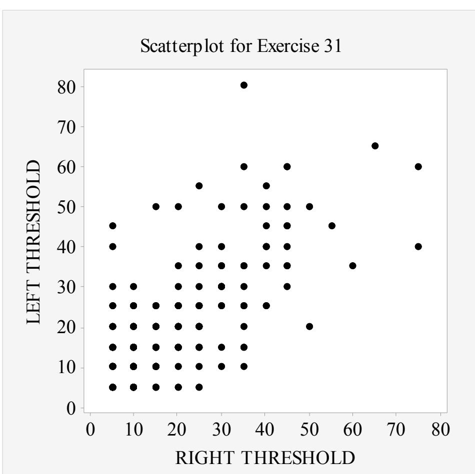
29. H_0: $\rho = 0$; H_1: $\rho \neq 0$; $r = -0.16217$; P-value = 0.0497; Critical values: $r = \pm 0.16197$; Reject H_0. There is sufficient evidence to support the claim that there is a linear correlation between pulse rates and systolic blood pressures of adult females. (In this exercise, extra decimal places are needed for r and the P-value. Table A-6 is not adequate to determine the critical values and P-value for this exercise.)





30. H_0: $\rho = 0$; H_1: $\rho \neq 0$; $r = 0.321$; P-value = 0.000; Critical values: $r = \pm 0.158$; Reject H_0. There is sufficient evidence to support the claim that there is a linear correlation between the diastolic and systolic blood pressures of adult females. (In this exercise, extra decimal places are needed for r and the P-value. Table A-6 is not adequate to determine the critical values and P-value for this exercise.)

31. H_0: $\rho = 0$; H_1: $\rho \neq 0$; $r = 0.748$; P-value = 0.000; Critical values: $r = \pm 0.105$; Reject H_0. There is sufficient evidence to support the claim that there is a linear correlation between right ear threshold measurements and left ear threshold measurements. (Table A-6 is not adequate to determine the critical values and P-value for this exercise.)



32. H_0: $\rho = 0$; H_1: $\rho \neq 0$; $r = 0.595$; P-value = 0.000; Critical values: $r = \pm 0.113$; Reject H_0. There is sufficient evidence to support the claim of a linear correlation between the visual acuity of the left eye and the right eye. (Table A-6 is not adequate to determine the critical values and P-value for this exercise.)

33. a. 0.911
b. 0.787
c. 0.9999 (largest)
d. 0.976
e. –0.948

y	x	x^2	$\log x$	$\sqrt{x}$	$1/x$
0.3	2	4	0.3010	1.4142	0.5
05	3	9	0.4771	1.7321	0.3333
1.3	20	400	1.3010	4.4721	0.05
1.7	50	2500	1.6990	7.0711	0.02
2.0	95	9025	1.9777	9.7468	0.0105

34. $r=\pm\dfrac{t}{\sqrt{t^2+n-2}}=\pm\dfrac{2.485}{\sqrt{2.485^2+27-2}}=\pm 0.445$

Section 10-2: Regression

1. a. $\hat{y}=98.3-0.001x$
b. $\hat{y}$ represents the predicted value of price from rating.

2. The first equation represents the regression line that best fits *sample* data, and the second equation represents the regression line that best fits all paired data in a population. The values b_0 and b_1 are statistics; the values β_0 and β_1 are parameters.

3. a. A residual is a value of $y-\hat{y}$, which is the difference between an observed value of y and a predicted value of y.
b. The regression line has the property that the sum of squares of the residuals is the lowest possible sum.

4. The value of r and the value of b_1 have the same sign. They are both positive or they are both negative or they are both 0. If r is positive, the regression line has a positive slope and rises from left to right. If r is negative, the slope of the regression line is negative and it falls from left to right.

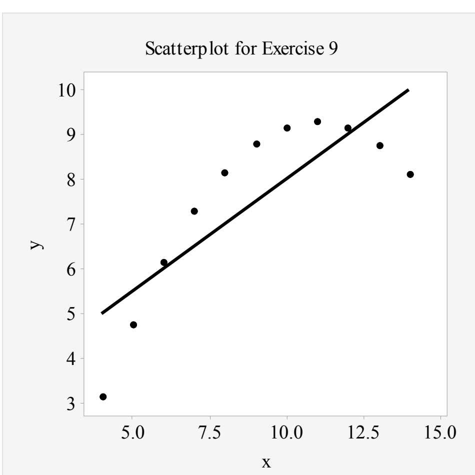
5. With no significant linear correlation, the best predicted value is $\bar{y}=161.69$ cm.

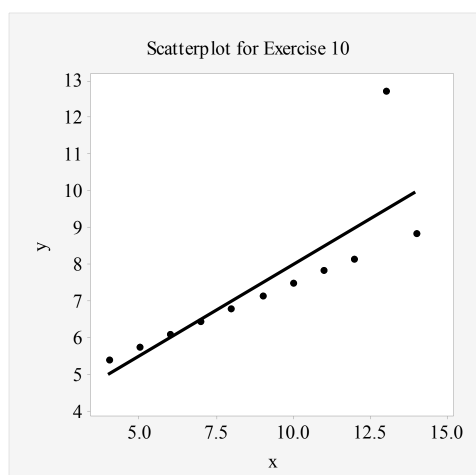
6. With a significant linear correlation, the best predicted value is $\hat{y}=-212+61.9(6.5)=190$ lb.

7. With a significant linear correlation, the best predicted value is $\hat{y}=-106+1.10(180)=92.0$ kg.

8. With no significant linear correlation, the best predicted value is $\bar{y}=6.75\,(1000\text{cells}/\mu\text{L})$.

9. $\hat{y}=3.00+0.500x$; The data have a pattern that is not a straight line.





10. $\hat{y} = 3.00 + 0.500x$; There is an outlier.

11. a. $\hat{y} = 0.264 + 0.906x$

b. $\hat{y} = 2 + 0x$ (or $\hat{y} = 2$)

c. The results are very different, indicating that one point can dramatically affect the regression equation.

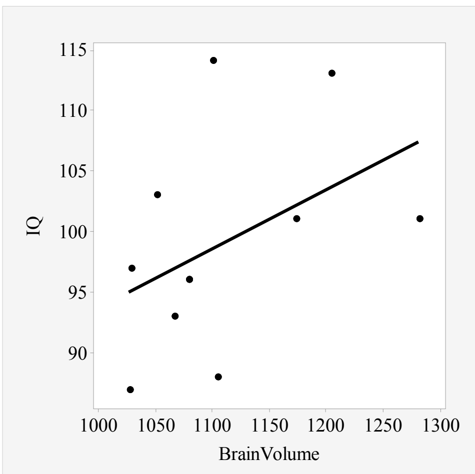
12. a. $\hat{y} = 0.0846 + 0.985x$

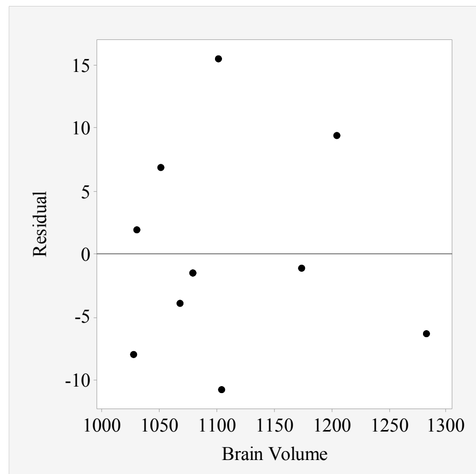
b. $\hat{y} = 1.5 + 0x$ (or $\hat{y} = 1.5$)

c. $\hat{y} = 9.5 + 0x$ (or $\hat{y} = 9.5$)

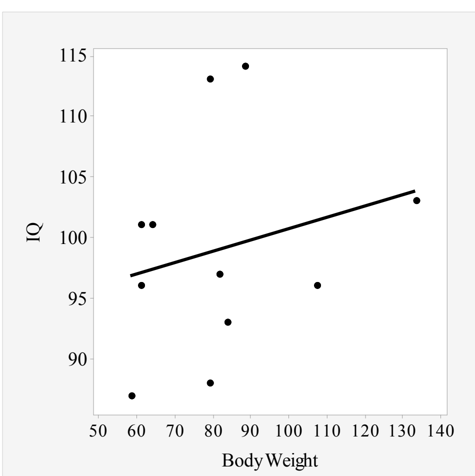
d. The results are very different, indicating that combinations of clusters can produce results that differ dramatically from results within each cluster alone.

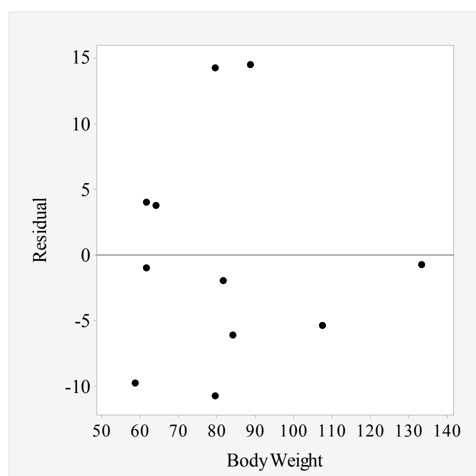
13. $\hat{y} = 44.9 + .0488x$; $r = 0.441$; P-value $= 0.174$; With no significant linear correlation, the best predicted value is $\overline{y} = 99.0$.



