Simple Linear Regression

Chapter 11

11.2 In a probabilistic model, the dependent variable is the variable that is to be modeled or predicted while the independent variable is the variable used to predict the dependent variable.

11.4 No. The random error component, ε, allows the values of the variable to fall above or below the line.

11.6 For all problems below, we use:

$$\text{Slope} = \frac{\text{"rise"}}{\text{"run"}} = \frac{y_2 - y_1}{x_2 - x_1} \qquad \text{If } y = \beta_0 + \beta_1 x\text{, then } \beta_0 = y - \beta_1 x\,.$$

a. $\text{Slope} = \dfrac{5-1}{5-1} = 1 = \beta_1$

Since a given point is (1, 1) and $\beta_1 = 1$, the y-intercept $= \beta_0 = 1 - 1(1) = 0$.

b. $\text{Slope} = \dfrac{0-3}{3-0} = -1 = \beta_1$

Since a given point is (0, 3) and $\beta_1 = -1$, the y-intercept is $\beta_0 = 3 - (-1)(0) = 3$.

c. $\text{Slope} = \dfrac{2-1}{4-(-1)} = \dfrac{1}{5} = .2 = \beta_1$

Since a given point is (–1, 1) and $\beta_1 = .2$, the y-intercept is $\beta_0 = 1 - .2(-1) = 1.2$.

d. $\text{Slope} = \dfrac{6-(-3)}{2-(-6)} = \dfrac{9}{8} = 1.125 = \beta_1$

Since a given point is (–6, –3) and $\beta_1 = 1.125$, the y-intercept is $\beta_0 = -3 - 1.125(-6) = 3.75$.

11.8 a. The equation for a straight line (deterministic) is $y = \beta_0 + \beta_1 x$.

If the line passes through (1, 1), then $1 = \beta_0 + \beta_1(1) = \beta_0 + \beta_1$

Likewise, through (5, 5), then $5 = \beta_0 + \beta_1(5)$

Solving for these two equations:

$$\begin{aligned} 1 &= \beta_0 + \beta_1 \\ -\big(5 &= \beta_0 + \beta_1(5)\big) \\ -4 &= \ -4\beta_1 \Rightarrow \beta_1 = 1 \end{aligned}$$

Substituting $\beta_1 = 1$ into the first equation, we get $1 = \beta_0 + 1 \Rightarrow \beta_0 = 0$

The equation is $y = 0 + 1x$ or $y = x$.

b. The equation for a straight line is $y = \beta_0 + \beta_1 x$. If the line passes through (0, 3), then $3 = \beta_0 + \beta_1(0)$, which implies $\beta_0 = 3$. Likewise, through the point (3, 0), then $0 = \beta_0 + 3\beta_1$ or $-\beta_0 = 3\beta_1$. Substituting $\beta_0 = 3$, we get $-3 = 3\beta_1$ or $\beta_1 = -1$. Therefore, the line passing through (0, 3) and (3, 0) is $y = 3 - x$.

c. The equation for a straight line is $y = \beta_0 + \beta_1 x$. If the line passes through (−1, 1), then $1 = \beta_0 + \beta_1(-1)$. Likewise through the point (4, 2), $2 = \beta_0 + \beta_1(4)$. Solving for these two equations:

$$\begin{aligned} 2 &= \beta_0 + \beta_1(4) \\ -\big(1 &= \beta_0 + \beta_1(-1)\big) \\ 1 &= \quad 5\beta_1 \Rightarrow \beta_1 = \frac{1}{5} = .2 \end{aligned}$$

Solving for β_0, $1 = \beta_0 + \frac{1}{5}(-1) \Rightarrow 1 = \beta_0 - \frac{1}{5} \Rightarrow \beta_0 = 1 + \frac{1}{5} = \frac{6}{5} = 1.2$

The equation, with $\beta_0 = 1.2$ and $\beta_1 = .2$, is $y = 1.2 + .2x$.

d. The equation for a straight line is $y = \beta_0 + \beta_1 x$. If the line passes through (−6, −3), then $-3 = \beta_0 + \beta_1(-6)$. Likewise, through the point (2, 6), $6 = \beta_0 + \beta_1(2)$. Solving these equations simultaneously.

$$\begin{aligned} 6 &= \beta_0 + \beta_1(2) \\ -\big(-3 &= \beta_0 + \beta_1(-6)\big) \\ 9 &= \quad 8\beta_1 \Rightarrow \beta_1 = \frac{9}{8} = 1.125 \end{aligned}$$

Solving for β_0, $6 = \beta_0 + 2(1.125) \Rightarrow 6 - 2.25 = \beta_0 \Rightarrow \beta_0 = 3.75$

Therefore, $y = 3.75 + 1.125x$.

11.10 a. $y = 4 + x$. The slope is $\beta_1 = 1$. The intercept is $\beta_0 = 4$.

b. $y = 5 - 2x$. The slope is $\beta_1 = -2$. The intercept is $\beta_0 = 5$.

c. $y = -4 + 3x$. The slope is $\beta_1 = 3$. The intercept is $\beta_0 = -4$.

d. $y = -2x$. The slope is $\beta_1 = -2$. The intercept is $\beta_0 = 0$.

e. $y = x$. The slope is $\beta_1 = 1$. The intercept is $\beta_0 = 0$.

f. $y = .5 + 1.5x$. The slope is $\beta_1 = 1.5$. The intercept is $\beta_0 = .5$.

11.12 Two properties of the line estimated using the method of least squares are:

1. the sum of the errors equals 0
2. the sum of squared errors (SSE) is smaller than for any other straight-line model

11.14 a.

x_i	y_i	x_i^2	x_iy_i
7	2	$7^2 = 49$	7(2) = 14
4	4	$4^2 = 16$	4(4) = 16
6	2	$6^2 = 36$	6(2) = 12
2	5	$2^2 = 4$	2(5) = 10
1	7	$1^2 = 1$	1(7) = 7
1	6	$1^2 = 1$	1(6) = 6
3	5	$3^2 = 9$	3(5) = 15

Totals:
$$\sum x_i = 7+4+6+2+1+1+3 = 24$$
$$\sum y_i = 2+4+2+5+7+6+5 = 31$$
$$\sum x_i^2 = 49+16+36+4+1+1+9 = 116$$
$$\sum x_i y_i = 14+16+12+10+7+6+15 = 80$$

b. $$\text{SS}_{xy} = \sum x_i y_i = \frac{\left(\sum x_i\right)\left(\sum y_i\right)}{n} = 80 - \frac{(24)(31)}{7} = 80 - 106.2857143 = -26.2857143$$

c. $$SS_{xx} = \sum x_i^2 - \frac{\left(\sum x_i\right)^2}{7} = 116 - \frac{(24)^2}{7} = 116 - 82.28571429 = 33.71428571$$

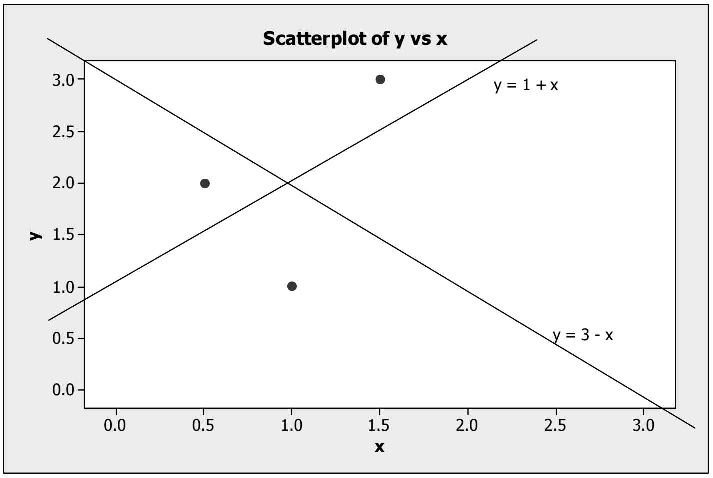
d. $$\hat{\beta}_1 = \frac{\text{SS}_{xy}}{\text{SS}_{xx}} = \frac{-26.2857143}{33.71428571} = -.779661017 \approx -.7797$$

e. $$\bar{x} = \frac{\sum x_i}{n} = \frac{24}{7} = 3.428571429 \qquad \bar{y} = \frac{\sum y_i}{n} = \frac{31}{7} = 4.428571429$$

f. $$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x} = 4.428571429 - (-.779661017)(3.428571429)$$
$$= 4.428571429 - (-2.673123487) = 7.101694916 \approx 7.102$$

g. The least squares line is $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x = 7.102 - .7797x$.

11.16 a.



b. Choose $y = 1 + x$ since it best describes the relation of x and y.

c.

y	x	$\hat{y} = 1 + x$	$y - \hat{y}$
2	.5	1 + .5 = 1.5	2 – 1.5 = .5
1	1.0	1 + 1 = 2.0	1 – 2.0 = –1.0
3	1.5	1 + 1.5 = 2.5	3 – 2.5 = .5
			Sum of errors = 0

y	x	$\hat{y} = 3 - x$	$y - \hat{y}$
2	.5	3 – .5 = 2.5	2 – 2.5 = –.5
1	1.0	3 – 1.0 = 2.0	1 – 2.0 = –1.0
3	1.5	3 – 1.5 = 1.5	3 – 1.5 = 1.5
			Sum of errors = 0

d. SSE = $\text{SSE} = \sum (y - \hat{y})^2$

SSE for 1st model: $y = 1 + x$, SSE = $(.5)^2 + (-1)^2 + (.5)^2 = 1.5$

SSE for 2nd model: $y = 3 - x$, SSE = $(-.5)^2 + (-1)^2 + (1.5)^2 = 3.5$

The best fitting straight line is the one that has the smallest sum of squares. The model $y = 1 + x$ has a smaller SSE, and therefore it verifies the visual check in part **a**.

e. Some preliminary calculations are:

$$\sum x_i = 3 \quad \sum y_i = 6 \quad \sum x_i y_i = 6.5 \quad \sum x_i^2 = 3.5$$

$$\text{SS}_{xy} = \sum x_i y_i - \frac{\left(\sum x_i\right)\left(\sum y_i\right)}{n} = 6.5 - \frac{(3)(6)}{3} = .5$$

$$\text{SS}_{xx} = \sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n} = 3.5 - \frac{(3)^2}{3} = .5$$

$$\hat{\beta}_1 = \frac{.5}{.5} = \frac{\text{SS}_{xy}}{\text{SS}_{xx}} = 1; \qquad \bar{x} = \frac{\sum x_i}{3} = \frac{3}{3} = 1; \qquad \bar{y} = \frac{\sum y_i}{3} = \frac{6}{3} = 2$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 2 - 1(1) = 1 \Rightarrow \hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x = 1 + x$$

The least squares line is the same as the first line given.

11.18 a. Yes. As the punishment use increases, the average payoff tends to decrease.

b. Negative.

c. Yes - the less the punishment use, the higher the average payoff.

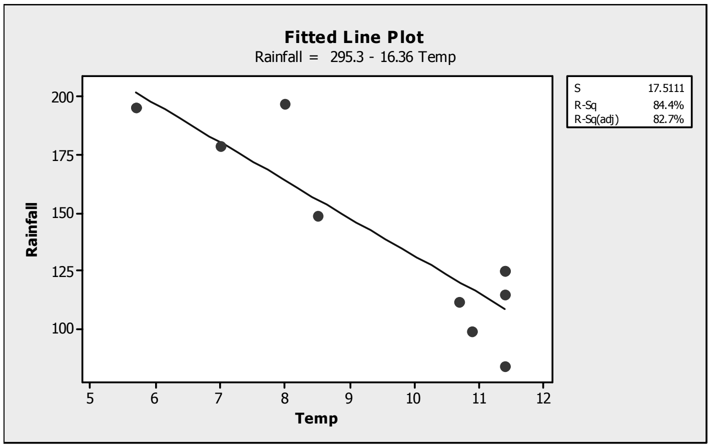
11.20 a. It appears that there is a positive linear trend. As the year of birth increases, the Z12-note entropy tends to increase.

b. The slope of the line is positive. As the year of birth increases, the Z12-note entropy tends to increase.

c. The line shown is the least squares line – it is the best line through the sample points. We do not know the values of β_0 and β_1 so we do not know the true line of means.

11.22 a. From the printout, the least squares prediction equation is $\hat{y} = 295.25 - 16.364x$.

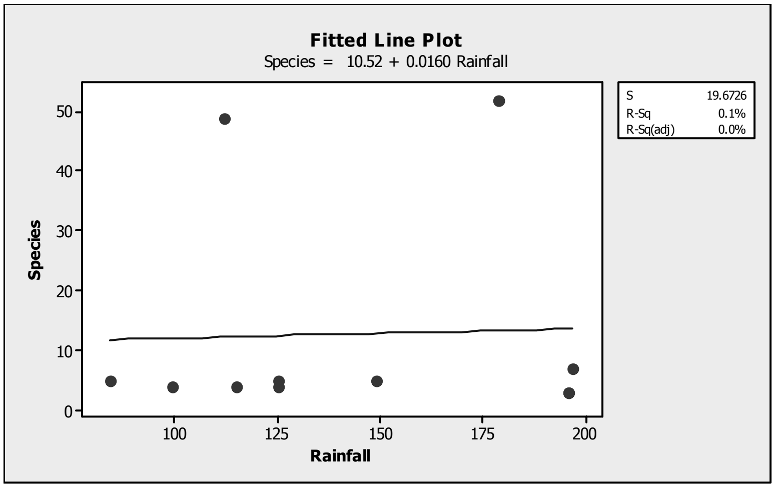
b. Using MINITAB, the scatterplot and the least square line are:



Since the data are fairly close the least squares prediction line, the line is a good predictor of annual rainfall.

c. From the printout, the least squares prediction equation is $\hat{y} = 10.52 + .016x$

Using MINITAB, the fitted regression plot and scatterplot are:



Since the data are not close to the least squares prediction line, the line is not a good predictor of ant species.

