Chapter 10 Regression and Correlation

Section 10.1

1. The points seem close to a straight line, so there is moderate or low linear correlation.

2. No straight line is realistically a good fit, so there is no linear correlation.

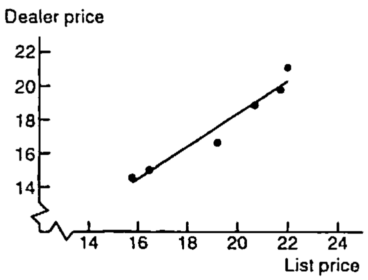
3. The points seem very close to a straight line, so there is high linear correlation.

4. The points seem close to a straight line, so there is moderate or low linear correlation.

5. The points seem very close to a straight line, so there is high linear correlation.

6. No straight line is realistically a good fit, so there is no linear correlation.

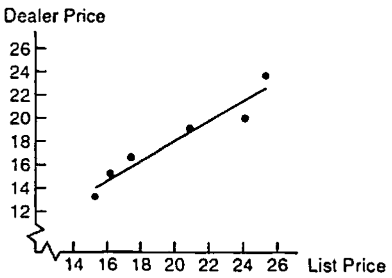
7. (a) List and Dealer Price Pontiac Grand Am (thousands of dollars)



   (b) Draw the line you think fits best. (Method to find equation is in Section 10.2.)

   (c) Since the points are very close to a straight line, the correlation is high.

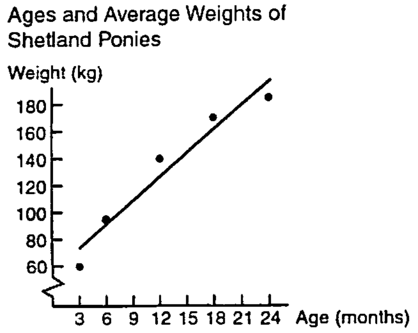
8. (a) List and Dealer Price Dodge Ram (thousands of dollars)



   (b) Draw the line you think fits best. (Method to find equation is in Section 10.2.)

   (c) Since the points are very close to a straight line, the correlation is high.

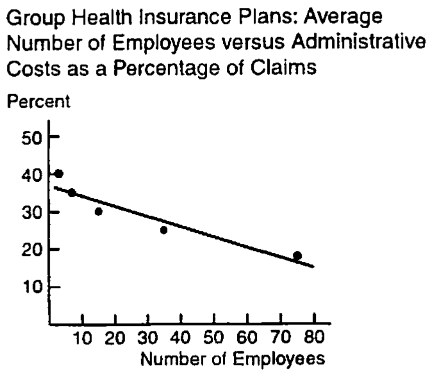
9. (a)



(b) Draw the line you think fits best. (Method to find equation is in Section 10.2.)

(c) Since the points are very close to a straight line, the correlation is high.

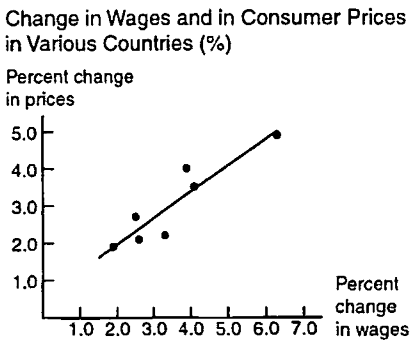
10. (a)



(b) Draw the line you think fits best. (Method to find equation is in Section 10.2.)

(c) Since the points are fairly close to a straight line, the correlation is moderate.

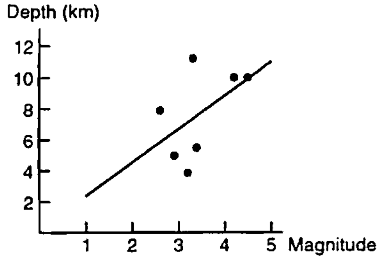
11. (a)



(b) Draw the line you think fits best. (Method to find equation is in Section 10.2.)

(c) Since the points are fairly close to a straight line, the correlation is moderate.

12. (a) Magnitude (Richter Scale) and Depth (km) of Earthquakes

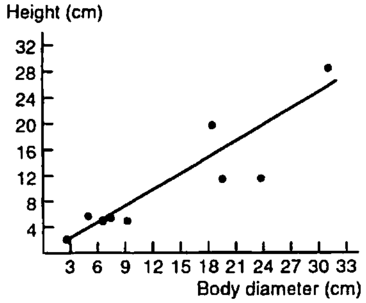


(b) Draw the line you think fits best. (Method to find equation is in Section 10.2.)