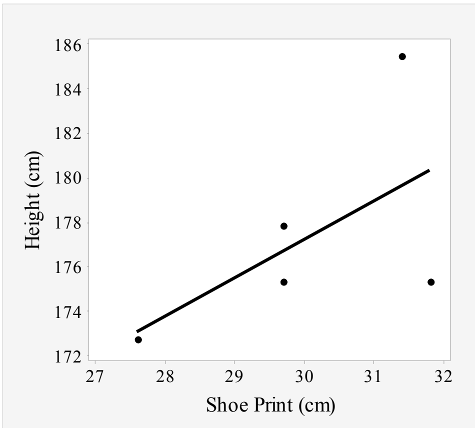


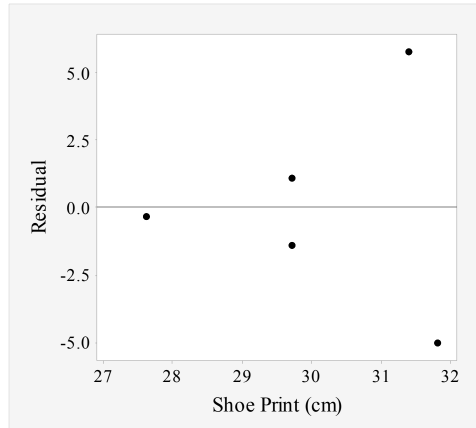
14. $\hat{y} = 91.4 + 0.0931x$; $r = 0.240$; P-value $= 0.478$; With no significant linear correlation, the best predicted value is $\overline{y} = 99.0$. The best predicted value is not close to the actual time of 88.



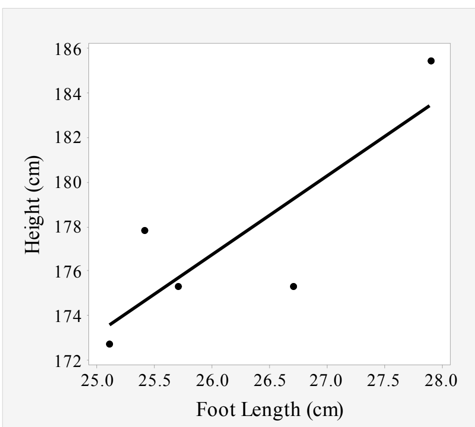


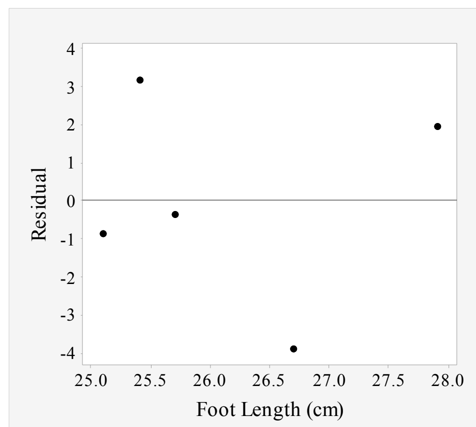
15. $\hat{y} = 125 + 1.73x$; $r = 0.591$; P-value $= 0.294$; With no significant linear correlation, the best predicted value is $\bar{y} = 177$ cm. Because the best predicted value is the mean height, it would not be helpful to police in trying to obtain a description of the male.



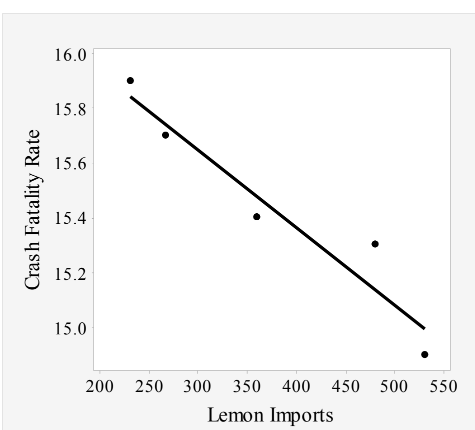


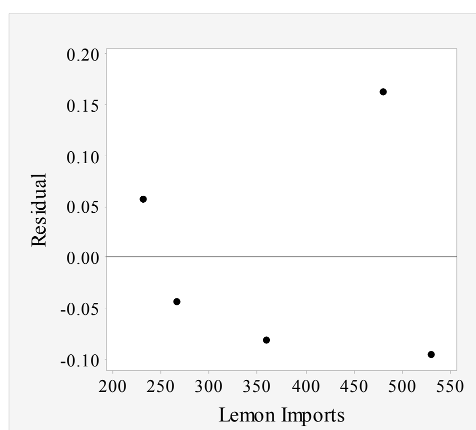
16. $\hat{y} = 85.1 + 3.52x$; $r = 0.827$; P-value $= 0.084$; With no significant linear correlation, the best predicted value is $\bar{y} = 177$ cm. Because the best predicted value is the mean height, it would not be helpful to police in trying to obtain a description of the male.



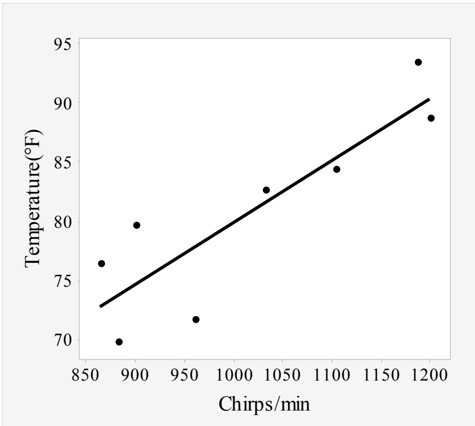


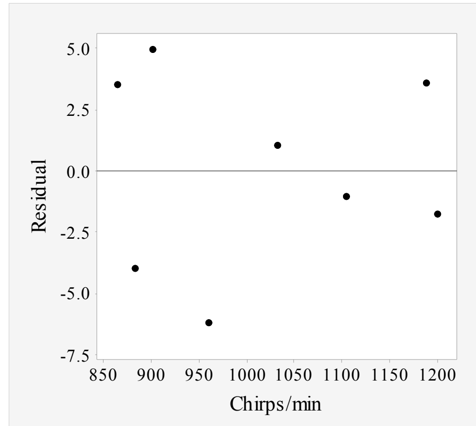
17. $\hat{y} = 16.5 - 0.00282x$; $r = -0.959$; P-value $= 0.010$; With a significant linear correlation, the best predicted value is $\hat{y} = 16.5 - 0.00282(500) = 15.1$ fatalities per 100,000 population. Common sense suggests that the prediction doesn't make much sense.



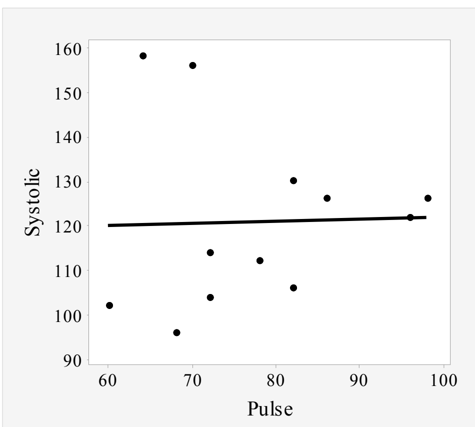


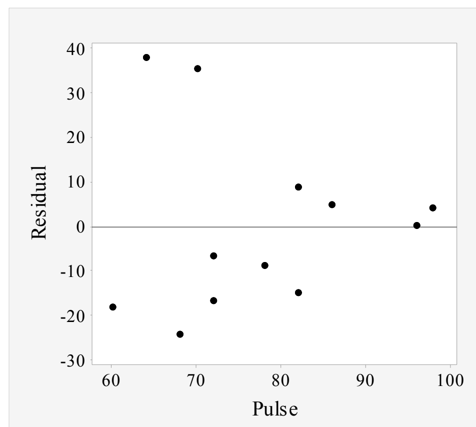
18. $\hat{y} = 27.6 + 0.0523x$; $r = 0.874$; P-value $= 0.005$; With a significant linear correlation, the best predicted value is $\hat{y} = 27.6 + 0.0523(3000) = 185°\text{F}$. The value of 3000 chirps in one minute is well beyond the scope of the listed sample data, so the extrapolation might be off by a considerable amount, especially if the cricket is dead from such a high temperature.