11.24 a. Some preliminary calculations are:

$$\sum x_i = 62 \qquad \sum y_i = 97.8 \qquad \sum x_i y_i = 1{,}087.78$$

$$\sum x_i^2 = 720.52 \qquad \sum y_i^2 = 1{,}710.2$$

$$SS_{xy} = \sum x_i y_i - \frac{\sum x_i \sum y_i}{n} = 1{,}087.78 - \frac{62(97.8)}{6} = 77.18$$

$$SS_{xx} = \sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n} = 720.52 - \frac{62^2}{6} = 79.8533333$$

$$\hat{\beta}_1 = \frac{SS_{xy}}{SS_{xx}} = \frac{77.18}{79.8533333} = .966521957$$

$$\hat{\beta}_o = \bar{y} - \hat{\beta}_1 \bar{x} = \frac{97.8}{6} - .966521957\left(\frac{62}{6}\right) = 6.312606442$$

The least squares prediction equation is $\hat{y} = 6.31 + .97x$

b. The y-intercept is 6.31. This value has no meaning because 0 is not in the observed range of the independent variable mean pore diameter.

c. The slope of the line is .97. For each unit increase in mean pore diameter, the mean porosity is estimated to increase by .97.

d. For x = 10, $\hat{y} = 6.31 + .97(10) = 16.01$.

11.26 a. The straight-line model is $y = \beta_0 + \beta_1 x + \varepsilon$

b. Some preliminary calculations are:

$\sum x_i$=1,292.7 $\sum y_i$=3,781.1 $\sum x_i y_i$= 218,291.63

$\sum x_i^2$=88,668.43 $\sum y_i^2$=651,612.45

$$\text{SS}_{xy} = \sum x_i y_i - \frac{\sum x_i \sum y_i}{n} = 218,291.63 - \frac{1,292.7(3,781.1)}{22} = -3,882.3686$$

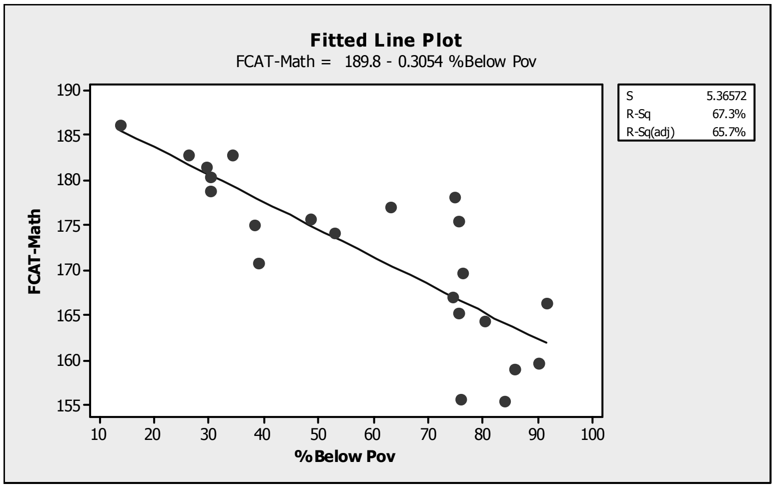
$$\text{SS}_{xx} = \sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n} = 88,668.43 - \frac{1,292.7^2}{22} = 12,710.55318$$

$$\hat{\beta}_1 = \frac{\text{SS}_{xy}}{\text{SS}_{xx}} = \frac{-3,882.3686}{12,710.55318} = -.305444503$$

$$\hat{\beta}_o = \overline{y} - \hat{\beta}_1 \overline{x} = \frac{3,781.1}{22} - (-.305444503)\left(\frac{1,292.7}{22}\right) = 189.815823$$

The least squares prediction equation is $\hat{y} = 189.816 - .305x$

c. Using MINITAB, the least squares line and the scatterplot are:



The relationship between the FCAT math scores and the percent of students below the poverty level appears to be negative. As the percent of students below the poverty line increases, the FCAT math score decreases. Since the data are fairly near the least squares line, it appears that the linear relationship is fairly strong.

d. $\hat{\beta}_0 = 189.816$. Since $x = 0$ is not in the observed range, $\hat{\beta}_0$ has no meaning other than the y-intercept.

$\hat{\beta}_1 = -.305$. For each unit increase in % below the poverty line, the mean FCAT-math score decreases by an estimated .305.

e. The straight-line model is $y = \beta_0 + \beta_1 x + \varepsilon$

Some preliminary calculations are:

$\sum x_i = 1,292.7$ $\sum y_i = 3,764.2$ $\sum x_i y_i = 217,738.81$

$\sum x_i^2 = 88,668.43$ $\sum y_i^2 = 645,221.16$

$$\text{SS}_{xy} = \sum x_i y_i - \frac{\sum x_i \sum y_i}{n} = 217,738.81 - \frac{1,292.7(3,764.2)}{22} = -3,442.16$$

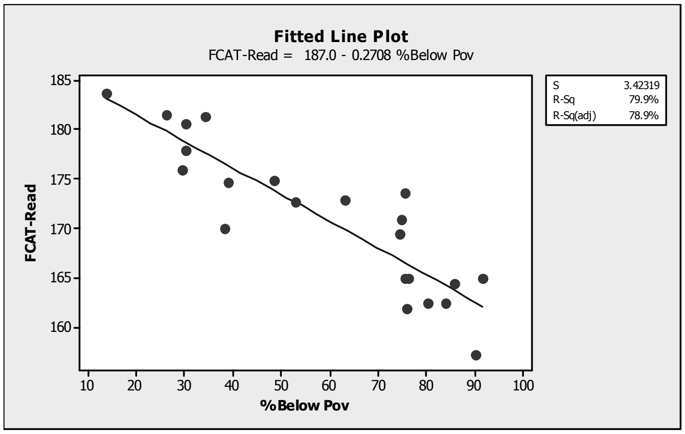
$$\text{SS}_{xx} = \sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n} = 88,668.43 - \frac{1,292.7^2}{22} = 12,710.55318$$

$$\hat{\beta}_1 = \frac{\text{SS}_{xy}}{\text{SS}_{xx}} = \frac{-3,442.16}{12,710.55318} = -.270811187$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = \frac{3,764.2}{22} - (-.270811187)\left(\frac{1,292.7}{22}\right) = 187.0126192$$

The least squares prediction equation is $\hat{y} = 187.013 - .271x$

Using MINITAB, the least squares line and the scatterplot are:



The relationship between the FCAT reading scores and the percent of students below the poverty level appears to be negative. As the percent of students below the poverty line increases, the FCAT read score decreases. Since the data are fairly near the least squares line, it appears that the linear relationship is fairly strong.

$\hat{\beta}_0 = 187.013$. Since $x = 0$ is not in the observed range, $\hat{\beta}_0$ has no meaning other than the y-intercept.

$\hat{\beta}_1 = -.271$. For each unit increase in % below the poverty line, the mean FCAT read score decreases by an estimated .271.

11.28 a. Some preliminary calculations are:

$$\sum x_i = 6167 \qquad \sum y_i = 135.8 \qquad \sum x_i^2 = 1{,}641{,}115 \qquad \sum x_i y_i = 34{,}764.5$$

$$\text{SS}_{xy} = \sum x_i y_i - \frac{\left(\sum x_i\right)\left(\sum y_i\right)}{n} = 34{,}764.5 - \frac{(6167)(135.8)}{24} = -130.441667$$

$$\text{SS}_{xx} = \sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n} = 1{,}641{,}115 - \frac{(6167)^2}{24} = 56{,}452.958$$

$$\hat{\beta}_1 = \frac{\text{SS}_{xy}}{\text{SS}_{xx}} = \frac{-130.441667}{56{,}452.958} = -.002310625 \approx -.0023$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x} = \frac{135.8}{24} - (-.002310625)\left(\frac{6167}{24}\right) = 6.2520679 \approx 6.25$$

The least squares line is $\hat{y} = 6.25 - .0023x$

b. $\hat{\beta}_0 = 6.25$. Since $x = 0$ is not in the observed range, $\hat{\beta}_0$ has no interpretation other than being the y-intercept.

$\hat{\beta}_1 = -.0023$. For each additional increase of 1 part per million of pectin, the mean sweetness index is estimated to decrease by .0023.

c. $\hat{y} = 6.25 - .0023(300) = 5.56$

11.30 Some preliminary calculations are:

$$\bar{y} = \frac{\sum x}{n} = \frac{103.07}{144} = .71576 \qquad \bar{x} = \frac{\sum y}{n} = \frac{792}{144} = 5.5$$

$$\text{SS}_{xy} = \sum xy - \frac{\sum x \sum y}{n} = 586.86 - \frac{792(103.07)}{144} = 19.975$$

$$\text{SS}_{xx} = \sum x^2 - \frac{\left(\sum x\right)^2}{n} = 5{,}112 - \frac{792^2}{144} = 756$$

$$\hat{\beta}_1 = \frac{\text{SS}_{xy}}{\text{SS}_{xx}} = \frac{19.975}{756} = .026421957$$

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1\overline{x} = \frac{103.07}{144} - (.026421957)\left(\frac{792}{144}\right) = .570443121$$

The estimated regression line is $\hat{y} = .5704 + .0264x$

$\hat{\beta}_0 = .5704$. Since $x = 0$ is not in the observed range, $\hat{\beta}_0$ has no interpretation other than being the y-intercept.

$\hat{\beta}_1 = -.0023$. For each additional unit increase in position, the mean proportion of words recalled is estimated to increase by .0264.

11.32 The four assumptions made about the probability distribution of ε in regression are:

1. The mean of the probability distribution of ε is 0.
2. The variance of the probability distribution of ε is constant for all settings of the independent variable x.
3. The probability distribution of ε is normal.
4. The values of ε associated with any two observed values of y are independent.

11.34 The graph in **b** would have the smallest s^2 because the width of the data points is the smallest.

11.36 a. $s^2 = \dfrac{\text{SSE}}{n-2} = \dfrac{.429}{12-2} = .0429$

b. $s = \sqrt{s^2} = \sqrt{.0429} = .2071$

c. We would expect most of the observations to be within 2*s* of the least squares line. This is:

$$2s = 2\sqrt{.0429} \approx .414$$

11.38 a. $s^2 = \dfrac{SSE}{n-2} = \dfrac{1.04}{28-2} = .04$ and $s = \sqrt{.04} = .2$.

b. We would expect most of the observations to fall within 2*s* or 2(.2) or .4 units of the least squares prediction line.

11.40 About 95% of the observations will fall within 2 standard deviations (2s) of their respective means. In this case, 2s = 2(.1) = .2 = *d*.

11.42 a. From Exercise 11.24, $\hat{\beta}_1 = .966521957$ and SS$_{xy}$ = 77.18.

Some preliminary calculations are:

$$SS_{yy} = \sum y_i^2 - \frac{\left(\sum y_i\right)^2}{n} = 1,710.2 - \frac{97.8^2}{6} = 116.06$$

$$SSE = SS_{yy} - \hat{\beta}_1 SS_{xy} = 116.06 - .966521957(77.18) = 41.4638$$

$$s^2 = \frac{SSE}{n-2} = \frac{41.4638}{6} = 6.9106$$

$$s = \sqrt{6.9106} = 2.6288$$

b. When x = 10, $\hat{y} = 6.313 + .9665(10) = 15.978$. The error of prediction is 2s = 2(2.6288) = 5.2576.

11.44 a. From Exercise 11.28, $\hat{\beta}_1 = -.002310625$ and SS_{xy} = -130.441667.

Some preliminary calculations are:

$$SS_{yy} = \sum y_i^2 - \frac{\left(\sum y_i\right)^2}{n} = 769.72 - \frac{135.8^2}{24} = 1.3183333$$

$$SSE = SS_{yy} - \hat{\beta}_1 SS_{xy} = 1.3183333 - (-.002310625)(-130.441667) = 1.016931523$$

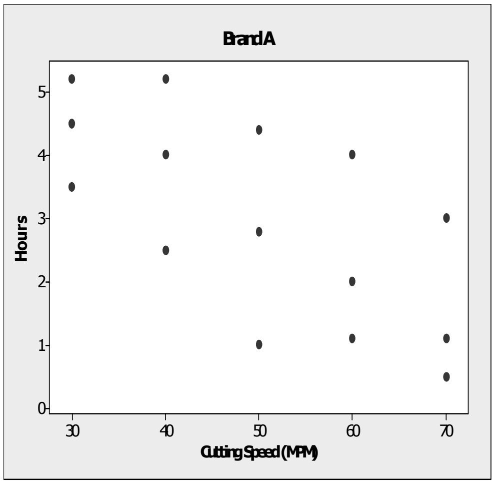
$$s^2 = \frac{SSE}{n-2} = \frac{1.016931523}{24-2} = .046224$$

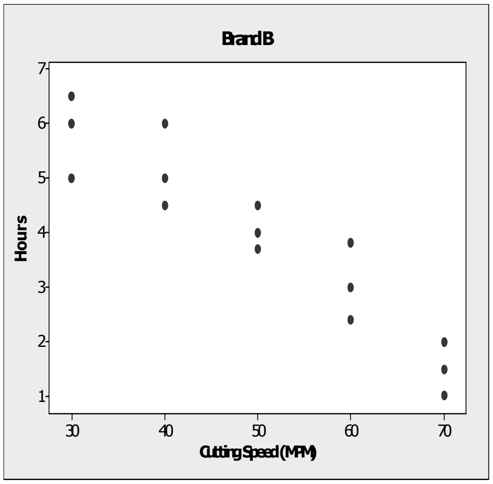
$$s = \sqrt{.046224} = .2150$$

b. The unit of s^2 is sweetness index squared. This number is very difficult to interpret in terms of the problem.

c. We would expect about 95% of the errors of prediction to fall within 2s = 2(.2150) = .43 units of 0 or between -.43 and .43.

11.46 Scatterplots of the two sets of data are:





Since the data points for Brand B are not spread apart as much as those for Brand A, it appears that Brand B would be a better predictor for useful life than Brand A.