(c) Since the points are not close to a straight line. the correlation is low.

Note: One possible reason why there appears to be little, if any, linear relationship is that the Richter scale is logarithmic. An increase of 1 on the Richter scale represents a 60-fold increase in energy.

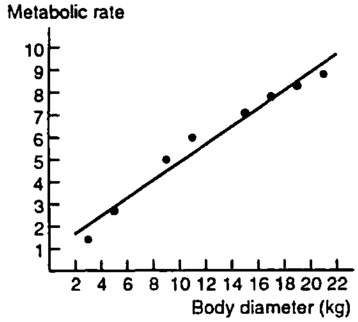
13. (a) Body Diameter and Weight of Prehistoric Pottery



(b) Draw the line you think fits best. (Method to find equation is in Section 10.2.)

(c) Since the points are fairly close to a straight line, the correlation is moderate.

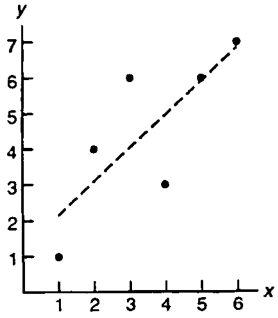
14. (a) Body Weight and Metabolic Rate of Children



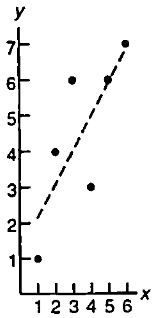
(b) Draw the line you think fits best. (Method to find equation is in Section 10.2.)

(c) Since the points are very close to a straight line, the correlation is high.

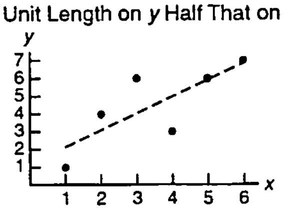
15. (a) Unit Length on *y* Same as That on *x*



(b) Unit Length on *y* Twice That on *x*



(c) Unit Length on *y* Half That on *x*

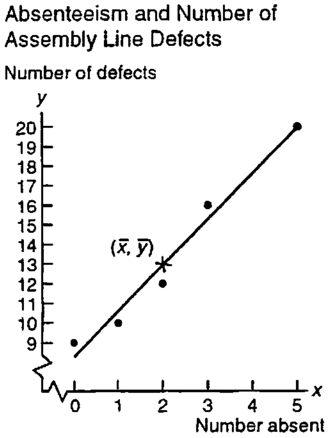


(d) Draw the lines you think best fit the data points.
Stretching the scale on the *y*-axis makes the line appear steeper. Shrinking the scale on the *y*-axis makes the line appear flatter. The slope of the line does not change. Only the appearance (visual impression) of slope changes as the scale of the *y*-axis changes.

Section 10.2

Note: In this section and the next two, answers may vary slightly, depending on how many significant digits are used throughout the calculations.

1. (a)



(b) $\bar{x} = \frac{\sum x}{n} = \frac{11}{5} = 2.2$

$\bar{y} = \frac{\sum y}{n} = \frac{67}{5} = 13.4$

$b = \frac{SS_{xy}}{SS_x} = \frac{34.6}{14.8} = 2.3378$

$a = \bar{y} - b\bar{x} = 13.4 - 2.3378(2.2) = 8.26$

$y = a + bx$ or $y = 8.26 + 2.338x$

(c) See figure of part (a).

(d) $$S_e = \sqrt{\frac{SS_y - bSS_{xy}}{n-2}} = \sqrt{\frac{83.2 - 2.3378(34.6)}{5-2}} = 0.878$$

(e) Use $x = 4$.

$y_p = 8.26 + 2.338(4) = 17.6$ defects

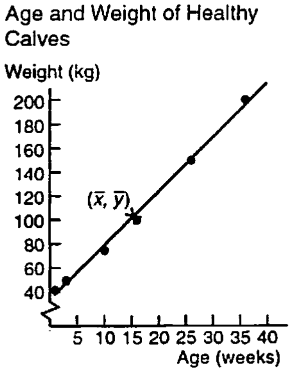
(f) $t_{0.95,3d.f.} = 3.182$

$$E = t_c S_e \sqrt{1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{SS_x}} = 3.182(0.878)\sqrt{1 + \frac{1}{5} + \frac{(4-2.2)^2}{14.8}} = 3.3$$

A 95% confidence interval is

$$y_p - E \le y \le y_p + E$$
$$17.6 - 3.3 \le y \le 17.6 + 3.3$$
$$14.3 \le y \le 20.9 \text{ defects}$$

2. (a)



(b) $\bar{x} = \frac{\sum x}{n} = \frac{92}{6} = 15.33$

$\bar{y} = \frac{\sum y}{n} = \frac{617}{6} = 102.83$

$b = \frac{SS_{xy}}{SS_x} = \frac{4181.3}{927.3} = 4.509$

$a = \bar{y} - b\bar{x} = 102.83 - 4.509(15.33) = 33.70$

$y = a + bx$ or $y = 33.70 + 4.51x$

(c) See figure of part (a).