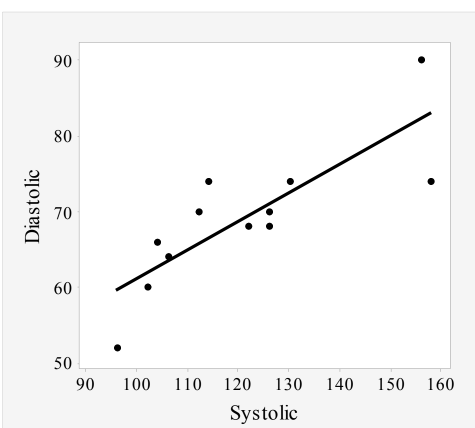


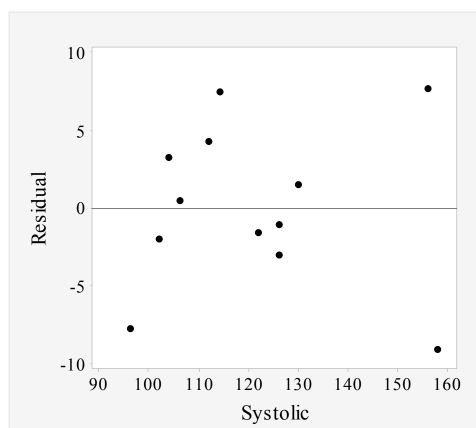
19. $\hat{y} = 117 + 0.0458x$; $r = 0.028$; P-value $= 0.932$; With no significant linear correlation, the best predicted value is $\bar{y} = 121.0$ mm Hg.



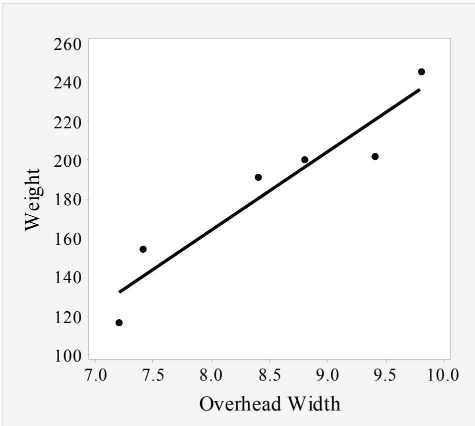


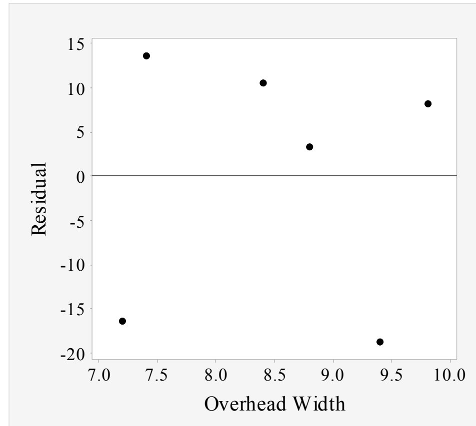
20. $\hat{y} = 23.6 + 0.377x$; $r = 0.819$; P-value $= 0.001$; With a significant linear correlation, the best predicted value is $\hat{y} = 23.6 + 0.377(120) = 68.84$ mm Hg.



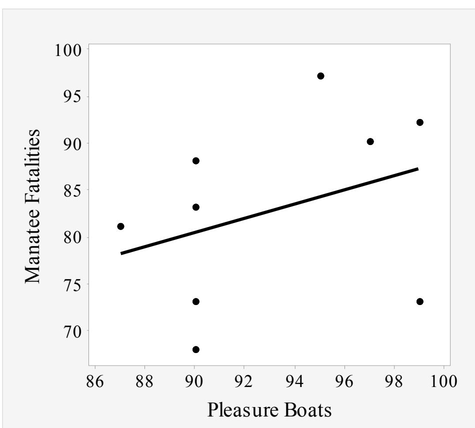


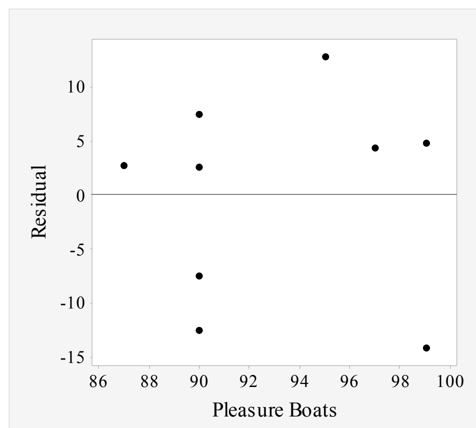
21. $\hat{y} = -157 + 40.2x$; $r = 0.948$; P-value $= 0.004$; With a significant linear correlation, the best predicted value is $\hat{y} = -157 + 40.2(2) = -76.6$ kg; The prediction is a negative weight that cannot be correct. The overhead width of 2 cm is well beyond the scope of the sample widths, so the extrapolation might be off by a considerable amount. Clearly, the predicted negative weight makes no sense.



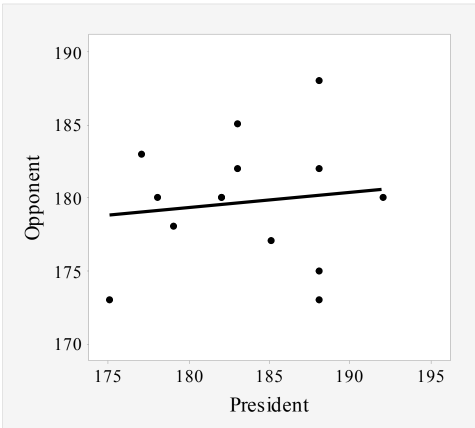


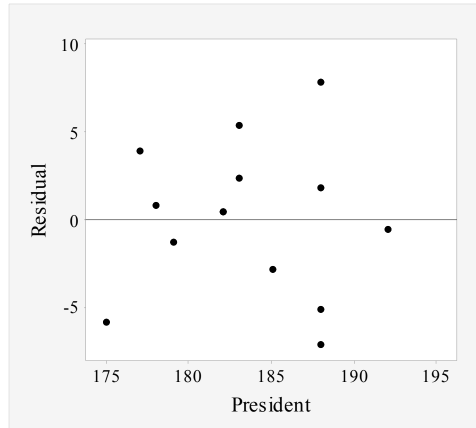
22. $\hat{y} = 13.6 + 0.744x$; $r = 0.341$; P-value $= 0.369$; With no significant linear correlation, the best predicted value is $\overline{y} = 83$ fatalities, which happens to be somewhat close to the predicted value of 79 fatalities. Because in this case, the best predicted number of fatalities is always the mean, the predicted values are not likely to be very good in general.



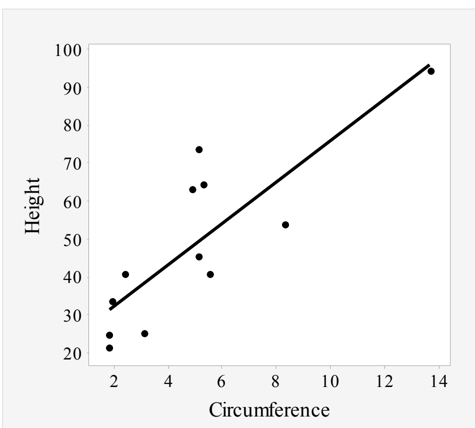


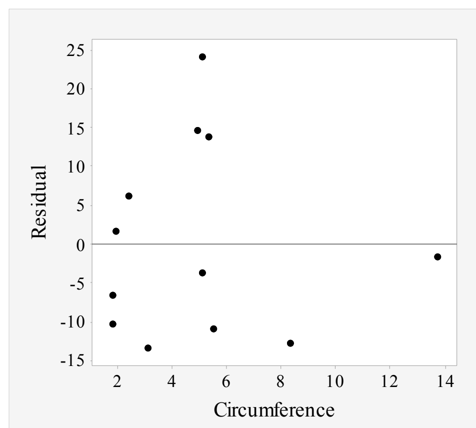
23. $\hat{y} = 162 + 0.0975x$; $r = 0.113$; P-value $= 0.700$ (Table: P-value > 0.05); With no significant linear correlation, the best predicted value is $\overline{y} = 179.7$ cm. Heights of opponents do not appear to be predicted well by using the heights of the presidents.