For Brand A,

$$\sum x_i = 750 \qquad \sum y_i = 44.8 \qquad \sum x_i y_i = 2022 \qquad \sum x_i^2 = 40{,}500 \qquad \sum y_i^2 = 168.7$$

$$\text{SS}_{xx} = \sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n} = 40{,}500 - \frac{750^2}{15} = 40{,}500 - 37{,}500 = 3000$$

$$\text{SS}_{xy} = \sum x_i y_i - \frac{\left(\sum x_i\right)\left(\sum y_i\right)}{n} = 2022 - \frac{(750)(44.8)}{15} = 2022 - 2240 = -218$$

$$\hat{\beta}_1 = \frac{\text{SS}_{xy}}{\text{SS}_{xx}} = \frac{-218}{3000} = -.07266667 \approx -.0727$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = \frac{44.8}{15} - (-.07266667)\left(\frac{750}{15}\right) = 2.9866667 + 3.633333 = 6.62$$

$$\hat{y} = 6.62 - .0727x$$

For Brand B,

$$\sum x_i = 750 \qquad \sum y_i = 58.9 \qquad \sum x_i y_i = 2622 \qquad \sum x_i^2 = 40{,}500 \qquad \sum y_i^2 = 270.89$$

$$\text{SS}_{xx} = \sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n} = 40{,}500 - \frac{(750)^2}{15} = 40{,}500 - 37{,}500 = 3000$$

$$\text{SS}_{xy} = \sum xy - \frac{\left(\sum x\right)\left(\sum y\right)}{n} = 2622 - \frac{(750)(58.9)}{15} = 2622 - 2945 = -323$$

$$\hat{\beta}_1 = \frac{\text{SS}_{xy}}{\text{SS}_{xx}} = \frac{-323}{3000} = -.10766667 \approx -.1077$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x} = \left(\frac{59.9}{15}\right) - (-.10766667)\left(\frac{750}{15}\right) = 3.92667 + 5.38333 = 9.31$$

$$\hat{y} = 9.31 - .1077x$$

For Brand A,

$$\text{SS}_{yy} = \sum y_i^2 - \frac{\left(\sum y_i\right)^2}{n} = 168.7 - \frac{(44.8)^2}{15} = 168.7 - 133.802667 = 34.8973333$$

$$\begin{aligned}\text{SSE} = \text{SS}_{yy} - \hat{\beta}_1\text{SS}_{xy} &= 34.8973333 - (-.07266667)(-218)\\ &= 34.8973333 - 15.8413333 = 19.056\end{aligned}$$

$$s^2 = \frac{\text{SSE}}{n-2} = \frac{19.056}{13} = 1.465846154 \qquad s = \sqrt{1.465846154} = 1.211$$

For Brand B,

$$\text{SS}_{yy} = \sum y_i^2 - \frac{\left(\sum y_i\right)^2}{n} 270.89 - \frac{(58.9)^2}{15} = 270.89 - 231.2806667 = 39.6093333$$

$$\begin{aligned}\text{SSE} = \text{SS}_{yy} - \hat{\beta}_1\text{SS}_{xy} &= 39.6093333 - (-.10766667)(-323)\\ &= 39.6093333 - 34.7763333 = 4.833\end{aligned}$$

$$s^2 = \frac{\text{SSE}}{n-2} = \frac{4.833}{13} = .37176923 \qquad s = \sqrt{.37176923} = .610$$

Since the standard deviation (s = .610) for Brand B is smaller than the standard deviation for Brand A (s = 1.211), Brand B would be a better predictor for the useful life for a given cutting speed.

11.48 The conditions required for valid inferences about the β's in simple linear regression are:

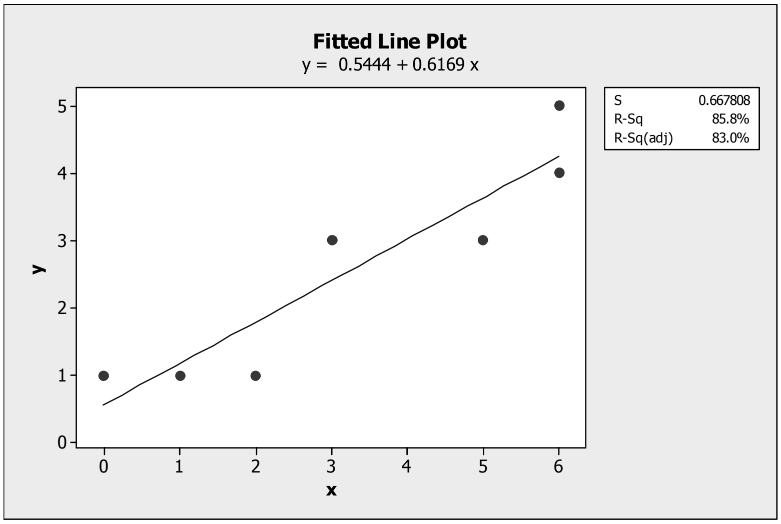
1. The mean of the probability distribution of ε is 0.
2. The variance of the probability distribution of ε is constant for all settings of the independent variable x.
3. The probability distribution of ε is normal.
4. The values of ε associated with any two observed values of y are independent.

11.50 a. For the confidence interval (22, 58) there is evidence of a positive linear relationship between y and x because the entire interval is composed of positive numbers.

b. For the confidence interval (–30, 111) there is no evidence of either a positive linear relationship or a negative linear relationship between *y* and *x* because 0 is contained in the interval.

c. For the confidence interval (–45, –7) there is evidence of a negative linear relationship between *y* and *x* because the entire interval is composed of negative numbers.

11.52 a. Using MINITAB, the scatterplot is:



b. Some preliminary calculations are:

$$\sum x = 23 \qquad \sum x^2 = 111 \qquad \sum xy = 81 \qquad \sum y = 18 \qquad \sum y^2 = 62$$

$$\text{SS}_{xy} = \sum xy - \frac{\sum x \sum y}{n} = 81 - \frac{23(18)}{7} = 21.85714286$$

$$\text{SS}_{xx} = \sum x^2 - \frac{\left(\sum x\right)^2}{n} = 111 - \frac{23^2}{7} = 35.42857143$$

$$\text{SS}_{yy} = \sum y^2 - \frac{\left(\sum y\right)^2}{n} = 62 - \frac{18^2}{7} = 15.71428571$$

$$\hat{\beta}_1 = \frac{\text{SS}_{xy}}{\text{SS}_{xx}} = \frac{21.85714286}{35.42857143} = .616935483 \approx .617$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = \frac{18}{7} - .616935483 \frac{23}{7} = .544354838 \approx .544$$

The least squares line is $\hat{y} = .544 + .617x$

c. The line is plotted on the graph in **a**.

d. To determine if *x* contributes information for the linear prediction of *y*, we test:

H_0: $\beta_1 = 0$
H_a: $\beta_1 \neq 0$

e. The test statistic is $t = \dfrac{\hat{\beta}_1 - 0}{\dfrac{s}{\sqrt{\text{SS}_{xx}}}} = \dfrac{.617 - 0}{\dfrac{.6678}{\sqrt{35.42857143}}} = 5.50$

where $\text{SSE} = \text{SS}_{yy} - \hat{\beta}_1 \text{SS}_{xy} = 15.71428571 - .616935483(21.85714286) = 2.22983872$

$s^2 = \dfrac{\text{SSE}}{n-2} = \dfrac{2.22983872}{7-2} = .44596774$ $\qquad$ $s = \sqrt{.44596774} = .6678$

The degrees of freedom are df = n – 2 = 7 – 2 = 5.

f. The rejection region requires $\alpha / 2 = .05 / 2 = .025$ in each tail of the *t* distribution with df = n – 2 = 7 – 2 = 5. From Table VI, Appendix A, $t_{.025}$ = 2.571. The rejection region is $t < -2.571$ or $t > 2.571$.

Since the observed value of the test statistic falls in the rejection region (t = 5.50 > 2.571), H_0 is rejected. There is sufficient evidence to indicate *x* contributes information for the linear prediction of *y* at $\alpha = .05$.

g. For confidence coefficient .95, $\alpha = 1 - .95 = .05$ and $\alpha / 2 = .05 / 2 = .025$. From Table VI, Appendix A, with df = n – 2 = 7 – 2 = 5, $t_{.025}$ = 2.571. The 95% confidence interval is:

$$\hat{\beta}_1 \pm t_{.025} s_{\hat{\beta}_1} \Rightarrow \hat{\beta}_1 \pm t_{.025} \frac{s}{\sqrt{\text{SS}_{xx}}} \Rightarrow .617 \pm 2.571 \frac{.6678}{\sqrt{35.42857143}}$$

$$\Rightarrow .617 \pm .288 \Rightarrow (.329,\ .905)$$

11.54 a. Since the *p*-value is greater than $\alpha = .05$ (p = .739 > .05), H_0 is not rejected. There is insufficient evidence to indicate that the ESLR score is linearly related to SG scores at $\alpha = .05$.

b. Since the *p*-value is less than $\alpha = .05$ (p = .012 < .05), H_0 is rejected. There is sufficient evidence to indicate that the ESLR score is linearly related to SR scores at $\alpha = .05$.

c. Since the *p*-value is less than $\alpha = .05$ (p = .022 < .05), H_0 is rejected. There is sufficient evidence to indicate that the ESLR score is linearly related to ER scores at $\alpha = .05$.

11.56 a. To determine whether driving accuracy decreases linearly as driving distance increases, we test:

H_0: $\beta_1 = 0$
H_a: $\beta_1 < 0$

b. From the results in Exercise 11.25, the test statistic is $t = -13.23$ and the p-value is $p = .000$.

c. Since the p-value is less than $\alpha = .01$ ($p = .000 < .01$), H_0 is rejected. There is sufficient evidence to indicate driving accuracy decreases linearly as driving distance increases at $\alpha = .01$.

11.58. a. From the results in Exercise 11.45 **a**, $\hat{\beta}_1 = .596$ and $s = 2.05688$. Also,

$$SS_{xx} = \sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n} = 379{,}604 - \frac{5{,}332^2}{75} = 534.3467\,.$$

For confidence coefficient .90, $\alpha = 1 - .90 = .10$ and $\alpha / 2 = .10 / 2 = .05$. From Table VI, Appendix A, with df = $n - 2 = 75 - 2 = 73$, $t_{.05} \approx 1.671$. The 90% confidence interval is:

$$\hat{\beta}_1 \pm t_{.05} s_{\hat{\beta}_1} \Rightarrow \hat{\beta}_1 \pm t_{.05}\frac{s}{\sqrt{\text{SS}_{xx}}} \Rightarrow .596 \pm 1.671\frac{2.05688}{\sqrt{534.3467}} \Rightarrow .596 \pm .149 \Rightarrow (.447,\ \ .745)$$

We are 90% confident that the change in the mean ideal partner's height for males for each unit increase in male student's height is between .447 and .745 inches.

b. From the results in Exercise 11.45 b, $\hat{\beta}_1 = .493$ and $s = 2.32153$. Also,

$$SS_{xx} = \sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n} = 300{,}768 - \frac{4{,}649^2}{72} = 584.6528\,.$$

For confidence coefficient .90, $\alpha = 1 - .90 = .10$ and $\alpha / 2 = .10 / 2 = .05$. From Table VI, Appendix A, with $df = n - 2 = 72 - 2 = 70$, $t_{.05} \approx 1.671$. The 90% confidence interval is:

$$\hat{\beta}_1 \pm t_{.05} s_{\hat{\beta}_1} \Rightarrow \hat{\beta}_1 \pm t_{.05}\frac{s}{\sqrt{\text{SS}_{xx}}} \Rightarrow .493 \pm 1.671\frac{2.32153}{\sqrt{584.6528}} \Rightarrow .493 \pm .160 \Rightarrow (.333,\ \ .653)$$

We are 90% confident that the change in the mean ideal partner's height for females for each unit increase in female student's height is between .333 and .653 inches.

c. The males have a greater increase in ideal partner's height for every 1 inch increase in student's height than females.

11.60 Some preliminary calculations are:

$$\bar{y} = \frac{\sum y}{n} = \frac{78.8}{16} = 4.925 \qquad \bar{x} = \frac{\sum x}{n} = \frac{247}{16} = 15.4375$$

$$\text{SS}_{xy} = \sum xy - \frac{\sum x \sum y}{n} = 1{,}264.6 - \frac{247(78.8)}{16} = 48.125$$

$$\text{SS}_{xx} = \sum x^2 - \frac{\left(\sum x\right)^2}{n} = 4{,}193 - \frac{247^2}{16} = 379.9375$$

$$\hat{\beta}_1 = \frac{\text{SS}_{xy}}{\text{SS}_{xx}} = \frac{48.125}{379.9375} = .12666557$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x} = \frac{78.8}{16} - (.12666557)\left(\frac{247}{16}\right) = 2.969600263$$

$$\text{SS}_{yy} = \sum y^2 - \frac{\left(\sum y\right)^2}{n} = 406.84 - \frac{78.8^2}{16} = 18.75$$

$$\text{SSE} = \text{SS}_{yy} - \hat{\beta}_1\left(\text{SS}_{xy}\right) = 18.75 - (.12666557)(48.125) = 12.65421994$$

$$s^2 = \frac{\text{SSE}}{n-2} = \frac{12.65421994}{16-2} = .903872817$$

$$s = \sqrt{s^2} = \sqrt{.903872817} = .95072226$$

To determine whether blood lactate level is linearly related to perceived recovery, we test:

H_0: $\beta_1 = 0$
H_a: $\beta_1 \neq 0$

The test statistic is $t = \dfrac{\hat{\beta}_1 - 0}{s_{\hat{\beta}}} = \dfrac{\hat{\beta}_1 - 0}{s\Big/\sqrt{\text{SS}_{xx}}} = \dfrac{.12667 - 0}{.95072\Big/\sqrt{379.9375}} = 2.597$

The rejection region requires $\alpha / 2 = .10 / 2 = .05$ in each tail of the t distribution. From Table VI, Appendix A, with df = n – 2 = 16 – 2 = 14, $t_{.05}$ = 1.761. The rejection region is $t < -1.761$ or $t > 1.761$.

Since the observed test statistic falls in the rejection region (t = 2.597 > 1.761), H_0 is rejected. There is sufficient evidence to indicate blood lactate level is linearly related to perceived recovery at $\alpha = .10$.