(d) $S_e = \sqrt{\frac{SS_y - bSS_{xy}}{n-2}}$

$= \sqrt{\frac{18940.8 - 4.509(4181.3)}{6-2}}$

$= 4.67$

(e) Use $x = 12$.

$y_p = 33.70 + 4.51(12) = 87.8$ kg

(f) $t_{0.90, 4d.f.} = 2.132$

$$E = t_c S_e \sqrt{1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{SS_x}}$$

$$= 2.132(4.67)\sqrt{1 + \frac{1}{6} + \frac{(12 - 15.3)^2}{927.3}}$$

$$= 10.81$$

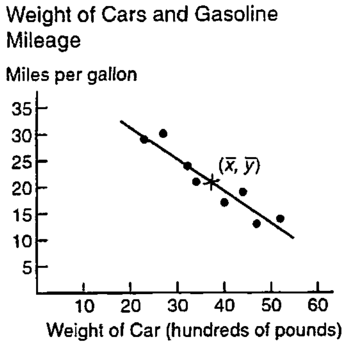
A 90% confidence interval for y is

$$y_p - E \le y \le y_p + E$$

$$87.8 - 10.81 \le y \le 87.8 + 10.81$$

$$76.99 \le y \le 98.61 \text{ kg}$$

3. (a)



(b) $\bar{x} = \frac{\sum x}{n} = \frac{299}{8} = 37.375$

$\bar{y} = \frac{\sum y}{n} = \frac{167}{8} = 20.875$

$b = \frac{SS_{xy}}{SS_x} = \frac{-427.625}{711.875} = -0.6007$

$a = \bar{y} - b\bar{x} = 20.875 - (-0.6007)(37.375) = 43.3263$

$y = a + bx$ or $y = 43.3263 - 0.6007x$

(c) See figure of part (a).

(d) $S_e = \sqrt{\frac{SS_y - bSS_{xy}}{n-2}}$

$= \sqrt{\frac{286.875 - (-0.6007)(-427.625)}{8-2}}$

$= 2.2361$

(e) Use $x = 38$.

$y_p = 43.3263 - 0.6007(38) = 20.5$ mpg

(f) $t_{0.80, 6 d.f.} = 1.440$

$E = t_c S_e \sqrt{1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{SS_x}}$

$= 1.44(2.2361)\sqrt{1 + \frac{1}{8} + \frac{(38-37.375)^2}{711.875}}$

$= 3.4$

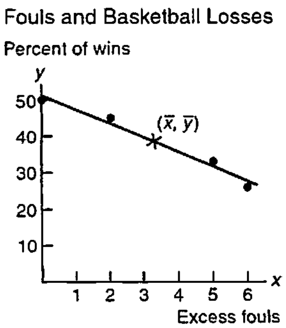
A 80% confidence interval for y is

$y_p - E \le y \le y_p + E$

$20.5 - 3.4 \le y \le 20.4 + 3.4$

$17.1 \le y \le 23.8$ mpg

4. (a)



(b) $\bar{x} = \frac{\sum x}{n} = \frac{13}{4} = 3.25$

$\bar{y} = \frac{\sum y}{n} = \frac{154}{4} = 38.5$

$b = \frac{SS_{xy}}{SS_x} = \frac{-89.5}{22.75} = -3.934$

$a = \bar{y} - b\bar{x} = 38.5 - (-3.934)(3.25) = 51.29$

$y = a + bx$ or $y = 51.29 - 3.934x$

(c) See figure of part (a).

(d) $S_e = \sqrt{\frac{SS_y - bSS_{xy}}{n-2}} = \sqrt{\frac{361 - (-3.934)(-89.5)}{4-2}} = 2.11$

(e) Use $x = 4$.