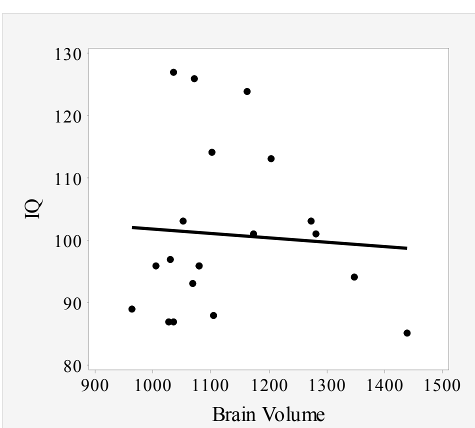


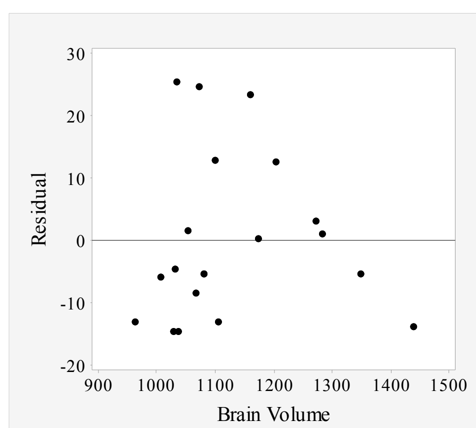
24. $\hat{y} = 21.6 + 5.40x$; $r = 0.831$; P-value $= 0.001$; With a significant linear correlation, the best predicted value is $\hat{y} = 21.6 + 5.40(5.0) = 48.6$ ft. It is much easier to measure the circumference of a tree to predict the height of the tree than having to measure the height of the tree directly.



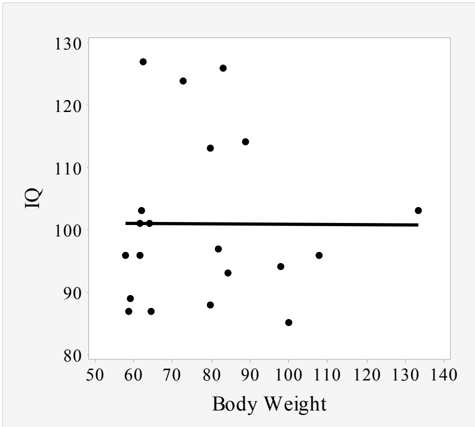


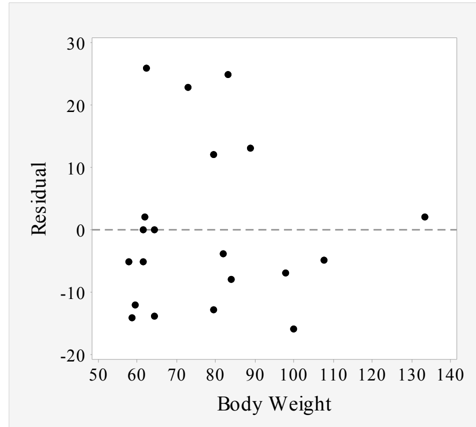
25. $\hat{y} = 109 - 0.00670x$; $r = -0.063$; P-value $= 0.791$; With no significant linear correlation, the best predicted value is $\overline{y} = 101.0$.



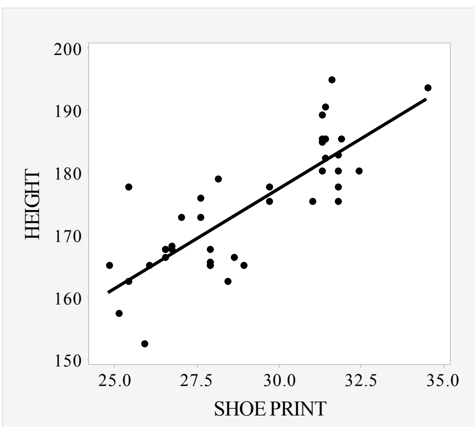


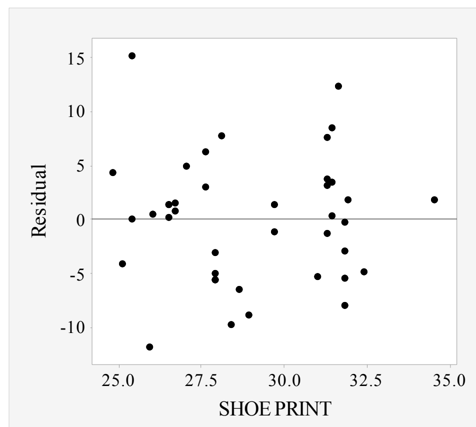
26. $\hat{y} = 101 - 0.00178x$; $r = -0.003$; P-value $= 0.991$; With no significant linear correlation, the best predicted value is $\bar{y} = 101.0$. The best predicted value is not close to the actual IQ of 88.



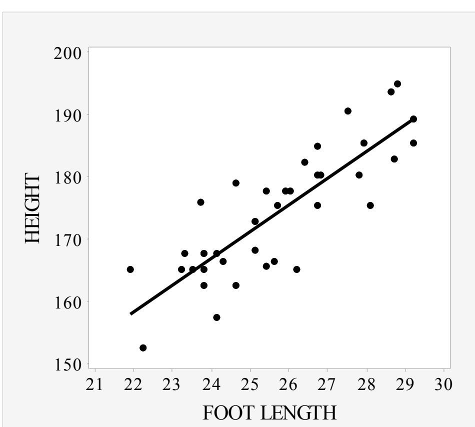


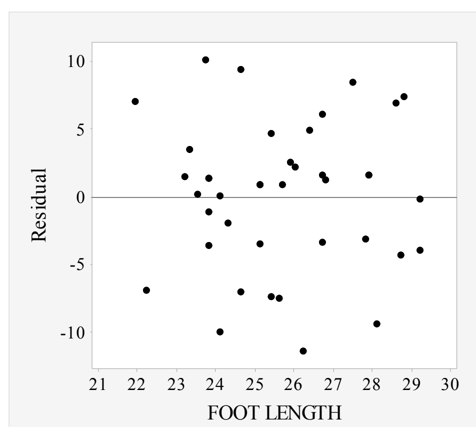
27. $\hat{y} = 93.5 + 2.85x$; $r = 0.594$; P-value $= 0.007$; With a significant linear correlation, the best predicted value is $\hat{y} = 93.5 + 2.85(31.3) = 183$ cm. Although there is a linear correlation, with $r = 0.594$, we see that it is not very strong, so an estimate of the height of a male might be off by a considerable amount.



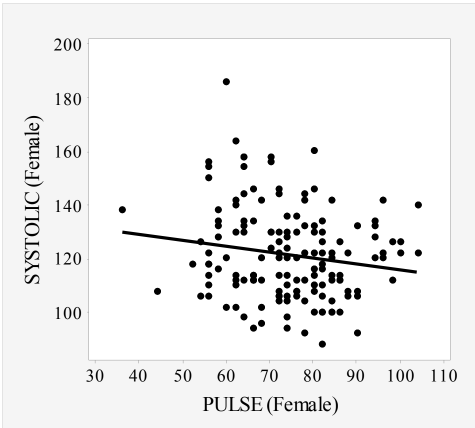


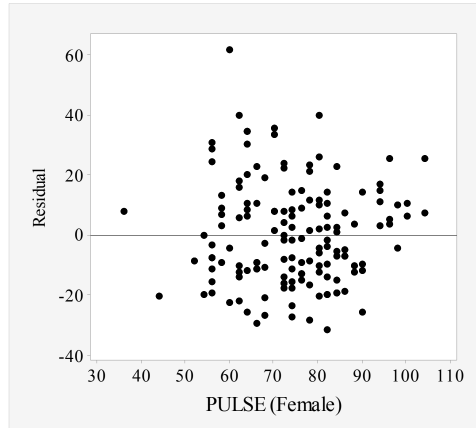
28. $\hat{y} = 87.1 + 3.50x$; $r = 0.706$; P-value $= 0.001$; With no significant linear correlation, the best predicted value is $\hat{y} = 87.1 + 3.50(28.0) = 185$ cm. Although there is a linear correlation, with $r = 0.706$, we see that it is not very close to 1, so an estimate of the height of a male might be off by a considerable amount.



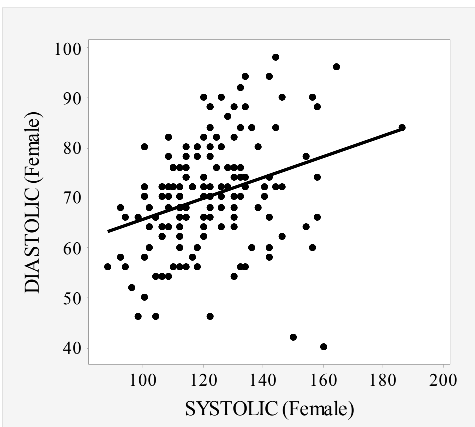


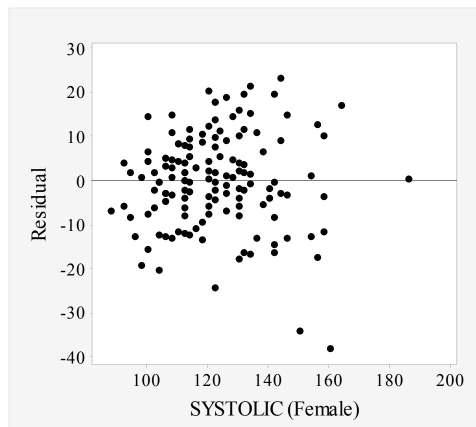
29. $\hat{y} = 138 - 0.223x$; $r = -0.162$; P-value $= 0.050$; With a significant linear correlation, the best predicted value is $\hat{y} = 138 - 0.223(80) = 120$ mm Hg.