11.62 Some preliminary calculations are:

$\sum x_i = 288$ $\sum y_i = 4.14$ $\sum x_i y_i = 80.96$ $\sum x_i^2 = 5{,}362$ $\sum y_i^2 = 1.663$

$$\text{SS}_{xy} = \sum x_i y_i - \frac{\sum x_i \sum y_i}{n} = 80.96 - \frac{288(4.14)}{16} = 6.44$$

$$\text{SS}_{xx} = \sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n} = 5{,}362 - \frac{288^2}{16} = 178$$

$$\hat{\beta}_1 = \frac{\text{SS}_{xy}}{\text{SS}_{xx}} = \frac{6.44}{178} = .036179775$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x} = \frac{4.14}{16} - (.036179775)\left(\frac{288}{16}\right) = -.392485955$$

$$\text{SS}_{yy} = \sum y_i^2 - \frac{\left(\sum y_i\right)^2}{n} = 1.663 - \frac{4.14^2}{16} = .591775$$

$$\text{SSE} = \text{SS}_{yy} - \hat{\beta}_1\text{SS}_{xy} = .591775 - (.036179775)(6.44) = .358777249$$

$$s^2 = \frac{\text{SSE}}{n-2} = \frac{.358777249}{16-2} = .025626946 \qquad s = \sqrt{s^2} = \sqrt{.025626946} = .160084185$$

To determine if people scoring higher in empathy show higher pain-related brain activity, we test:

H_0: $\beta_1 = 0$
H_a: $\beta_1 > 0$

The test statistic is $t = \dfrac{\hat{\beta}_1 - 0}{s_{\hat{\beta}_1}} = \dfrac{.0362}{\left(\dfrac{.1601}{\sqrt{178}}\right)} = 3.017$

Since no α level was specified in the Exercise, we will use $\alpha = .05$. The rejection region requires $\alpha = .05$ in the upper tail of the t distribution with df = $n - 2 = 16 - 2 = 14$. From Table VI, Appendix A, $t_{.05} = 1.761$. The rejection region is $t > 1.761$.

Since the observed value of the test statistic falls in the rejection region ($t = 3.017 > 1.761$), H_0 is rejected. There is sufficient evidence to indicate the people scoring higher in empathy show higher pain-related brain activity at $\alpha = .05$.

11.64 Using the calculations from Exercise 11.30 and these calculations:

$$\text{SS}_{yy} = \sum y^2 - \frac{\left(\sum y\right)^2}{n} = 83.474 - \frac{103.07^2}{144} = 9.70021597$$

$$\text{SSE} = \text{SS}_{yy} - \hat{\beta}_1\left(\text{SS}_{xy}\right) = 9.70021597 - (.026421957)(19.975) = 9.172437366$$

$$s^2 = \frac{\text{SSE}}{n-2} = \frac{9.172437366}{144-2} = .064594629$$

$$s = \sqrt{s^2} = \sqrt{.064594629} = .254154735$$

To determine if there is a linear trend between the proportion of names recalled and position, we test:

H_0: $\beta_1 = 0$
H_a: $\beta_1 \neq 0$

The test statistic is $t = \dfrac{\hat{\beta}_1 - 0}{s_{\hat{\beta}_1}} = \dfrac{\hat{\beta}_1 - 0}{s \big/ \sqrt{\text{SS}_{xx}}} = \dfrac{.02642}{.25415 \big/ \sqrt{756}} = 2.858$

The rejection region requires $\alpha / 2 = .01 / 2 = .005$ in each tail of the t distribution. From Table VI, Appendix A, with df = n – 2 = 144 – 2 = 142, $t_{.005} \approx$ 2.576. The rejection region is $t < -2.576$ or $t > 2.576$.

Since the observed test statistic falls in the rejection region (t = 2.858 > 2.576), H_0 is rejected. There is sufficient evidence to indicate the proportion of names recalled is linearly related to position at $\alpha = .01$.

11.66 Using MINITAB, the results of fitting the regression model are:

Regression Analysis: Mass versus Time

```
The regression equation is
Mass = 5.22 - 0.114 Time


Predictor      Coef   SE Coef       T      P
Constant     5.2207    0.2960   17.64  0.000
Time       -0.11402   0.01032  -11.05  0.000


S = 0.857257   R-Sq = 85.3%   R-Sq(adj) = 84.6%


Analysis of Variance

Source          DF       SS      MS       F      P
Regression       1   89.794  89.794  122.19  0.000
Residual Error  21   15.433   0.735
Total           22  105.227
```

To determine if the mass of the spill tends to diminish linearly as elapsed time increases, we test:

H_0: $\beta_1 = 0$
H_a: $\beta_1 < 0$

From the printout, the test statistic is t = -11.05.

The rejection region requires α = .05 in the lower tail of the t-distribution with $df = n - 2$ = 23 – 2 = 21. From Table VI, Appendix A, $t_{.05}$ = 1.721. The rejection region is t < -1.721.

Since the observed value of the test statistic falls in the rejection region ($t = -11.05 < -1.721$), H_0 is rejected. There is sufficient evidence to indicate the mass of the spill tends to diminish linearly as elapsed time increases at $\alpha = .05$.

For confidence level .95, $\alpha = .05$ and $\alpha / 2 = .05 / 2 = .025$. From Table VI, Appendix A with $df = n - 2 = 23 - 2 = 21$, $t_{.025} = 2.080$.

The confidence interval is:

$$\hat{\beta}_1 \pm t_{.025} s_{\hat{\beta}_1} \Rightarrow -.114 \pm 2.080(.0103) \Rightarrow -.114 \pm .0214 \Rightarrow (-.1354, -.0926)$$

We are 95% confident that for each additional minute of elapsed time, the mean spill mass will decrease anywhere from .0926 and .1354 pounds.

11.68 a. If $r = .7$, there is a positive linear relationship between x and y. As x increases, y tends to increase. The slope is positive.

b. If $r = -.7$, there is a negative linear relationship between x and y. As x increases, y tends to decrease. The slope is negative.

c. If $r = 0$, there is a 0 slope. There is no linear relationship between x and y.

d. If $r^2 = .64$, then r is either .8 or $-.8$. The relationship between x and y could be either positive or negative.

11.70 The statement "A value of the correlation coefficient near 1 or near -1 implies a casual relationship between x and y." is a false statement. High values of the sample correlation do not infer causal relationships, but just strong linear relationships between the variables.

11.72 From Exercises 11.14 and 11.37,

$$r^2 = 1 - \frac{\text{SSE}}{\text{SS}_{yy}} = 1 - \frac{1.22033896}{21.7142857} = 1 - .0562 = .9438$$

94.38% of the total sample variability around $\bar{y}$ is explained by the linear relationship between y and x.

11.74 a. The linear model would be $E(y) = \beta_0 + \beta_1 x$.

b. $r = .68$. There is a moderate, positive linear relationship between RMP and SET ratings.

c. The slope is positive since the value of r is positive.

d. Since the p-value is very small ($p = .001$), we would reject H_0 for any value of α greater than .001. There is sufficient evidence to indicate a significant linear relationship between RMP and SET for $\alpha > .001$.

e. $r^2 = .68^2 = .4624$. 46.24% of the total sample variability around the sample mean RMP values is explained by the linear relationship between RMP and SET values.

11.76 a. From the printout, the value of r^2 is .5315. 53.15% of the sample variability around the sample mean total catch is explained by the linear relationship between total catch and search frequency.

b. Yes. We can look at the estimate of β_1. From the printout, $\hat{\beta}_1 = -171.57265$. Since the estimate is negative, the total catch is negatively linearly related to search frequency.

11.78 **Radiata Pine**: $r^2 = .84$. 84% of the total sample variability around the sample mean stress is explained by the linear relationship between stress and the natural logarithm of number of blade cycles.

Hoop Pine: $r^2 = .90$. 90% of the total sample variability around the sample mean stress is explained by the linear relationship between stress and the natural logarithm of number of blade cycles.

11.80 a. $r = .84$. Since the value is fairly close to 1, there is a moderately strong positive linear relationship between the magnitude of a QSO and the redshift level.

b. The relationship between r and the estimated slope of the line is that they will both have the same sign. If r is positive, the slope of the line is positive. If r is negative, the slope of the line is negative.

c. $r^2 = .84^2 = .7056$. 70.56% of the total sample variability around the sample mean magnitude of a QSO is explained by the linear relationship between magnitude of a QSO and redshift level.

11.82 a. $r = .41$. Since the value is not particularly close to 1, there is a moderately weak positive linear relationship between average earnings and height for those in sales occupations.

b. $r^2 = .41^2 = .1681$. 16.81% of the total sample variability around the sample mean average earnings is explained by the linear relationship between average earnings and height for those in sales occupations.

c. To determine whether average earnings and height are positively correlated, we test:

H_0: $\rho = 0$
H_a: $\rho > 0$

d. The test statistic is $t = \dfrac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \dfrac{.41\sqrt{117-2}}{\sqrt{1-.41^2}} = 4.82$.

e. The rejection region requires $\alpha = .01$ in the upper tail of the t distribution with df = $n - 2$ = 117 – 2 = 115. From Table VI, Appendix A, $t_{.01} \approx 2.358$. The rejection region is $t > 2.358$.

Since the observed value of the test statistic falls in the rejection region ($t = 4.82 > 2.358$), H_0 is rejected. There is sufficient evidence to indicate that average earnings and height are positively correlated for sales occupations at $\alpha = .01$.

f. We will select Managers.

$r = .35$. Since the value is not particularly close to 1, there is a moderately weak positive linear relationship between average earnings and height for managers.

$r^2 = .35^2 = .1225$. 12.25% of the total sample variability around the sample mean average earnings is explained by the linear relationship between average earnings and height for managers.

To determine whether average earnings and height are positively correlated, we test:

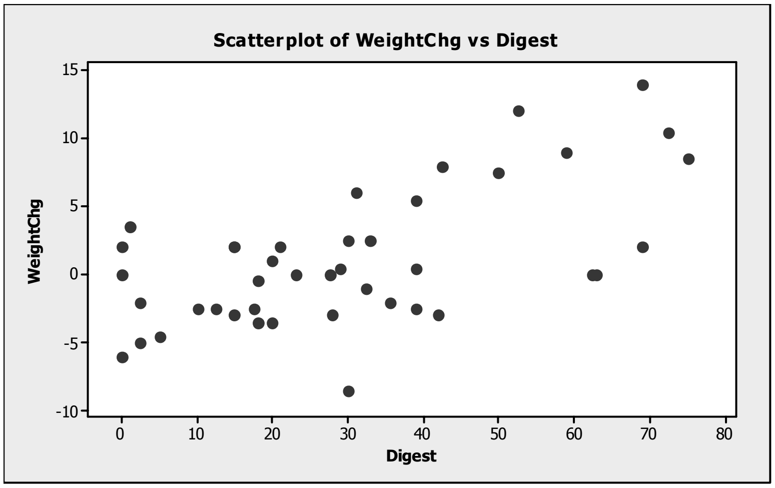
H_0: $\rho = 0$
H_a: $\rho > 0$

The test statistic is $t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{.35\sqrt{455-2}}{\sqrt{1-.35^2}} = 7.95$

The rejection region requires $\alpha = .01$ in the upper tail of the t distribution with df = $n - 2 = 455 - 2 = 453$. From Table VI, Appendix A, $t_{.01} \approx 2.326$. The rejection region is $t > 2.326$.

Since the observed value of the test statistic falls in the rejection region ($t = 7.95 > 2.326$), H_0 is rejected. There is sufficient evidence to indicate that average earnings and height are positively correlated for managers at $\alpha = .01$.

11.84 a. Using MINITAB, the plot of weight change and digestions efficiency is:



Yes. There appears to be a positive linear trend. As digestion efficiency (%) increases, weight change (%) tends to increase.

b. Using MINITAB, the results are:

Correlations: WeightChg, Digest

```
Pearson correlation of WeightChg and Digest = 0.612
P-Value = 0.000
```

Thus, $r = .612$. Since the value is near .5, there is a moderate positive linear relationship between weight change (%) and digestion efficiency (%).

c. To determine if weight change is correlated to digestion efficiency, we test:

H_0: $\rho = 0$
H_a: $\rho \neq 0$

The test statistic is $t = \dfrac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \dfrac{.612\sqrt{42-2}}{\sqrt{1-.612^2}} = 4.89$

The rejection region requires $\alpha / 2 = .01/2 = .005$ in each tail of the t distribution with df $= n - 2 = 42 - 2 = 40$. From Table VI, Appendix A, $t_{.005} = 2.704$. The rejection region is $t < -2.704$ or $t > 2.704$.

Since the observed value of the test statistic falls in the rejection region ($t = 4.89 > 2.704$), H_0 is rejected. There is sufficient evidence to indicate weight change is correlated to digestion efficiency at $\alpha = .01$.

d. Using MINITAB, the results for all observations except the trials using duck chow are:

Correlations: WeightChg2, Digest2

```
Pearson correlation of WeightChg2 and Digest2 = 0.309
P-Value = 0.080
```

Thus, $r = .309$. Since the value is near 0, there is a very weak positive linear relationship between weight change (%) and digestion efficiency (%).

To determine if weight change is correlated to digestion efficiency, we test:

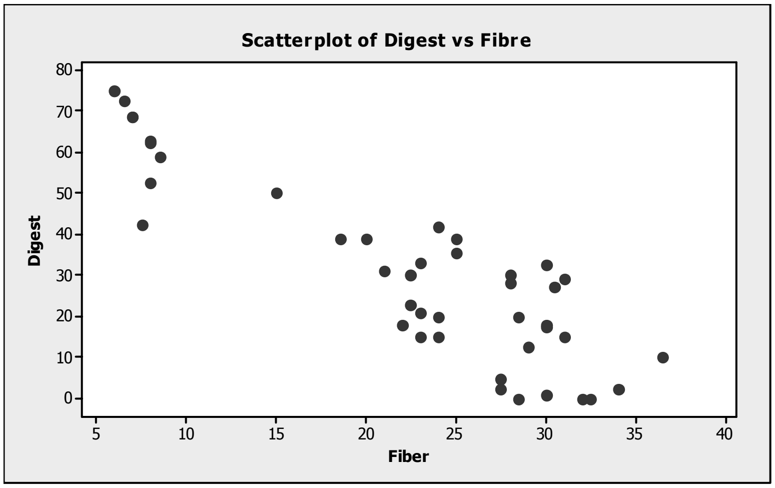
H_0: $\rho = 0$
H_a: $\rho \neq 0$

The test statistic is $t = \dfrac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \dfrac{.309\sqrt{33-2}}{\sqrt{1-.309^2}} = 1.81$

The rejection region requires $\alpha / 2 = .01/2 = .005$ in each tail of the t distribution with df $= n - 2 = 33 - 2 = 31$. From Table VI, Appendix A, $t_{.005} \approx 2.75$. The rejection region is $t < -2.75$ or $t > 2.75$.