$y_p = 51.29 - 3.934(4) = 35.55\%$

(f) $t_{0.80, 2d.f.} = 1.886$

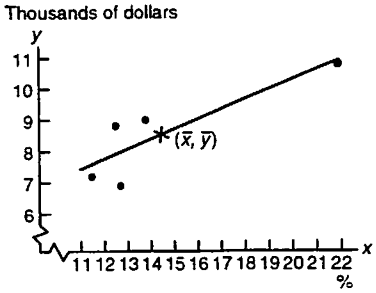
$$E = t_c S_e \sqrt{1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{SS_x}}$$

$$= 1.886(2.11)\sqrt{1 + \frac{1}{4} + \frac{(4-3.25)^2}{22.75}} = 4.49$$

A 80% confidence interval for y is

$$y_p - E \le y \le y_p + E$$
$$35.55 - 4.49 \le y \le 35.55 + 4.49$$
$$31.06 \le y \le 40.04\%$$

5. (a) Education and Income in Small Cities



(b) $\bar{x} = \frac{\sum x}{n} = \frac{72.4}{5} = 14.48$

$\bar{y} = \frac{\sum y}{n} = \frac{42.7}{5} = 8.54$

$b = \frac{SS_{xy}}{SS_x} = \frac{22.854}{71.448} = 0.320$

$a = \bar{y} - b\bar{x} = 8.54 - 0.320(14.48) = 3.91$

$y = a + bx$ or $y = 3.91 + 0.320x$

(c) See figure of part (a).

Note that the regression line would be much steeper if (21.9, 10.8) were eliminated from the data set [which would also affect $(\bar{x}, \bar{y})$]. Not all outliers (this point is an outlier in both x (probably) and y) have this effect; however, when the parameter estimates a and b depend heavily on a particular observation, as is the case here, the point is called "influential," and conclusions drawn are shaky at best when influential observations remain in the data. For further information, refer to a more advanced textbook such as Applied Regression Analysis by Draper and Smith.

(d) $$S_e = \sqrt{\frac{SS_y - bSS_{xy}}{n-2}} = \sqrt{\frac{9.872 - 0.320(22.854)}{5-2}} = 0.924$$

(e) Use $x = 20$.

$y_p = 3.91 + 0.320(20) = 10.31$ i.e., 10.31 thousand dollars

(f) $t_{0.80,3.d.f} = 1.638$

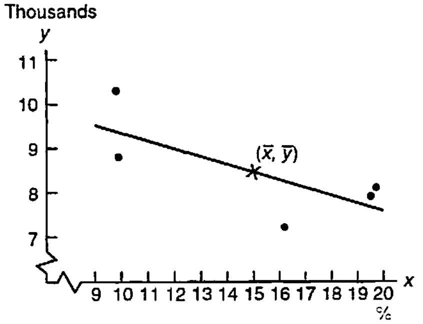
$$E = t_c S_e \sqrt{1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{SS_x}} = 1.638(0.924)\sqrt{1 + \frac{1}{5} + \frac{(20-14.48)^2}{71.448}} = 1.93$$

A 80% confidence interval for y is

$$y_p - E \le y \le y_p + E$$
$$10.31 - 1.93 \le y \le 10.31 + 1.93$$
$$8.4 \le y \le 12.2$$

or 8.4 thousand dollars to 12.2 thousand dollars

6. (a) Percentage of 16 to 19-Year-Olds Not in School and per Capita Income (thousands of dollars)



(b) $$\bar{x} = \frac{\Sigma x}{n} = \frac{75.1}{5} = 15.02$$

$$\bar{y} = \frac{\Sigma y}{n} = \frac{42.3}{5} = 8.46$$

$$b = \frac{SS_{xy}}{SS_x} = \frac{-17.026}{96.828} = -0.1758$$

$a = \bar{y} - b\bar{x} = 8.46 - (-0.1758)15.02 = 11.10$

$y = a + bx$ or $y = 11.10 - 0.176x$

(c) See figure of part (a).

(d) $S_e = \sqrt{\frac{SS_y - bSS_{xy}}{n-2}}$

$= \sqrt{\frac{5.532 - (-0.1758)(-17.026)}{5-2}}$

$= 0.9199$

(e) Use $x = 17$.