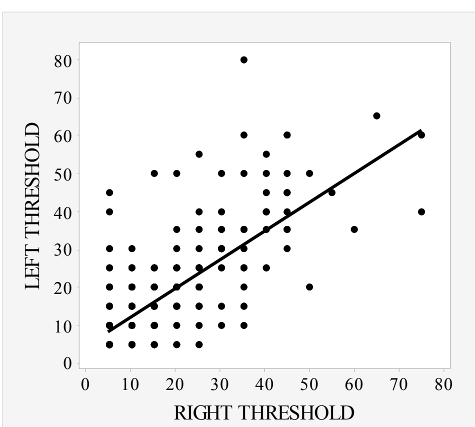


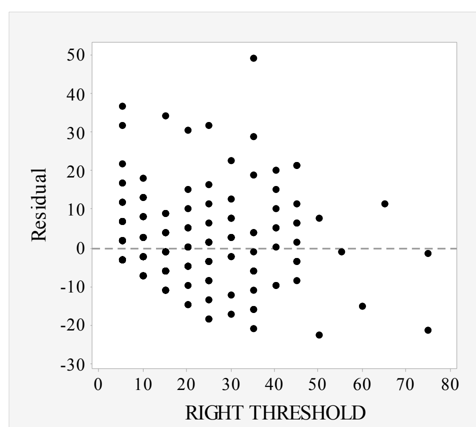
30. $\hat{y} = 44.7 + 0.209x$; $r = 0.321$; P-value $= 0.000$; With a significant linear correlation, the best predicted value is $\hat{y} = 44.7 + 0.209(120) = 69.8$ mm Hg.



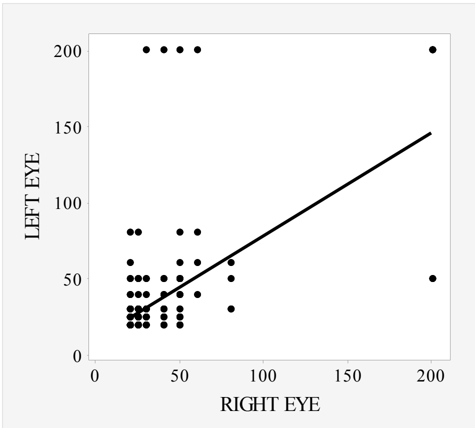


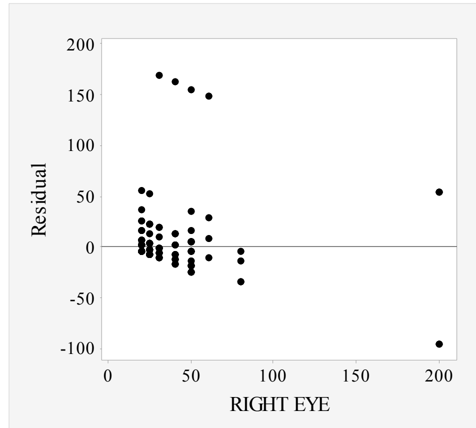
31. $\hat{y} = 4.67 + 0.758x$; $r = 0.748$; P-value $= 0.000$; With a significant linear correlation, the best predicted value is $\hat{y} = 4.67 + 0.758(20) = 19.8$.





32. $\hat{y}=10.7+0.678x$; $r=0.595$; P-value $=0.000$; With a significant linear correlation, the best predicted value is $\hat{y}=10.7+0.678(50)=44.6$.





33. a. Using $\hat{y}=3.02+.0602x$, the sum of squares of the residuals is 12.28.

b. Using $\hat{y}=-3+0.0500x$, the sum of squares of the residuals is 15.71, which is larger than 12.28, which is the sum of squares of the residuals for the regression line.

x	$\hat{y}=3.02+0.0602x$	$(\hat{y}-\overline{y})^2$
56.0	6.3912	0.259
82.0	7.9564	0.021
78.0	7.7156	1.731
86.0	8.1972	3.599
88.0	8.3176	6.669
	Sum:	12.279

x	$\hat{y}=3+0.0500x$	$(\hat{y}-\overline{y})^2$
56.0	5.8	1.21
82.0	7.1	1
78.0	6.9	0.25
86.0	7.3	1
88.0	7.4	12.25
	Sum:	15.71

Section 10-3: Prediction Intervals and Variation

1. The value of $s_e=16.27555$ cm is the standard error of estimate, which is a measure of the differences between the observed weights and the weights predicted from the regression equation. It is a measure of the variation of the sample points about the regression line.

2. We have 95% confidence that the limits of 59.0 kg and 123.6 kg contain the value of the weight for a male with a height of 180 cm. The major advantage of using a prediction interval is that it provides us with a range of likely weights, so we have a sense of how accurate the predicted weight is likely to be. The terminology of *prediction interval* is used for an interval estimate of a variable, whereas the terminology of *confidence interval* is used for an interval estimate of a population parameter.

3. The coefficient of determination is $r^2=0.155$. We know that 15.5% of the variation in weight is explained by the linear correlation between height and weight, and 84.5% of the variation in weight is explained by other factors and/or random variation.

4. For the paired weights, $s_e=0$ because there is an exact conversion formula. For a student who weighs 100 lb, the predicted weight is 45.4 kg, and there is no prediction interval because the conversion yields an exact result.

5. $r^2=(0.874)^2=0.764$; 76.4% of the variation in temperature is explained by the linear correlation between chirps and temperature, and 23.6% of the variation in temperature is explained by other factors and/or random variation.

6. $r^2 = (0.885)^2 = 0.783$; 78.3% of the variation in waist size is explained by the linear correlation between weight and waist size, and 21.7% of the variation in waist size is explained by other factors and/or random variation.

7. $r^2 = (0.934)^2 = 0.872$; 87.2% of the variation in weights of bears is explained by the linear correlation between neck size and weight, and 12.8% of the variation in weights is explained by other factors and/or random variation.

8. $r^2 = (0.783)^2 = 0.613$; 61.3% of the variation in weight is explained by the linear correlation between head width and weight, and 38.7% of the variation in weight is explained by other factors and/or random variation.

9. $r = 0.850$; Critical values $(\alpha = 0.05)$: $r = \pm 0.404$ (Table: $r \approx \pm 0.396$); Yes, there is sufficient evidence to support a claim of a linear correlation between registered boats and manatee fatalities.

10. $r^2 = (0.850)^2 = 0.723$, or 72.3%

11. The best predicted value is 70.5 manatees.

12. The 95% prediction interval estimate is 50.0 manatees $< y <$ 90.9 manatees. We have 95% confidence that the limits of 50.0 and 90.9 contain the number of manatee fatalities in a year with 850,000 registered boats.

13. 99% CI: 42.7 manatees $< y <$ 98.3 manatees

$\hat{y} = -49.049 + 1.406(85) = 70.46$

$$E = t_{\alpha/2} s_e \sqrt{1 + \frac{1}{n} + \frac{n(x_0 - \bar{x})^2}{n(\Sigma x^2) - (\Sigma x)^2}} = 2.819(9.6605)\sqrt{1 + \frac{1}{24} + \frac{24(85 - 85.25)^2}{24(127822) - (2046)^2}} = 27.79$$

14. 95% CI: 67.7 manatees $< y <$ 109.8 manatees

$\hat{y} = -49.049 + 1.406(98) = 88.74$

$$E = t_{\alpha/2} s_e \sqrt{1 + \frac{1}{n} + \frac{n(x_0 - \bar{x})^2}{n(\Sigma x^2) - (\Sigma x)^2}} = 2.074(9.6605)\sqrt{1 + \frac{1}{24} + \frac{24(98 - 85.25)^2}{24(127822) - (2046)^2}} = 20.41$$

15. 95% CI: 65.1 manatees $< y <$ 106.8 manatees

$\hat{y} = -49.049 + 1.406(96) = 85.93$

$$E = t_{\alpha/2} s_e \sqrt{1 + \frac{1}{n} + \frac{n(x_0 - \bar{x})^2}{n(\Sigma x^2) - (\Sigma x)^2}} = 2.074(9.6605)\sqrt{1 + \frac{1}{24} + \frac{24(96 - 85.25)^2}{24(127822) - (2046)^2}} = 20.42$$

16. 95% CI: 45.5 manatees $< y <$ 101.1 manatees

$\hat{y} = -49.049 + 1.406(87) = 73.27$

$$E = t_{\alpha/2} s_e \sqrt{1 + \frac{1}{n} + \frac{n(x_0 - \bar{x})^2}{n(\Sigma x^2) - (\Sigma x)^2}} = 2.819(9.6605)\sqrt{1 + \frac{1}{24} + \frac{24(87 - 85.25)^2}{24(127822) - (2046)^2}} = 27.79$$