Since the observed value of the test statistic does not fall in the rejection region ($t = 1.81 \ngtr 2.75$), H_0 is not rejected. There is insufficient evidence to indicate weight change is correlated to digestion efficiency at $\alpha = .01$ for those not using duck chow.

e. a. Using MINITAB, the plot of digestion efficiency and fibre is:



Yes. There appears to be a negative linear trend. As fiber (%) increases, digestion efficiency (%) tends to decrease.

b. Using MINITAB, the results are:

Correlations: Digest, Fibre

```
Pearson correlation of Digest and Fibre = -0.880
P-Value = 0.000
```

Thus, $r = -.880$. Since the value is fairly near -1, there is a fairly strong negative linear relationship between digestion efficiency (%) and fibre (%).

c. To determine if digestion efficiency is related to fibre, we test:

$$H_0\text{: } \rho = 0$$
$$H_a\text{: } \rho \neq 0$$

The test statistic is $t = \dfrac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \dfrac{-.88\sqrt{42-2}}{\sqrt{1-(-.88)^2}} = -11.72$

The rejection region requires $\alpha / 2 = .01 / 2 = .005$ in each tail of the t distribution with df = $n - 2 = 42 - 2 = 40$. From Table VI, Appendix A, $t_{.005} = 2.704$. The rejection region is $t < -2.704$ or $t > 2.704$.

Since the observed value of the test statistic falls in the rejection region ($t = -11.72 < -2.704$), H_0 is rejected. There is sufficient evidence to indicate digestion efficiency is related to fibre at $\alpha = .01$.

d. Using MINITAB, the results for all observations except the trials using Duck Chow are:

Correlations: Digest2, Fibre2

```
Pearson correlation of Digest2 and Fibre2 = -0.646
P-Value = 0.000
```

Thus, $r = -.646$. Since the value is slightly bigger than .5, there is a moderately strong negative linear relationship between digestion efficiency (%) and fibre (%).

To determine if digestion efficiency is correlated to fibre , we test:

H_0: $\rho = 0$

H_a: $\rho \neq 0$

The test statistic is $t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{-.646\sqrt{33-2}}{\sqrt{1-(-.646)^2}} = -4.71$

The rejection region requires $\alpha/2 = .01/2 = .005$ in each tail of the t distribution with df = n – 2 = 33 – 2 = 31. From Table VI, Appendix A, $t_{.005} \approx 2.75$. The rejection region is $t < -2.75$ or $t > 2.75$.

Since the observed value of the test statistic falls in the rejection region ($t = -4.71 < -2.75$), H_0 is rejected. There is sufficient evidence to indicate digestion efficiency and fibre are correlated at $\alpha = .01$ for those not using duck chow.

11.86 Using the values computed in Exercises 11.30 and 11.64:

$$r = \frac{\text{SS}_{xy}}{\sqrt{\text{SS}_{xx}\text{SS}_{yy}}} = \frac{19.975}{\sqrt{756(9.700121597)}} = .2333$$

Because r is fairly close to 0, there is a very weak positive linear relationship between the proportion of names recalled and position.

$r^2 = .2333^2 = .0544$.

5.44% of the total sample variability around the sample mean proportion of names recalled is explained by the linear relationship between proportion of names recalled and position.

11.88 a. Since there was an inverse relationship, the value of r must be negative.

b. If the result was significant, then the test statistic must fall in the rejection region. For a one tailed test, $\alpha = .05$ must fall in the lower tail of the t distribution with df = n – 2 = 337 – 2 = 335. From Table VI, Appendix A, $t_{.05} \approx 1.645$. The rejection region is $t < -1.645$.

Using the equation given, then:

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} < -1.645$$

$$\Rightarrow \frac{r^2(n-2)}{1-r^2} > (-1.645)^2$$

$$\Rightarrow r^2(337-2) > 2.706025(1-r^2)$$

$$\Rightarrow r^2(335) + 2.706025r^2 > 2.706025$$

$$\Rightarrow r^2(337.706025) > 2.706025$$

$$\Rightarrow r^2 > \frac{2.706025}{337.706025} = .00801296$$

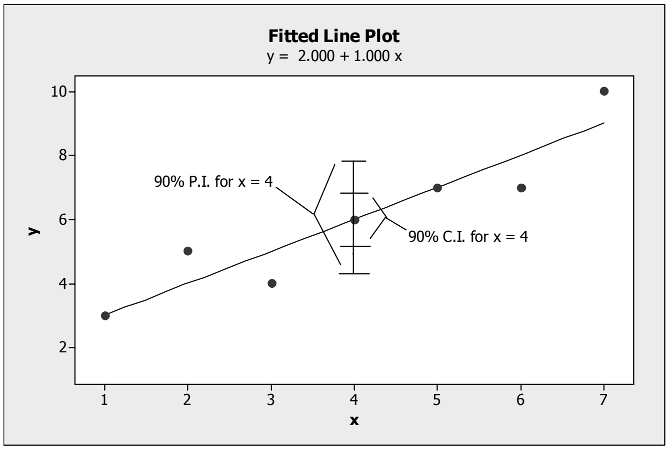
$$\Rightarrow r < -\sqrt{.00801296} = -.0895$$

11.90 The statement "For a given *x*, a confidence interval for *E*(*y*) will always be wider than a prediction interval for *y*." is false. The prediction interval for *y* will always be wider than the confidence interval for *E*(*y*) for a given value of *x*.

11.92 a. If a jeweler wants to predict the selling price of a diamond stone based on its size, he would use a prediction interval for *y*.

b. If a psychologist wants to estimate the average IQ of all patients that have a certain income level, he would use a confidence interval for *E*(*y*).

11.94 a. Using MINITAB, the plot is:



b. Some preliminary calculations are:

$\sum x_i = 28 \qquad \sum x_i^2 = 140 \qquad \sum x_i y_i = 196 \qquad \sum y_i = 42 \qquad \sum y_i^2 = 284$

$$\text{SS}_{xy} = \sum x_i y_i - \frac{\sum x_i \sum y_i}{n} = 196 - \frac{28(42)}{7} = 28$$

$$\text{SS}_{xx} = \sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n} = 140 - \frac{28^2}{7} = 28$$

$$\text{SS}_{yy} = \sum y_i^2 - \frac{\left(\sum y_i\right)^2}{n} = 284 - \frac{42^2}{7} = 32$$

$$\hat{\beta}_1 = \frac{\text{SS}_{xy}}{\text{SS}_{xx}} = \frac{28}{28} = 1 \qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = \frac{42}{7} - 1\left(\frac{28}{7}\right) = 6 - 4 = 2$$

The least squares line is $\hat{y} = 2 + x$.

c. $SSE = SS_{yy} - \hat{\beta}_1 SS_{xy} = 32 - 1(28) = 4$

$$s^2 = \frac{\text{SSE}}{n-2} = \frac{4}{5} = .8$$

d. The form of the confidence interval is $\hat{y} \pm t_{\alpha/2}\, s \sqrt{\frac{1}{n} + \frac{(x_\text{p} - \bar{x})^2}{\text{SS}_{xx}}}$

where $s = \sqrt{s^2} = \sqrt{.8} = .8944$. For $x_\text{p} = 4$, $\hat{y} = 2 + 4 = 6$, and $\bar{x} = \frac{28}{7} = 4$.

For confidence coefficient .90, $\alpha = 1 - .90 = .10$ and $\alpha/2 = .10/2 = .05$. From Table VI, Appendix A, $t_{.05} = 2.015$ with df = $n - 2 = 7 - 2 = 5$.

The 90% confidence interval is:

$$6 \pm 2.015(.8944)\sqrt{\frac{1}{7} + \frac{(4-4)^2}{28}} \Rightarrow 6 \pm .681 \Rightarrow (5.319,\ 6.681)$$

e. The form of the prediction interval is $\hat{y} \pm t_{\alpha/2}\, s \sqrt{1 + \frac{1}{n} + \frac{(x_\text{p} - \bar{x})^2}{\text{SS}_{xx}}}$

The 90% prediction interval is:

$$6 \pm 2.015(.8944)\sqrt{1 + \frac{1}{7} + \frac{(4-4)^2}{28}} \Rightarrow 6 \pm 1.927 \Rightarrow (4.073,\ 7.927)$$

f. The 95% prediction interval for y is wider than the 95% confidence interval for the mean value of y when $x_\text{p} = 4$.

The error of predicting a particular value of y will be larger than the error of estimating the mean value of y for a particular x value. This is true since the error in estimating the mean value of *y* for a given *x* value is the distance between the least squares line and the true line of means, while the error in predicting some future value of *y* is the sum of two errors–the error of estimating the mean of *y* plus the random error that is a component of the value of *y* to be predicted.

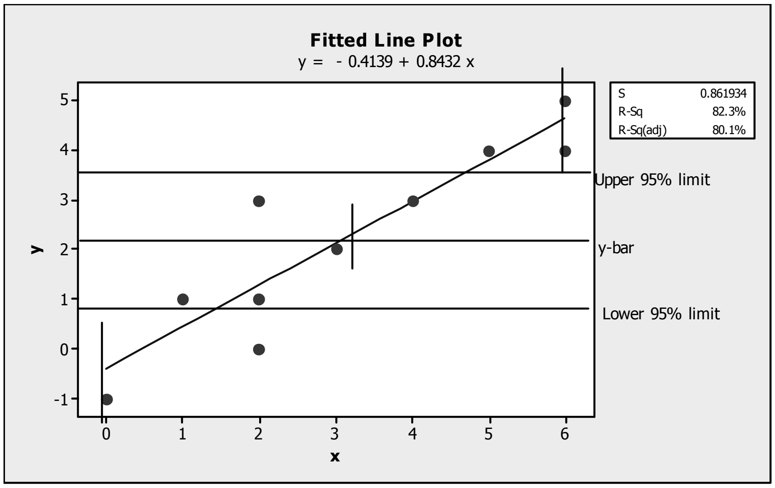
11.96 a. The form of the confidence interval is $\bar{y} \pm t_{\alpha/2}\frac{s}{\sqrt{n}}$ where $\bar{y} = \frac{\sum y}{n} = \frac{22}{10} = 2.2$,

$$s^2 = \frac{\sum y^2 - \frac{\left(\sum y\right)^2}{n}}{n-1} = \frac{82 - \frac{22^2}{10}}{10-1} = 3.733\text{, and } s = \sqrt{s^2} = \sqrt{3.733} = 1.932$$

For confidence coefficient .95, $\alpha = 1 - .95 = .05$ and $\alpha/2 = .05/2 = .025$. From Table VI, Appendix A, with $df = n - 1 = 10 - 1 = 9$, $t_{.025} = 2.262$. The 95% confidence interval is:

$$\bar{y} \pm t_{\alpha/2}\frac{s}{\sqrt{n}} \Rightarrow 2.2 \pm 2.262\frac{1.932}{\sqrt{10}} \Rightarrow 2.2 \pm 1.382 \Rightarrow (.818,\ \ 3.582)$$

b. Using MINITAB, the plot of the data is:



c. The intervals calculated in Exercise 11.95 are:

For $x_p = 6$, the 95% confidence interval is (3.526, 5.762)
For $x_p = 3.2$, the 95% confidence interval is (1.655, 2.913)
For $x_p = 0$, the 95% confidence interval is (–1.585, 0.757)

These intervals are all much narrower than the interval found in part a. They are also quite different, depending on the value of *x*. Thus, *x* appears to contribute information about the mean value of *y*.

d. Some preliminary calculations are:

$$\sum x_i = 31 \qquad \sum y_i = 22 \qquad \sum x_i y_i = 101 \qquad \sum x_i^2 = 135 \qquad \sum y_i^2 = 82$$

$$\text{SS}_{xy} = \sum x_i y_i - \frac{\sum x_i \sum y_i}{n} = 101 - \frac{31(22)}{10} = 32.8$$

$$\text{SS}_{xx} = \sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n} = 135 - \frac{31^2}{10} = 38.9$$

$$\hat{\beta}_1 = \frac{\text{SS}_{xy}}{\text{SS}_{xx}} = \frac{32.8}{38.9} = .84318766$$

$$\text{SS}_{yy} = \sum y_i^2 - \frac{\left(\sum y_i\right)^2}{n} = 82 - \frac{22^2}{10} = 33.6$$

$$SSE = \text{SS}_{yy} - \hat{\beta}_1 \text{SS}_{xy} = 33.6 - (.84318766)(32.8) = 5.943444752$$

$$s^2 = \frac{SSE}{n-2} = \frac{5.943444752}{10-2} = .742930594 \qquad s = \sqrt{s^2} = \sqrt{.742930594} = .861934216$$

H_0: $\beta_1 = 0$
H_a: $\beta_1 \neq 0$

The test statistic is $t = \dfrac{\hat{\beta}_1 - 0}{s_{\hat{\beta}_1}} = \dfrac{.843 - 0}{.862 / \sqrt{38.9}} = 6.100$

The rejection region requires $\alpha / 2 = .05 / 2 = .025$ in each tail of the t distribution with $df = n - 2 = 10 - 2 = 8$. From Table VI, Appendix A, $t_{.025} = 2.306$. The rejection region is $t < -2.306$ or $t > 2.306$.

Since the observed value of the test statistic falls in the rejection region ($t = 6.100 > 2.306$), H_0 is rejected. There is sufficient evidence to indicate that the straight-line model contributes information for the prediction of y at $\alpha = .05$.

11.98 a. The researchers should use a prediction interval to estimate the actual ELSR score based on a value of the independent variable of x = 50%.

b. The researchers should use a confidence interval for the mean ELSR score based on a value of the independent variable of x = 70%.