17. a. 618.9541
    b. 304.7126
    c. 95% CI: 56.0 mm Hg $< y <$ 81.6 mm Hg
    $\hat{y} = 23.6 + 0.377(120) = 68.8$
    $$E = t_{\alpha/2} s_e \sqrt{1 + \frac{1}{n} + \frac{n(x_0 - \bar{x})^2}{n(\Sigma x^2) - (\Sigma x)^2}} = 2.2283(5.52008)\sqrt{1 + \frac{1}{12} + \frac{12(120 - 121)^2}{12(180{,}048) - (1452)^2}} = 12.8$$

18. a. 3714.141
    b. 1659.426
    c. 99% CI: 0.607 ft $< y <$ 85.8 ft; The interval is wide enough to be of no practical use.
    $\hat{y} = 21.6 + 5.401(4.0) = 43.2$
    $$E = t_{\alpha/2} s_e \sqrt{1 + \frac{1}{n} + \frac{n(x_0 - \bar{x})^2}{n(\Sigma x^2) - (\Sigma x)^2}} = 3.1693(12.88187)\sqrt{1 + \frac{1}{12} + \frac{12(4.0 - 4.91)^2}{12(416.41) - (58.9)^2}} = 42.6$$

19. a. 352.7278
    b. 109.3722
    c. 90% CI: 71.09°F $< y <$ 88.71°F
    $\hat{y} = 27.628 + 0.05227(1000) = 79.90$
    $$E = t_{\alpha/2} s_e \sqrt{1 + \frac{1}{n} + \frac{n(x_0 - \bar{x})^2}{n(\Sigma x^2) - (\Sigma x)^2}} = 1.943(4.2695)\sqrt{1 + \frac{1}{8} + \frac{8(1000 - 1016.25)^2}{8(8391204) - (8130)^2}} = 8.81$$

20. a. 8880.12
    b. 991.1515
    c. 99% CI: 125.0 kg $< y <$ 284.5 kg
    $\hat{y} = -156.879 + 40.182(9.0) = 204.76$
    $$E = t_{\alpha/2} s_e \sqrt{1 + \frac{1}{n} + \frac{n(x_0 - \bar{x})^2}{n(\Sigma x^2) - (\Sigma x)^2}} = 4.604(15.7413)\sqrt{1 + \frac{1}{6} + \frac{6(9.0 - 8.5)^2}{6(439) - (51)^2}} = 79.79$$

21. a. 95% CI: 6.5 $(1000\text{cells}/\mu\text{L}) < \bar{y} < 7.2$ $(1000\text{cells}/\mu\text{L})$
    $\hat{y} = 4.06 + 0.0345(80) = 6.82$
    $$E = t_{\alpha/2} s_e \sqrt{\frac{1}{n} + \frac{n(x_0 - \bar{x})^2}{n(\Sigma x^2) - (\Sigma x)^2}} = 1.976(1.92241)\sqrt{\frac{1}{147} + \frac{147(80.0 - 74.0)^2}{147(828{,}832) - (10{,}884)^2}} = 0.341$$
    b. 95% CI: 6.2 $(1000\text{cells}/\mu\text{L}) < \bar{y} < 6.8$ $(1000\text{cells}/\mu\text{L})$
    $\hat{y} = 4.06 + 0.0345(70) = 6.48$
    $$E = t_{\alpha/2} s_e \sqrt{\frac{1}{n} + \frac{n(x_0 - \bar{x})^2}{n(\Sigma x^2) - (\Sigma x)^2}} = 1.976(1.92241)\sqrt{\frac{1}{147} + \frac{147(70.0 - 74.0)^2}{147(828{,}832) - (10{,}884)^2}} = 0.326$$

Section 10-4: Multiple Regression

1. The response variable is weight and the predictor variables are length and chest size.
2. No, it is not better to use the regression equation with the three predictor variables of length, chest size, and neck size. The adjusted R^2 value of 0.925 is just a little less than 0.933, so in this case it is better to use two predictor variables instead of three.
3. The unadjusted R^2 increases (or remains the same) as more variables are included, but the adjusted R^2 is adjusted for the number of variables and sample size. The unadjusted R^2 incorrectly suggests that the best multiple regression equation is obtained by including all of the available variables, but by taking into account the sample size and number of predictor variables, the adjusted R^2 is much more helpful in weeding out variables that should not be included.
4. 92.8% of the variation in weights of bears can be explained by the variables of length and chest size, so 7.2% of the variation in weights can be explained by other factors and/or random variation.
5. $\text{Son} = 18.0 + 0.504\,\text{Father} + 0.277\,\text{Mother}$
6. a. $P\text{-value} < 0.0001$

   b. $R^2 = 0.3649$

   c. adjusted $R^2 = 0.3552$
7. A P-value less than 0.0001 is low, but the values of R^2 (0.3649) and adjusted R^2 (0.3552) are not high. Although the multiple regression equation fits the sample data best, it is not a good fit, so it should not be used for predicting the height of a son based on the height of his father and the height of his mother.
8. Predicted height: $\text{Son} = 18.0 + 0.504(70) + 0.277(60) = 70$ in.; This result is not likely to be a good predicted value because the multiple regression equation is not a good model (based on the results from Exercise 7).
9. WAIST (waist circumference) because it has the highest adjusted R^2 (0.782) and all the P-values are the same (0.000).
10. WAIST (waist circumference) and ARM (arm circumference) because they have the best combination of small P-value (0.000) and highest adjusted R^2 (0.878), but WAIST (waist circumference) and HT (height) are almost the same.
11. Some subjective judgment is involved here. A good choice would be to use all three variables because the adjusted R^2 value of 0.939 is considerably higher than any of the other values of adjusted R^2. With this reasoning, the best regression equation is $\hat{y} = -147 + 0.632\,\text{HT} + 0.697\,\text{WAIST} + 1.58\,\text{ARM}$.
12. Predicted city fuel consumption is $\hat{y} = -147 + 0.632(186) + 0.697(107.8) + 1.58(37.0) = 104.1$ kg. (Based on the result from Exercise 11.) The predicted value is a good estimate, given the adjusted R^2 value of 0.939, the P-value of 0.000, and the large sample size.
13. The best regression equation is $\hat{y} = 0.127 + 0.0878x_1 - 0.0250x_2$, where x_1 represents tar and x_2 represents carbon monoxide. It is best because it has the highest adjusted R^2 value of 0.927 and the lowest P-value of 0.000. It is a good regression equation for predicting nicotine content because it has a high value of adjusted R^2 and a low P-value. Possible models:

    $\hat{y} = 0.080 + 0.0633x_1$, adjusted $R^2 = 0.877$

    $\hat{y} = 0.328 + 0.0397x_2$, adjusted $R^2 = 0.437$

    $\hat{y} = 0.127 + 0.0878x_1 - 0.0250x_2$, adjusted $R^2 = 0.927$

14. The best regression equation is $\hat{y} = 0.251 + 0.101x_1 - 0.0454x_2$, where x_1 represents tar and x_2 represents carbon monoxide. It is best because it has the highest adjusted R^2 value of 0.908 and the lowest P-value of 0.000. It is a good regression equation for predicting nicotine content because it has a high value of adjusted R^2 and a low P-value. Possible models:

    $\hat{y} = 0.139 + 0.0567x_1$, adjusted $R^2 = 0.752$

    $\hat{y} = 0.385 + 0.0325x_2$, adjusted $R^2 = 0.283$

    $\hat{y} = 0.251 + 0.101x_1 - 0.0454x_2$, adjusted $R^2 = 0.908$

15. The best regression equation is $\hat{y} = 109 - 0.00670x_1$, where x_1 represents volume. It is best because it has the highest adjusted R^2 value of 0.0513 and the lowest P-value of 0.791. The three regression equations all have adjusted values of R^2 that are very close to 0, so none of them are good for predicting IQ. It does not appear that people with larger brains have higher IQ scores. Possible models:

    $\hat{y} = 109 - 0.00670x_1$, adjusted $R^2 = 0.0513$

    $\hat{y} = 101 - 0.00178x_2$, adjusted $R^2 = 0.0555$

    $\hat{y} = 108 - 0.00694x_1 + 0.00722x_2$, adjusted $R^2 = 0.113$

16. The best regression equation is $\hat{y} = -10.0 + 0.567x_1 + 0.532x_2$, where x_1 represents verbal IQ score and x_2 represents performance IQ score. It is best because it has the highest adjusted R^2 value of 0.999 and the lowest P-value of 0.000. Because the adjusted R^2 is so close to 1, it is likely that predicted values will be very accurate. Possible models:

    $\hat{y} = 11.5 + 0.940x_1$, adjusted $R^2 = 0.762$

    $\hat{y} = 10.5 + 0.806x_2$, adjusted $R^2 = 0.814$

    $\hat{y} = -10.0 + 0.567x_1 + 0.532x_2$, adjusted $R^2 = 0.999$

17. For H_0: $\beta_1 = 0$, Test statistic: $t = \dfrac{0.769317 - 0}{0.0711414} = 10.813917$; $P\text{-value} < 0.0001$; Reject H_0 and conclude that the regression coefficient of $b_1 = 0.769$ should be kept.

    For H_0: $\beta_2 = 0$, Test statistic: $t = \dfrac{1.009510 - 0}{0.0338123} = 29.856$; $P\text{-value} < 0.0001$; Reject H_0 and conclude that the regression coefficient of $b_2 = 1.01$ should be kept.

    It appears that the regression equation should include both independent variables of height and waist circumference.

18. $0.629 < \beta_1 < 0.910$; $0.943 < \beta_2 < 1.08$; Neither confidence interval includes 0, so neither of the two variables should be eliminated from the regression equation.

$$\text{CI for } \beta_1$$
$$b_1 - E < \beta_1 < b_1 + E$$
$$b_1 - t_{\alpha/2}s_{b_1} < \beta_1 < b_1 + t_{\alpha/2}s_{b_1}$$
$$0.7693 - 1.976(0.0711414) < \beta_1 < 0.7693 + 1.976(0.0711414)$$
$$0.629 < \beta_1 < 0.910$$

18. (continued)

$$\text{CI for } \beta_2$$
$$b_2 - E < \beta_2 < b_2 + E$$
$$b_2 - t_{\alpha/2} s_{b_2} < \beta_2 < b_2 + t_{\alpha/2} s_{b_2}$$
$$1.0095 - 1.976(0.033812) < \beta_2 < 1.0095 + 1.976(0.033812)$$
$$0.943 < \beta_2 < 1.08$$

Section 10-5: Dummy Variables and Logistic Regression

1. A dummy variable is a variable having only two values and those values (such as 0 and 1) are used to represent the different categories of a qualitative variable.

2. Linear regression is used to predict the value of a continuous variable while logistic regression is used to predict the probability of an outcome in a binomial event.

3. With simple logistic regression, there is only one predictor variable, but with multiple logistic regression, there are two or more predictor variables.

4. False, a dummy variable that is a predictor variable does not necessitate the use of logistic regression.

5. Length and weight are predictor variables. The response variable is sex, which is a dummy variable.

6. Because the regression equation has the high overall P-value of 0.218, the predicted value is not likely to be very accurate.

7. The probability that the bear is a male is $\dfrac{e^{2.40-0.0553(60)+0.00826(300)}}{1+e^{2.40-0.0553(60)+0.00826(300)}} = 0.826$. The probability that the bear is a female is $1-0.826 = 0.174$.

8. The probability that the bear is a male is $\dfrac{e^{2.40-0.0553(40)+0.00826(50)}}{1+e^{2.40-0.0553(40)+0.00826(50)}} = 0.646$. The probability that the bear is a female is $1-0.646 = 0.354$. The bear is most likely a cub or juvenile, given its size.

9. $\hat{y} = 3.06 + 82.4(\text{SEX}) + 2.91(\text{AGE})$

   a. Female: $\hat{y} = 3.06 + 82.4(0) + 2.91(20) = 61$ lb

   b. Male: $\hat{y} = 3.06 + 82.4(1) + 2.91(20) = 144$ lb

   The sex of the bear does appear to have an effect on its weight. The regression equation indicates that the predicted weight of a male bear is about 82 lb more than the predicted weight of a female bear with other characteristics being the same.

10. $\hat{y} = -56.2 + 0.828(\text{HEIGHT}) - 2.37(\text{SEX})$

   a. Female: $\hat{y} = -56.2 + 0.828(170) - 2.37(0) = 84.6$ kg

   b. Male: $\hat{y} = -56.2 + 0.828(170) - 2.37(1) = 82.2$ kg

   One would expect a man to weigh more than a woman of the same height, so the results do not make particular sense. Also, the weights seem rather high for a person standing approximately 5 foot 6 inches.

11. $\ln\left(\dfrac{p}{1-p}\right) = -41.2 + 0.250(\text{Height}) - 0.00856(\text{Weight})$

   The probability of a male is $\dfrac{e^{-41.2+0.250(170)-0.00856(90)}}{1+e^{-41.2+0.250(170)-0.00856(90)}} = 0.629$.

12. $\ln\left(\frac{p}{1-p}\right) = -38.3 + 0.245(\text{Height}) - 0.0289(\text{Pulse})$; The probability of a male is $\frac{e^{-38.3+0.245(170)-0.0289(60)}}{1+e^{-38.3+0.245(170)-0.0289(60)}} = 0.834$. The negative coefficient for PULSE shows that higher pulse rates make males less likely

13. $\ln\left(\frac{p}{1-p}\right) = -101.5 + 0.191(\text{Foot Length}) + 0.301(\text{Height})$

    The probability of a male is $\frac{e^{-101.5+0.191(28)+0.301(190)}}{1+e^{-101.5+0.191(28)+0.301(190)}} = 0.9999$.

14. $\ln\left(\frac{p}{1-p}\right) = -70.8 + 0.993(\text{Shoe Size}) + 0.350(\text{Height})$; The probability of a male is $\frac{e^{-70.8+0.993(9)+0.350(170)}}{1+e^{-70.8+0.993(9)+0.350(170)}} = 0.086$. The P-value for shoe size in this model is 0.245, whereas the P-value for foot length in Exercise 13 is 0.013, and the model in Exercise 13 has a higher adjusted R^2 value. The predictions using this model would be less accurate.

Chapter Quick Quiz

1. Conclude that there is not sufficient evidence to support the claim of a linear correlation between the systolic blood pressure measurements of the right and left arms
2. None of the given values change when the variables are switched.
3. The value of r does not change if all values of one of the variables are multiplied by the same constant.
4. Because r must be between –1 and 1 inclusive, the value of 1.500 is the result of an error in the calculations.
5. The best predicted number of burglaries is 163.2 mm Hg, which is the mean of the five measurements for the left arm.
6. The best predicted number of burglaries would be $\hat{y} = 43.6 + 1.31(100) = 174.6$ mm Hg.
7. $r^2 = (0.867)^2 = 0.752$
8. false
9. false
10. $r = -1$

Review Exercises

1. a. $r = 0.962$; P-value $= 0.000$ (Table: P-value < 0.01); Critical values ($\alpha = 0.05$): $r = \pm 0.707$; There is sufficient evidence to support the claim that there is a linear correlation between the amount of tar and the amount of nicotine.

   b. $(0.962)^2 = 0.925$, or 92.5%

   c. $\hat{y} = -0.758 + 0.0920x$

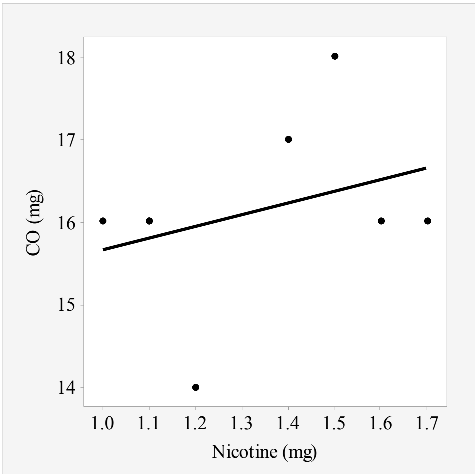
   d. The predicted value is $\hat{y} = -0.758 + 0.0920(23) = 1.358$ mg or 1.4 mg rounded, which is close to the actual amount of 1.3 mg.

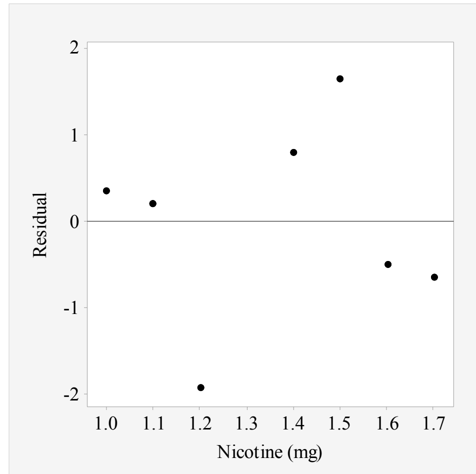
2. a. The scatterplot (see part c) shows a pattern with nicotine and carbon monoxide both increasing from left to right, but it is a very weak pattern and the points are not very close to a straight-line pattern, so it appears that there is not sufficient sample evidence to support the claim of a linear correlation between amounts of nicotine and carbon monoxide.

2. (continued)

b. H_0: $\rho = 0$; H_1: $\rho \neq 0$; $r = 0.329$; P-value $= 0.427$ (Table: P-value > 0.05); Critical values $(\alpha = 0.05)$: $r = \pm 0.707$; Fail to reject H_0. There is not sufficient evidence to support the claim that there is a linear correlation between amount of nicotine and amount of carbon monoxide.