11.100 a. From the printout, the 95% prediction interval for driving accuracy for a driving distance of x = 300 yards is (56.724, 65.894). We are 95% confident that the actual driving accuracy for a golfer driving the ball 300 yards is between 56.724 and 65.894.

b. From the printout, the 95% confidence interval for mean driving accuracy for a driving distance of x = 300 yards is (60.586, 62.032). We are 95% confident that the mean driving accuracy for all golfers driving the ball 300 yards is between 60.586 and 62.032.

c. If we are interested in the average driving accuracy of all PGA golfers who have a driving distance of 300 yards, we would use the confidence interval for the mean driving accuracy. The confidence interval for the mean estimates the mean while to prediction interval for the actual value estimates a single value, not a mean.

11.102 a. From Exercise 11.27, SS_{xy} = 242,380, SS_{xx} = 1,150, $\hat{\beta}_0 = 1,469.351449$, and $\hat{\beta}_1 = 210.7652174$.

The least squares prediction equation is: $\hat{y} = 1,469.351 + 210.765x$

When x = 10, $\hat{y} = 1,469.351 + 210.765(10) = 3577.001$

b. Some preliminary calculations are:

$$\sum y = 98,494 \quad \sum y^2 = 456,565,950$$

$$SS_{yy} = \sum y_i^2 - \frac{\left(\sum y_i{}^2\right)}{n} = 456,565,950 - \frac{(98,494)^2}{24} = 52,354,781.8$$

$$SSE = SS_{yy} - \hat{\beta}_1 SS_{xy} = 52,354,781.8 - 210.7652174(242,380) = 1,269,508.407$$

$$s^2 = \frac{SSE}{n-2} = \frac{1,269,508.407}{24-2} = 57,704.92759$$

$$s = \sqrt{57,704.92759} = 240.2185$$

For confidence coefficient .90, $\alpha = .10$ and $\alpha/2 = .10/2 = .05$. From Table VI, Appendix A, with $df = n - 2 = 24 - 2 = 22$, $t_{.05} = 1.717$. The 90% prediction interval is:

$$\hat{y} \pm t_{\alpha/2} s\sqrt{1 + \frac{1}{n} + \frac{\left(x_p - \bar{x}\right)^2}{SS_{xx}}} \Rightarrow 3,577.001 \pm 1.717\left(240.2185\right)\sqrt{1 + \frac{1}{24} + \frac{\left(10 - 12.5\right)^2}{1,150}}$$

$$\Rightarrow 3,577.001 \pm 422.057 \Rightarrow (3,154.944, 3,999.058\ )$$

We are 90% confident that the actual sound wave frequency is between 3,154.944 and 3,999.058 when the resonance is 10.

c. A resonance of 30 is outside the observed range (we observed values of range from 1 to 24). Thus, we do not know what the relationship between resonance and frequency is outside the observed range. If the relationship stays the same outside the observed range, then there would be no danger in using the above formula. However, if the relationship between resonance and frequency changes outside the observed range, then using the above formula will lead to unreliable estimations.

11.104 a. From Exercises 11.30, 11.64 and 11.84, $\bar{x} = 5.5$, SS_{xx} = 756, s = .25415, and $\hat{y} = .5704 + .0264x$.

For x = 5, $\hat{y} = .5704 + .0264(5) = .7024$

For confidence coefficient .99, $\alpha = .01$ and $\alpha / 2 = .01 / 2 = .005$. From Table VI, Appendix A, with df = n – 2 = 144 – 2 = 142, $t_{.005} \approx 2.617$. The 99% confidence interval is:

$$\hat{y} \pm t_{\alpha/2} s\sqrt{\frac{1}{n}+\frac{\left(x_p-\bar{x}\right)^2}{\mathrm{SS}_{xx}}} \Rightarrow .7024 \pm 2.617(.2542)\sqrt{\frac{1}{144}+\frac{(5-5.5)^2}{756}}$$

$$\Rightarrow .7024 \pm .0567 \Rightarrow (.6457, .7591)$$

We are 99% confident that the mean recall of all those in the 5th position is between .6457 and .7591.

b. For confidence coefficient .99, $\alpha = .01$ and $\alpha / 2 = .01 / 2 = .005$. From Table VI, Appendix A, with df = n – 2 = 144 – 2 = 142, $t_{.005} \approx 2.617$. The 99% prediction interval is:

$$\hat{y} \pm t_{\alpha/2} s\sqrt{1+\frac{1}{n}+\frac{\left(x_p-\bar{x}\right)^2}{SS_{xx}}} \Rightarrow .7024 \pm 2.617(.2542)\sqrt{1+\frac{1}{144}+\frac{(5-5.5)^2}{756}}$$

$$\Rightarrow .7024 \pm .6677 \Rightarrow (.0347, 1.3701)$$

We are 99% confident that the actual recall of a person in the 5th position is between .0347 and 1.3701. Since the proportion of names recalled cannot be larger than 1, the actual proportion recalled will be between .0347 and 1.000.

c. The prediction interval in part **b** is wider than the confidence interval in part **a**. The prediction interval will always be wider than the confidence interval. The confidence interval for the mean is an interval for predicting the mean of all observations for a particular value of x. The prediction interval is a confidence interval for the actual value of the dependent variable for a particular value of x.

11.106 a. From MINITAB, the output is:

```
The regression equation is
weight = - 3.17 + 0.141 digest

Predictor        Coef       StDev          T        P
Constant       -3.171       1.068      -2.97    0.005
digest        0.14147     0.02889       4.90    0.000

S = 4.003       R-Sq = 37.5%     R-Sq(adj) = 35.9%

Analysis of Variance

Source            DF          SS          MS         F        P
Regression         1      384.24      384.24     23.98    0.000
Residual Error    40      640.88       16.02
Total             41     1025.12


Predicted Values

     Fit  StDev Fit         95.0% CI             95.0% PI
  -1.049      0.757   (  -2.579,   0.481)  (  -9.282,   7.185)
```

The least squares equation is $\hat{y} = -3.17 + .141x$.

b. To determine if digestion efficiency contributes to the estimation of weight change, we test:

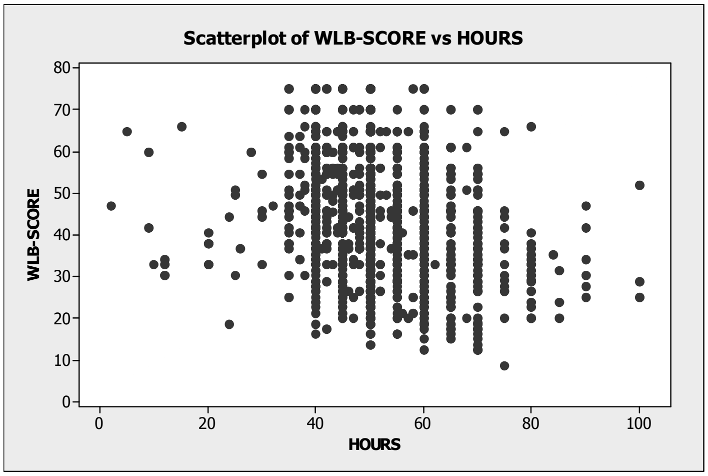
H_0: $\beta_1 = 0$
H_1: $\beta_1 \neq 0$

The test statistic is $t = 4.90$ and the p-value is $p < .001$.

Since the p-value is so small ($p < .001$), H_0 is rejected for any reasonable value of α. There is sufficient evidence to indicate that the model can be used to predict weight change for any reasonable value of α.

c. The 95% confidence interval, from the output, is: (−2.579, .481). We can be 95% confident that the mean weight change for all baby snow geese with digestion efficiency of 15% is between −2.579% and .481%.

11.108 Using MINITAB, a scatterplot of the data is:



From the plot, it looks like there could be a negative linear relationship between the WLB scores and the number of hours worked.

The descriptive statistics for the variables are:

Descriptive Statistics: WLB-SCORE, HOURS

```
Variable      N    Mean   StDev  Minimum      Q1  Median      Q3  Maximum
WLB-SCORE  2087  45.070  12.738    8.540  36.750  44.510  54.740   75.220
HOURS      2087  50.264   9.742    2.000  45.000  50.000  55.000  100.000
```

Using MINITAB, the results of fitting the regression model are:

Regression Analysis: WLB-SCORE versus HOURS

```
The regression equation is
WLB-SCORE = 62.5 - 0.347 HOURS


Predictor       Coef  SE Coef       T      P
Constant      62.499    1.414   44.22  0.000
HOURS       -0.34673  0.02761  -12.56  0.000


S = 12.2845   R-Sq = 7.0%   R-Sq(adj) = 7.0%


Analysis of Variance

Source            DF      SS     MS       F      P
Regression         1   23803  23803  157.73  0.000
Residual Error  2085  314647    151
Total           2086  338451
```

The fitted regression line is $\hat{y} = 62.5 - .347x$.

To determine if there is a negative linear relationship between WLB-scores and number of hours worked, we test:

H_0: $\beta_1 = 0$
H_a: $\beta_1 < 0$

The test statistic is t = -12.56 and the p-value is p = 0.000/2 = 0.000. Since the p-value is so small, H_0 is rejected for any reasonable value of α . There is sufficient evidence to indicate a negative linear relationship between WLB-scores and number of hours worked. The more hours worked per week, the lower the WLB-score.

From the printout, the value of r^2 is 7%. Thus, only 7% of the total sample variability around the sample mean WLB-scores is explained by the linear relationship between the WLB-scores and the number of hours worked. With only 7% of the variability explained, we know that there might be other factors not considered that can help explain the variability in the WLB-scores.

11.110 A probabilistic model contains 2 parts – a deterministic component and a random error component. The deterministic component (or deterministic model) allows for the prediction of y exactly from x. If you know the value of x, then you know the value of y. The random error component allows for the prediction of y to not be exactly determined by the value of x.

11.112 The five steps in a simple linear regression analysis are:

1. Hypothesize the deterministic component of the model that relates the mean, $E(y)$, to the independent variable x.

2. Use the sample data to estimate unknown parameters in the model.

3. Specify the probability distribution of the random error term and estimate the standard deviation of the distribution.

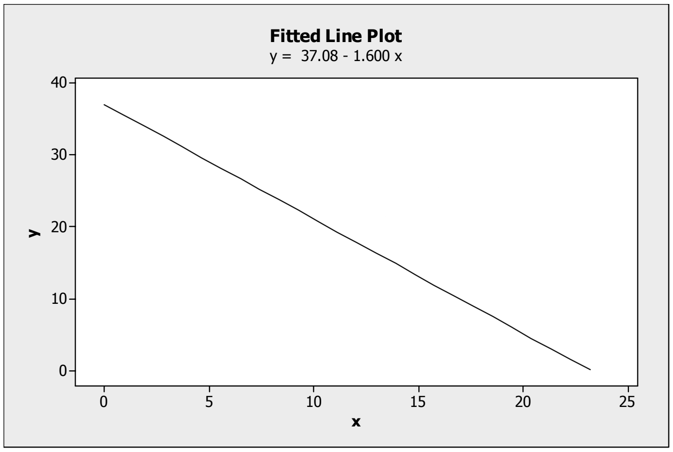
4. Statistically evaluate the usefulness of the model.

5. When satisfied that the model is useful, use it for prediction, estimation, and other purposes.

11.114 a. $\hat{\beta}_1 = \frac{\text{SS}_{xy}}{\text{SS}_{xx}} = \frac{-88}{55} = -1.6$, $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x} = 35 - (-1.6)(1.3) = 37.08$

The least squares line is $\hat{y} = 37.08 - 1.6x$.

b.



c. $\text{SSE} = \text{SS}_{yy} - \hat{\beta}_1\text{SS}_{xy} = 198 - (-1.6)(-88) = 57.2$

d. $s^2 = \frac{\text{SSE}}{n-2} = \frac{57.2}{15-2} = 4.4$

e. For confidence coefficient .90, $\alpha = 1 - .90 = .10$ and $\alpha / 2 = .10 / 2 = .05$. From Table VI, Appendix A, with df = n – 2 = 15 – 2 = 13, $t_{.05}$ = 1.771. The 90% confidence interval for β_1 is:

$$\hat{\beta}_1 \pm t_{\alpha/2}\frac{s}{\sqrt{\text{SS}_{xx}}} \Rightarrow -1.6 \pm 1.771\frac{\sqrt{4.4}}{\sqrt{55}} \Rightarrow -1.6 \pm .501 \Rightarrow (-2.101, -1.099)$$

We are 90% confident the change in the mean value of y for each unit change in x is between –2.101 and –1.099.

f. For x_p = 15, $\hat{y} = 37.08 - 1.6(15) = 13.08$

The 90% confidence interval is:

$$\hat{y} \pm t_{\alpha/2}\, s\sqrt{\frac{1}{n} + \frac{(x_p - \overline{x})^2}{\text{SS}_{xx}}} \Rightarrow 13.08 \pm 1.771\left(\sqrt{4.4}\right)\sqrt{\frac{1}{15} + \frac{(15-1.3)^2}{55}}$$

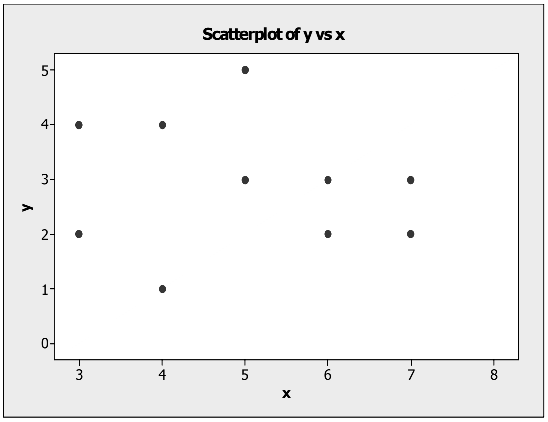
$$\Rightarrow 13.08 \pm 6.929 \Rightarrow (6.151,\ 20.009)$$

g. The 90% prediction interval is:

$$\hat{y} \pm t_{\alpha/2}\, s\sqrt{1+\frac{1}{n} + \frac{(x_p - \overline{x})^2}{\text{SS}_{xx}}} \Rightarrow 13.08 \pm 1.771\left(\sqrt{4.4}\right)\sqrt{1+\frac{1}{15} + \frac{(15-1.3)^2}{55}}$$

$$\Rightarrow 13.08 \pm 7.862 \Rightarrow (5.218,\ 20.942)$$

11.116 a. Using MINITAB, a scatterplot of the data is:



b. Some preliminary calculations are:

$$\sum x = 50 \qquad \sum x^2 = 270 \qquad \sum xy = 143 \qquad \sum y = 29 \qquad \sum y^2 = 97$$

$$\text{SS}_{xy} = \sum xy - \frac{\sum x \sum y}{n} = 143 - \frac{50(29)}{10} = -2$$

$$\text{SS}_{xx} = \sum x^2 - \frac{\left(\sum x\right)^2}{n} = 270 - \frac{50^2}{10} = 20$$

$$\text{SS}_{yy} = \sum y^2 - \frac{\left(\sum y\right)^2}{n} = 97 - \frac{29^2}{10} = 12.9$$

$$r = \frac{\text{SS}_{xy}}{\sqrt{\text{SS}_{xx}\text{SS}_{yy}}} = \frac{-2}{\sqrt{20(12.9)}} = -.1245 \qquad r^2 = (-.1245)^2 = .0155$$

c. To determine if *x* and *y* are linearly correlated, we test:

H_0: $\rho = 0$
H_a: $\rho \neq 0$

The test statistic is $t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{-.1245\sqrt{10-2}}{\sqrt{1-(-.1245)^2}} = -.35$

The rejection requires $\alpha / 2 = .10 / 2 = .05$ in the each tail of the *t* distribution with df = $n - 2 = 10 - 2 = 8$. From Table VI, Appendix A, $t_{.05} = 1.86$. The rejection region is $t < -1.86$ or $t > 1.86$.