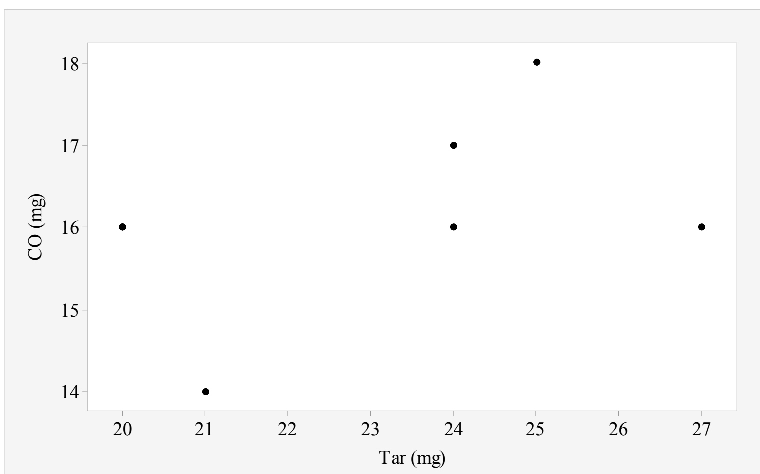
c. $\hat{y} = 14.2 + 1.42x$





d. The predicted value is $\bar{y} = 16.1$ mg, which is close to the actual amount of 15 mg.

3. a. The scatterplot shows a pattern with amounts of tar and carbon monoxide both increasing from left to right, but it is a very weak pattern and the points are not very close to a straight-line pattern, so it appears that there is not sufficient sample evidence to support the claim of a linear correlation between amounts of tar and carbon monoxide.



b. Three of the original pairs of sample data are the same $(20, 16)$, so those three points are at the same location and they appear to be one point.

c. $r = 0.437$; P-value $= 0.279$ (Table: P-value < 0.05); Critical values $(\alpha = 0.05)$: $r = \pm 0.707$; There is not sufficient evidence to support the claim that there is a linear correlation between amounts of tar and carbon monoxide.

d. $\hat{y} = 12.0 + 0.181x$

e. The predicted value is $\bar{y} = 16.1$ mg, which is close to the actual amount of 15 mg.

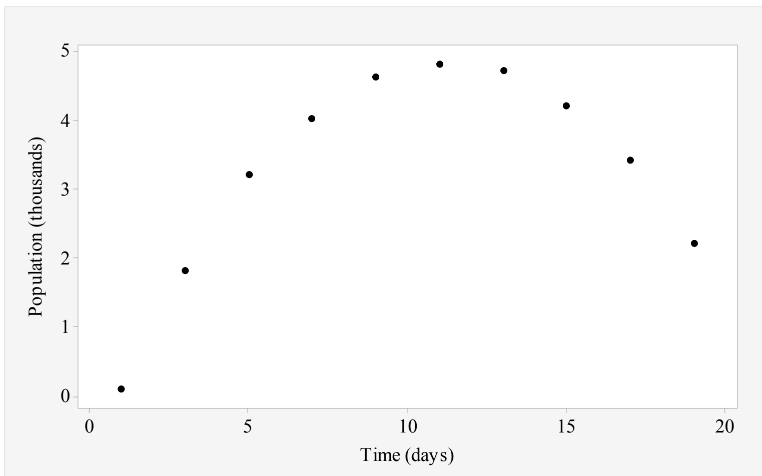
4. a. NICOTINE $= -0.443 + 0.0968\,\text{TAR} - 0.0262\,\text{CO}$, or $\hat{y} = -0.443 + 0.0968x_1 - 0.0262x_2$, where x_1 represents tar and x_2 represents carbon monoxide.

   b. $R^2 = 0.936$; adjusted $R^2 = 0.910$; P-value $= 0.001$

   c. With high values of R^2 and adjusted R^2 and a small P-value of 0.001, it appears that the regression equation can be used to predict the amount of nicotine given the amounts of tar and carbon monoxide.

   d. The predicted value is $\hat{y} = -0.443 + 0.0968(23) - 0.0262(15) = 1.39$ mg or 1.4 mg rounded, which is close to the actual value of 1.3 mg of nicotine.

5. H_0: $\rho = 0$; H_1: $\rho \neq 0$; $r = 0.450$; P-value $= 0.192$ (Table: P-value < 0.05); Critical values ($\alpha = 0.05$): $r = \pm 0.632$; Fail to reject H_0. There is not sufficient evidence to support the claim that there is a linear correlation between time and population size. Although there is no *linear* correlation between time and population size, the scatterplot shows a very distinct pattern revealing that time and population size are associated by a function that is not linear.



6. $\ln\left(\dfrac{p}{1-p}\right) = -44.4 + 0.557(8\text{ AM Temp}) - 0.098(12\text{ AM Temp})$; The probability that the subject smokes is $\dfrac{e^{-44.4+0.557(98.25)-0.098(98.25)}}{1+e^{-44.4+0.557(98.25)-0.098(98.25)}} = 0.667$. Because the regression equation has the high overall P-value of 0.521, the predicted value is not likely to be very accurate.

Cumulative Review Exercises

1. $\bar{x} = \dfrac{4+14+3+1+4+7+3+4}{8} = 3.3$ lb

$$s = \sqrt{\frac{(4-3.3)^2 + (14-3.3)^2 + (3-3.3)^2 + (1-3.3)^2 + (4-3.3)^2 + (7-3.3)^2 + (3-3.3)^2 + (4-3.3)^2}{8-1}} = 5.7 \text{ lb}$$

Before	183	212	177	209	155	162	167	170
After	179	198	180	208	159	155	164	166
Difference	4	14	3	1	4	7	3	4

2. The highest weight before the diet is 212 lb, which converts to $z = \dfrac{212-179.4}{21.0} = 1.55$. The highest weight is not significantly high because its z score of 1.55 shows that it is within 2 standard deviations of the mean.

3. H_0: $\mu_d = 0$; H_1: $\mu_d > 0$; difference = Before − After;
Test statistic: $t = 1.613$; P-value = 0.0754 (Table: P-value > 0.05); Critical value: $t = 1.895$; Fail to reject H_0. There is not sufficient evidence to support the claim that the diet is effective.

$$t = \frac{\bar{d} - \mu_d}{s_d / \sqrt{n}} = \frac{3.3 - 0}{5.7/\sqrt{8}} = 1.613 \text{ (df = 7)}$$

4. 95% CI: $\bar{x} \pm t_{\alpha/2} \frac{s}{\sqrt{n}} = 179.4 \pm 2.365 \cdot \frac{21.0}{\sqrt{8}} \Rightarrow 161.8 \text{ lb} < \mu < 197.0 \text{ lb}$; We have 95% confidence that the interval limits of 161.8 lb and 197.0 lb contain the true value of the mean of the population of all subjects before the diet.

5. a. H_0: $\rho = 0$; H_1: $\rho \neq 0$; $r = 0.965$; P-value = 0.0001; Critical values $(\alpha = 0.05)$: $r = \pm 0.707$; There is sufficient evidence to support the claim that there is a linear correlation between before and after weights.

   b. $r = 1$

   c. $r = 1$

   d. The effectiveness of the diet is determined by the amounts of weight lost, but the linear correlation coefficient is not sensitive to different amounts of weight loss. Correlation is not a suitable tool for testing the effectiveness of the diet.

6. a. $z_{x=3500} = \frac{3500 - 3420}{495} = 0.16$; which has an area of $1 - 0.5636 = 0.4364$, or 43.64% to the right (Tech: 43.58%).

   b. The z score for 10% to the left is –1.28, which corresponds to a weight of $-1.28 \cdot 495 + 3420 = 2786.4$ g (Tech: 2785.5 g).

   c. $z_{x=2450} = \frac{2450 - 3420}{495} = -1.96$ and $z_{x=4390} = \frac{4390 - 3420}{495} = 1.96$; which has an area of 0.9550 between them, so $1 - 0.9500 = 0.0500$, or 5.00% require special treatment. It is the case that many of the babies do require special treatment.

7. a. No, correlation can be used to investigate an association between the two variables, not whether differences between values of the two variables are significant.

   b. H_0: $\mu_d = 0$; H_1: $\mu_d \neq 0$; difference = Sitting − Supine;
   Test statistic: $t = 1.185$; P-value = 0.2663 (Table: P-value > 0.20); Critical values: $t = \pm 2.262$; Fail to reject H_0. There is not sufficient evidence to warrant rejection of the claim that position has no effect. Position does not appear to have a significant effect.

$$t = \frac{\bar{d} - \mu_d}{s_d / \sqrt{n}} = \frac{0.140 - 0}{0.380/\sqrt{10}} = 1.185 \text{ (df = 9)}$$

Subject	A	B	C	D	E	F	G	H	I	J
Sitting	4.66	5.70	5.37	3.34	3.77	7.43	4.15	6.21	5.90	5.77
Supine	4.63	6.34	5.72	3.23	3.60	6.96	3.66	5.81	5.61	5.33
Difference	0.03	0.64	0.35	0.11	0.17	0.47	0.49	0.40	0.29	0.44

8. There must be an error, because the rates of 13.7% and 10.6% are not possible with sample sizes of 100.