Since the observed value of the test statistic does not fall in the rejection region ($t = -.35 \not< -1.86$), H_0 is not rejected. There is insufficient evidence to indicate that *x* and *y* are linearly correlated at $\alpha = .10$.

11.118 a. A straight-line model would be: $y = \beta_0 + \beta_1 x + \varepsilon$.

b. Yes, the data points are all clustered around the line.

c. From the printout, the least squares prediction line is: $\hat{y} = 184 + 1.20x$

The estimated *y*-intercept is $\hat{\beta}_o = 184$. Since 0 is not in the observed range of values of the appraised value (*x*), the y-intercept has no meaning.

The estimated slope is $\hat{\beta}_1 = 1.20$. For each additional dollar of appraised value the mean selling price is estimated to increase by 1.20 dollars.

d. From the printout, the test statistic is $t = 53.70$ and the *p*-value is $p = 0.000$. For a one-tailed test, the *p*-value will be $p/2 = 0.00/2 = 0.000$. Since the *p*-value is less than $\alpha = .01$ ($p = 0.000 < .01$), H_0 is rejected. There is sufficient evidence to indicate a positive linear relationship between appraised property value and sale price at $\alpha = .01$.

e. From the printout, r^2 = R-Sq = 97.4%. 97.4% of the total sample variability around the sample mean appraised value is explained by the linear relationship between sale price and appraised value.

From the printout, $r = .987$. Since this value is close to 1, there is a strong positive linear relationship between sale price and appraised value.

f. The prediction interval for the actual sale price when the appraised value is 400,000 is (390,085, 569,930). We are 95% confident that the actual selling price for a home appraised at $400,000 is between $390,085 and $569,930.

11.120 a. The straight-line model is $y = \beta_0 + \beta_1 x + \varepsilon$

b. The least squares prediction equation is $\hat{y} = 2.522 + 7.261x$

c. $\hat{\beta}_1 = 7.261$. For each additional day of duration, the mean number of arrests is estimated to increase by 7.261.

$\hat{\beta}_0 = 2.522$. Since $x = 0$ is not in the observed range of the duration in days, $\hat{\beta}_0$ has no interpretation other than the y-intercept.

d. From the printout, s = 16.913. We would expect most of the observations to fall within 2s or 2(16.913) or 33.826 units of the least squares prediction line.

e. From the printout, r^2 = .361. Thus, 36.1% of the total sample variability around the sample mean number of arrests is explained by the linear relationship between the number of arrests and the duration of the sit-ins.

f. To determine whether the number of arrests is positively linearly related to duration, we test:

H_0: $\beta_1 = 0$
H_a: $\beta_1 > 0$

The test statistic is t = 1.302 and the p-value is p = .284/2 = .142. Since the p-value is greater than $\alpha = .10$, H_0 is not rejected. There is insufficient evidence to indicate a positive linear relationship exists between the number of arrests and duration for $\alpha = .10$.

11.122 **SG Score**: r^2 = .002. .2% of the total sample variability around the sample mean ESLR scores is explained by the linear relationship between ESLR scores and the SG scores.

SR Score: r^2 = .099. 9.9% of the total sample variability around the sample mean ESLR scores is explained by the linear relationship between ESLR scores and the SR scores.

ER Score: r^2 = .078. 7.8% of the total sample variability around the sample mean ESLR scores is explained by the linear relationship between ESLR scores and the ER scores.

11.124 a. The straight-line model is $y = \beta_0 + \beta_1 x + \varepsilon$

b. Some preliminary calculations are:

$\sum x_i = 51.4$ $\sum y_i = 45.5$ $\sum x_i y_i = 210.49$ $\sum x_i^2 = 227.5$ $\sum y_i^2 = 214.41$

$$\text{SS}_{xy} = \sum x_i y_i - \frac{\sum x_i \sum y_i}{n} = 210.49 - \frac{51.4(45.5)}{15} = 54.5766667$$

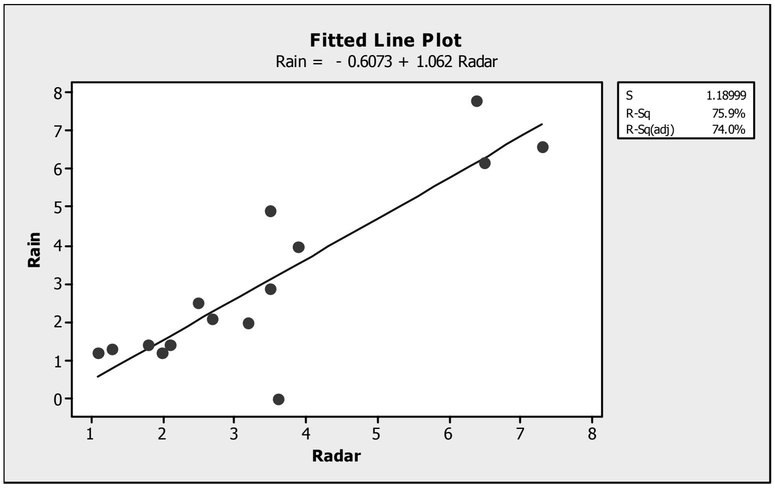
$$\text{SS}_{xx} = \sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n} = 227.5 - \frac{51.4^2}{15} = 51.3693333$$

$$\hat{\beta}_1 = \frac{\text{SS}_{xy}}{\text{SS}_{xx}} = \frac{54.5766667}{51.3693333} = 1.062436734$$

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1\overline{x} = \frac{45.5}{15} - (1.062436734)\left(\frac{51.4}{15}\right) = -.607283208$$

The least squares prediction equation is $\hat{y} = -.607 + 1.062x$

c. Using MINITAB, the graph is:



There appears to be a positive linear relationship between the two variables. As the radar rainfall increases, the rain gauge values also increase.

d. $\hat{\beta}_0 = -.607$. Since $x = 0$ is not in the observed range, $\hat{\beta}_0$ has no meaning other than the y-intercept.

$\hat{\beta}_1 = 1.062$. For each unit increase in radar rainfall, the mean rain gauge rainfall increases by an estimated 1.062.

e. Some preliminary calculations are:

$$SS_{yy} = \sum y_i^2 - \frac{\left(\sum y_i{}^2\right)}{n} = 214.41 - \frac{(45.5)^2}{15} = 76.393333$$

$$SSE = SS_{yy} - \hat{\beta}_1 SS_{xy} = 76.393333 - 1.062436734(54.5766667) = 18.40907778$$

$$s^2 = \frac{SSE}{n-2} = \frac{18.40907778}{15-2} = 1.416082906$$

$$s = \sqrt{1.416082906} = 1.18999$$

We would expect most of the observations to fall within 2*s* or 2(1.18999) or 2.37998 units of the least squares prediction line.

f. To determine if rain gauge amounts are linearly related to radar rain estimates, we test:

H_0: $\beta_1 = 0$
H_a: $\beta_1 \neq 0$

The test statistic is $t = \dfrac{\hat{\beta}_1 - 0}{s_{\hat{\beta}_1}} = \dfrac{1.062}{\dfrac{1.18999}{\sqrt{51.3693333}}} = 6.396$.

The rejection region requires $\alpha / 2 = .01/2 = .005$ in each tail of the t distribution with $df = n - 2 = 15 - 2 = 13$. From Table VI, Appendix A, $t_{.005} = 3.012$. The rejection region is $t < -3.012$ or $t > 3.012$.

Since the observed value of the test statistic falls in the rejection region ($t = 6.396 > 3.012$), H_0 is rejected. There is sufficient evidence to indicate that rain gauge amounts are linearly related to radar rain estimates at $\alpha = .01$.

g. For confidence level .99, $\alpha = 1 - .99 = .01$ and $\alpha / 2 = .01/2 = .005$. From Table VI, Appendix A with $df = n - 2 = 15 - 2 = 13$, $t_{.005} = 3.012$.

The confidence interval is:

$$\hat{\beta}_1 \pm t_{.005} s_{\hat{\beta}_1} \Rightarrow 1.062 \pm 3.012\left(\frac{1.18999}{\sqrt{51.3693333}}\right) \Rightarrow 1.062 \pm .5000 \Rightarrow (.562, 1.562)$$

We are 99% confident that for each unit increase in radar rain estimate, the mean value of rain gauge amount is estimated to increase from .562 to 1.562 units.

h. The straight-line model is $y = \beta_0 + \beta_1 x + \varepsilon$

Some preliminary calculations are:

$\sum x_i = 46.7$ $\quad \sum y_i = 45.5$ $\quad \sum x_i y_i = 207.89$ $\quad \sum x_i^2 = 210.21$ $\quad \sum y_i^2 = 214.41$

$$\text{SS}_{xy} = \sum x_i y_i - \frac{\sum x_i \sum y_i}{n} = 207.89 - \frac{46.7(45.5)}{15} = 66.2333333$$

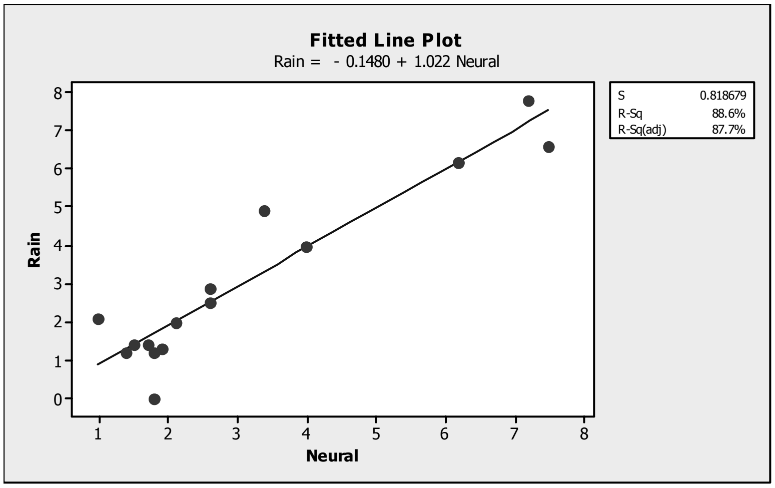
$$\text{SS}_{xx} = \sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n} = 210.21 - \frac{46.7^2}{15} = 64.8173333$$

$$\hat{\beta}_1 = \frac{\text{SS}_{xy}}{\text{SS}_{xx}} = \frac{66.2333333}{64.8173333} = 1.021846008$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = \frac{45.5}{15} - (1.021846008)\left(\frac{46.7}{15}\right) = -.148013904$$

The least squares prediction equation is $\hat{y} = -.148 + 1.022x$

Using MINITAB, the graph is:



There appears to be a positive linear relationship between the two variables. As the neural network rainfall increases, the rain gauge values also increase.

$\hat{\beta}_0 = -.148$. Since $x = 0$ is not in the observed range, $\hat{\beta}_0$ has no meaning other than the y-intercept.

$\hat{\beta}_1 = 1.022$. For each unit increase in neural network rainfall, the mean rain gauge rainfall increases by an estimated 1.022.

Some preliminary calculations are:

$$SS_{yy} = \sum y_i^2 - \frac{\left(\sum y_i{}^2\right)}{n} = 214.41 - \frac{(45.5)^2}{15} = 76.393333$$

$$SSE = SS_{yy} - \hat{\beta}_1 SS_{xy} = 76.393333 - 1.021846008(66.2333333) = 8.713065771$$

$$s^2 = \frac{SSE}{n-2} = \frac{8.713065771}{15-2} = .670235828$$

$$s = \sqrt{.670235828} = .81868$$

We would expect most of the observations to fall within $2s$ or 2(.81868) or 1.63736 units of the least squares prediction line.

f. To determine if rain gauge amounts are linearly related to radar rain estimates, we test:

H_0: $\beta_1 = 0$
H_a: $\beta_1 \neq 0$

The test statistic is $t = \frac{\hat{\beta}_1 - 0}{s_{\hat{\beta}_1}} = \frac{1.022}{\frac{.81868}{\sqrt{64.8173333}}} = 10.050$.

The rejection region requires $\alpha / 2 = .01/2 = .005$ in each tail of the t distribution with $df = n - 2 = 15 - 2 = 13$. From Table VI, Appendix A, $t_{.005} = 3.012$. The rejection region is $t < -3.012$ or $t > 3.012$.

Since the observed value of the test statistic falls in the rejection region ($t = 10.050 > 3.012$), H_0 is rejected. There is sufficient evidence to indicate that rain gauge amounts are linearly related to neural network rain estimates at $\alpha = .01$.

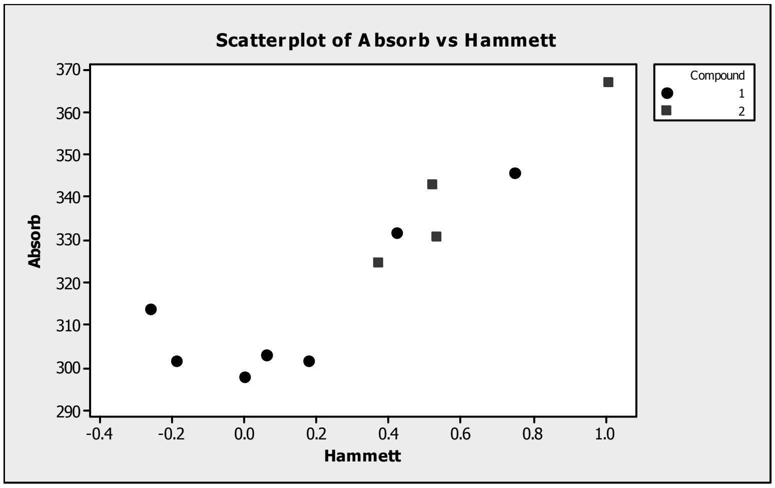
For confidence level .99, $\alpha = 1 - .99 = .01$ and $\alpha / 2 = .01/2 = .005$. From Table VI, Appendix A with $df = n - 2 = 15 - 2 = 13$, $t_{.005} = 3.012$.

The confidence interval is:

$$\hat{\beta}_1 \pm t_{.005} s_{\hat{\beta}_1} \Rightarrow 1.022 \pm 3.012\left(\frac{.81868}{\sqrt{64.8173333}}\right) \Rightarrow 1.022 \pm .306 \Rightarrow (.716, 1.328)$$

We are 99% confident that for each unit increase in neural network rain estimate, the mean value of rain gauge amount is estimated to increase from .716 to 1.328 units.

11.126 a. Using MINITAB, a scattergram of the data is:



It appears that the relationship between the Hammett constant and the maximum absorption is fairly similar for both compounds. For both compounds, there appears to be a positive linear relationship between the Hammett constant and the maximum absorption.

b. Using MINITAB, the results for compound 1 are:

Regression Analysis: Absorb1 versus Hammett1

```
The regression equation is
Absorb1 = 308 + 41.7 Hammett1


Predictor     Coef  SE Coef      T      P
Constant   308.137    4.896  62.94  0.000
Hammett1     41.71    13.82   3.02  0.029


S = 11.9432   R-Sq = 64.6%   R-Sq(adj) = 57.5%


Analysis of Variance

Source          DF      SS      MS     F      P
Regression       1  1299.7  1299.7  9.11  0.029
Residual Error   5   713.2   142.6
Total            6  2012.9
```

The least squares prediction line is $\hat{y} = 308.137 + 41.71x$.

c. To determine if the model is adequate for compound 1, we test:

H_0: $\beta_1 = 0$
H_a: $\beta_1 \neq 0$

From the printout, the test statistic is $t = 3.02$ and the p-value is $p = .029$. Since the p-value is not less than $\alpha = .01$, H_0 is not rejected. There is insufficient evidence to indicate the model is adequate for compound 1 at $\alpha = .01$.

d. Using MINITAB, the results for compound 2 are:

Regression Analysis: Absorb2 versus Hammett2

```
The regression equation is
Absorb2 = 303 + 64.1 Hammett2


Predictor     Coef  SE Coef      T      P
Constant   302.588    8.732  34.65  0.001
Hammett2     64.05    13.36   4.79  0.041


S = 6.43656   R-Sq = 92.0%   R-Sq(adj) = 88.0%


Analysis of Variance

Source          DF       SS      MS      F      P
Regression       1   952.14  952.14  22.98  0.041
Residual Error   2    82.86   41.43
Total            3  1035.00
```

The least squares prediction line is $\hat{y} = 302.588 + 64.05x$.

To determine if the model is adequate for compound 2, we test:

H_0: $\beta_1 = 0$
H_a: $\beta_1 \neq 0$

From the printout, the test statistic is $t = 4.79$ and the p-value is $p = .041$. Since the p-value is not less than $\alpha = .01$, H_0 is not rejected. There is insufficient evidence to indicate the model is adequate for compound 2 at $\alpha = .01$.

11.128 a. Yes. For the men, as the year increases, the winning time tends to decrease. The straight-line model is $y = \beta_0 + \beta_1 x + \varepsilon$. We would expect the slope to be negative.

b. Yes. For the women, as the year increases, the winning time tends to decrease. The straight-line model is $y = \beta_0 + \beta_1 x + \varepsilon$. We would expect the slope to be negative.

c. Since the slope of the women's line is steeper that that for the men, the slope of the women's line will be greater in absolute value.

d. No. The gathered data is from 1880 to 2000. Using this data to predict the time for the year 2020 would be very risky. We have no idea what the relationship between time and year will be outside the observed range. Thus, we would not recommend using this model.

e. The women's model is more likely to have the smaller estimate of σ. The women's observed points are closer to the women's line than the men's observed points are to the men's line.

11.130 a. A straight line model is $y = \beta_0 + \beta_1 x + \varepsilon$.

b. The researcher hypothesized that therapists with more years of formal dance training will report a higher perceived success rate in cotherapy relationships. This indicates that $\beta_1 > 0$.

c. $r = -.26$. Because this value is fairly close to 0, there is a weak negative linear relationship between years of formal training and reported success rate.

d. To determine if there is a positive linear relationship between years of formal training and reported success rate, we test:

H_0: $\beta_1 = 0$
H_a: $\beta_1 > 0$

The test statistic is $t = \dfrac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \dfrac{-.26\sqrt{136-2}}{\sqrt{1-(-.26^2)}} = -3.12$

The rejection region requires $\alpha = .05$ in the upper tail of the t distribution with df = $n - 2$ = 136 − 2 = 134. From Table VI, Appendix A, $t_{.05} \approx 1.658$. The rejection region is $t > 1.658$.

Since the observed value of the test statistic does not fall in the rejection region ($t = -8.66 \not> 1.658$), H_0 is not rejected. There is insufficient evidence to indicate that there is a positive linear relationship between years of formal training and perceived success rates at $\alpha = .05$.

11.132 a. The equation for the straight-line model relating duration to frequency is $y = \beta_0 + \beta_1 x + \varepsilon$.

b. Some preliminary calculations are:

$$\sum x_i = 394 \qquad \sum y_i = 1287 \qquad \sum x_i y_i = 30,535 \qquad \sum x_i^2 = 28,438 \qquad \sum y_i^2 = 203,651$$

$$\bar{y} = \frac{\sum y}{n} = \frac{1,287}{11} = 117 \qquad \bar{x} = \frac{\sum x}{n} = \frac{394}{11} = 35.818$$

$$\text{SS}_{xy} = \sum xy - \frac{\sum x \sum y}{n} = 30,535 - \frac{394(1,287)}{11} = -15,563$$

$$\text{SS}_{xx} = \sum x^2 - \frac{\left(\sum x\right)^2}{n} = 28,438 - \frac{394^2}{11} = 14,325.63636$$

$$\hat{\beta}_1 = \frac{\text{SS}_{xy}}{\text{SS}_{xx}} = \frac{-15,563}{14,325.63636} = -1.086374079$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = \frac{1,287}{11} - (-1.086374079)\left(\frac{394}{11}\right) = 155.9119443$$

The least squares prediction equation is $\hat{y} = 155.912 - 1.086x$.

c. Some preliminary calculations are:

$$\text{SS}_{yy} = \sum y^2 - \frac{\left(\sum y\right)^2}{n} = 203,651 - \frac{1,287^2}{11} = 53,072$$

$$\text{SSE} = \text{SS}_{yy} - \hat{\beta}_1\left(\text{SS}_{xy}\right) = 53,072 - (-1.086374079)(-15,563) = 36,164.76021$$

$$s^2 = \frac{\text{SSE}}{n-2} = \frac{36,164.76021}{11-2} = 4,018.30669$$

$$s = \sqrt{s^2} = \sqrt{4,018.30669} = 63.39011508$$

To determine if there is a linear relationship between duration and frequency, we test:

H_0: $\beta_1 = 0$
H_a: $\beta_1 \neq 0$

The test statistic is $t = \dfrac{\hat{\beta}_1 - 0}{s_{\hat{\beta}}} = \dfrac{\hat{\beta}_1 - 0}{s / \sqrt{\text{SS}_{\text{xx}}}} = \dfrac{-1.086 - 0}{63.3901 / \sqrt{14{,}325.63636}} = -2.051$

The rejection region requires $\alpha / 2 = .05 / 2 = .025$ in each tail of the t distribution. From Table VI, Appendix A, with df = $n - 2 = 11 - 2 = 9$, $t_{.025} = 2.262$. The rejection region is $t < -2.262$ or $t > 2.262$.

Since the observed test statistic does not fall in the rejection region ($t = -2.051 \not< -2.262$), H_0 is not rejected. There is insufficient evidence to indicate that duration and frequency are linearly related at $\alpha = .05$.

d. For $x = 25$, the predicted duration is $\hat{y} = 155.912 = 1.086(25) = 128.762$.

For confidence coefficient .95, $\alpha = 1 - .95 = .05$ and $\alpha / 2 = .05 / 2 = .025$. From Table VI, Appendix A, with df = $n - 2 = 11 - 2 = 9$, $t_{.025} = 2.262$. The 95% prediction interval is:

$$\hat{y} \pm t_{\alpha/2} s \sqrt{1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{\text{SS}_{xx}}} \Rightarrow 128.762 \pm 2.262(63.3901)\sqrt{1 + \frac{1}{11} + \frac{(25 - 35.818)^2}{14{,}325.63636}}$$

$$\Rightarrow 128.762 \pm 150.324 \Rightarrow (-21.562,\ \ 279.086)$$

We are 95% confident that the actual duration of a person who participates 25 times a year is between –21.562 and 279.086 days. Since the duration cannot be negative, the actual duration will be between 0 and 279.086.

11.134 Using MINITAB, the regression analysis is:

```
The regression equation is
y = 5.35 + 0.530 x

Predictor         Coef       StDev          T        P
Constant        5.3480      0.1635      32.71    0.000
x               0.5299      0.9254       0.57    0.575

S = 0.4115      R-Sq = 2.0%      R-Sq(adj) = 0.0%

Analysis of Variance

Source            DF          SS          MS         F        P
Regression         1      0.0555      0.0555      0.33    0.575
Residual Error    16      2.7091      0.1693
Total             17      2.7646
```

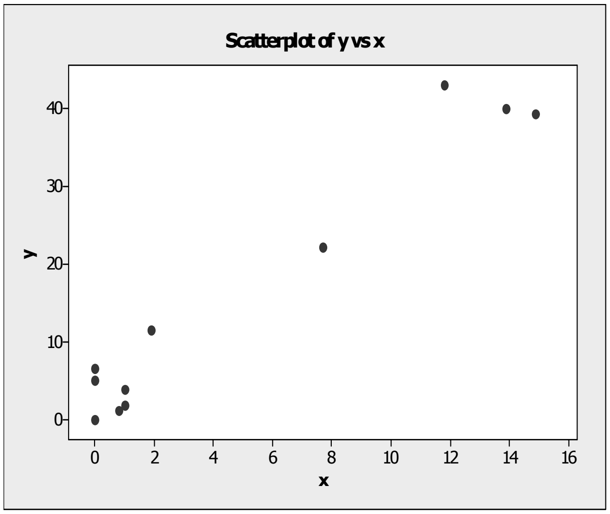
The fitted regression line is $\hat{y} = 5.35 + .530x$. Thus, if the takeoff error was reduced by .1 meters, we would estimate that the best jumping distance would change by .530(–.1) = –.053 meters.

Generally, we would expect that the smaller the takeoff error, the longer the jump. From this data, the coefficient corresponding to the takeoff error is positive, indicating that there is a positive linear relationship between jumping distance and takeoff error. However, examining the output indicates that there is insufficient evidence of a linear relationship between jumping distance and takeoff error (t = .57 and p = .575). In addition, the R-square is very small (R^2 = 2.0%), again indicating that takeoff error is not linearly related to jumping distance.

11.136 Answers may vary. Possible answer:

The scaffold-drop survey provides the most accurate estimate of spall rate in a given wall segment. However, the drop areas were not selected at random from the entire complex; rather, drops were made at areas with high spall concentrations. Therefore, if the photo spall rates could be shown to be related to drop spall rates, then the 83 photo spall rates could be used to predict what the drop spall rates would be.

Construct a scattergram for the data.



The scattergram shows a positive linear relationship between the photo spall rate (x) and the drop spall rate (y).

Find the prediction equation for drop spall rate. The MINITAB output shows the results of the analysis.

```
The regression equation is
drop = 2.55 + 2.76 photo

Predictor        Coef       StDev          T        P
Constant        2.548       1.637       1.56    0.154
photo          2.7599      0.2180      12.66    0.000

S = 4.164       R-Sq = 94.7%     R-Sq(adj) = 94.1%

Analysis of Variance

Source            DF          SS          MS         F         P
Regression         1      2777.5      2777.5    160.23     0.000
Residual Error     9       156.0        17.3
Total             10      2933.5
```

The least squares prediction line is $\hat{y} = 2.55 + 2.76x$.

Conduct a formal statistical hypothesis test to determine if the photo spall rates contribute information for the prediction of drop spall rates.

H_0: $\beta_1 = 0$
H_a: $\beta_1 \neq 0$

The test statistic is $t = 12.66$ and the p-value is $p = .000$.

Since the p-value is so small ($p = .000$), H_0 is rejected for any reasonable value of α. There is sufficient evidence to indicate that photo spall rates contribute information for the prediction of drop spall rates for any reasonable value of α.

One could now use the 83 photos spall rates to predict values for 83 drop spall rates. Then use this information to estimate the true spall rate at a given wall segment and estimate to total spall damage.