$y_p = 11.10 - 0.176(17) = 8.11$ thousand dollars

(f) $t_{0.75,3\,d.f.} = 1.423$

$E = t_c S_e \sqrt{1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{SS_x}}$

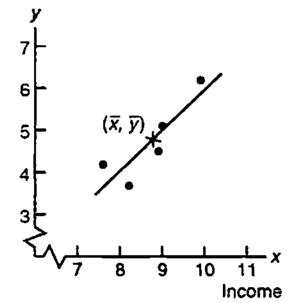
$= 1.423(0.9199)\sqrt{1 + \frac{1}{5} + \frac{(17-15.02)^2}{96.828}}$

$= 1.46$

A 75% confidence interval for y is

$$y_p - E \le y \le y_p + E$$
$$8.11 - 1.46 \le y \le 8.11 + 1.46$$
$$6.65 \le y \le 9.57 \quad \text{thousand dollars}$$

7. (a) Per Capita Income and Per Capita Retail Sales in Small Cities (thousands of dollars)



(b) $\bar{x} = \frac{\sum x}{n} = \frac{43.6}{5} = 8.72$

$\bar{y} = \frac{\sum y}{n} = \frac{23.7}{5} = 4.74$

$b = \frac{SS_{xy}}{SS_x} = \frac{2.926}{3.028} = 0.966314$

$a = \bar{y} - b\bar{x} = 4.74 - (0.966314)(8.72) = -3.69$

$y = a + bx$ or $y = -3.69 + 0.966x$

(c) See figure of part (a).

(d) $S_e = \sqrt{\frac{SS_y - bSS_{xy}}{n-2}}$

$= \sqrt{\frac{3.692 - 0.966314(2.926)}{5-2}}$

$= 0.5368$

(e) Use $x = 9.5$.

$y_p = -3.69 + 0.966(9.5) = 5.49$ thousand dollars

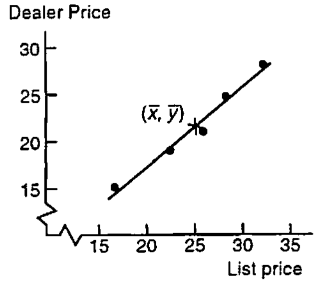
(f) $t_{0.80,3d.f.} = 1.638$

$$E = t_c S_e \sqrt{1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{SS_x}}$$
$$= 1.638(0.5368)\sqrt{1 + \frac{1}{5} + \frac{(9.5-8.72)^2}{3.028}}$$
$$= 1.04$$

A 80% confidence interval for y is

$$y_p - E \le y \le y_p + E$$
$$5.49 - 1.04 \le y \le 5.49 + 1.04$$
$$4.45 \le y \le 6.53 \quad \text{thousand dollars}$$

8. (a) List and Dealer Price for GMC Sierra (thousands of dollars)



(b) $\bar{x} = \frac{\sum x}{n} = \frac{125.6}{5} = 25.12$

$\bar{y} = \frac{\sum y}{n} = \frac{108.1}{5} = 21.62$

$b = \frac{SS_{xy}}{SS_x} = \frac{117.978}{138.608} = 0.85116$

$a = \bar{y} - b\bar{x} = 21.62 - 0.85116(25.12) = 0.239$

$y = a + bx$ or $y = 0.239 + 0.851x$

(c) See figure of part (a).

(d) $$S_e = \sqrt{\frac{SS_y - bSS_{xy}}{n-2}}$$
$$= \sqrt{\frac{103.168 - (0.85116)(117.978)}{5-2}}$$
$$= 0.957$$

(e) Use $x = 23.5$.

$y_p = 0.239 + 0.851(23.5) = 20.24$ thousand dollars

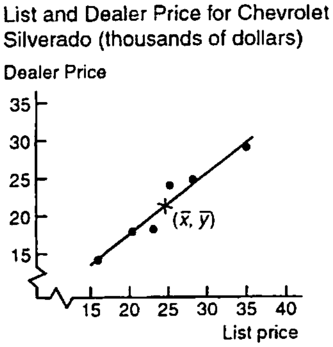
(f) $t_{0.75,3d.f.} = 1.423$

$$E = t_c S_e \sqrt{1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{SS_x}}$$
$$= 1.423(0.957)\sqrt{1 + \frac{1}{5} + \frac{(23.5-25.12)^2}{138.608}}$$
$$= 1.504$$

A 75% confidence interval for y is

$$y_p - E \le y \le y_p + E$$
$$20.24 - 1.504 \le y \le 20.24 + 1.504$$
$$18.74 \le y \le 21.74 \quad \text{thousand dollars}$$

9. (a)



(b) $\bar{x} = \frac{\sum x}{n} = \frac{148.4}{6} = 24.73$

$\bar{y} = \frac{\sum y}{n} = \frac{128.1}{6} = 21.35$

$b = \frac{SS_{xy}}{SS_x} = \frac{174.99}{217.053} = 0.8062$

$a = \bar{y} - b\bar{x} = 21.35 - 0.8062(24.73) = 1.41$

$y = a + bx$ or $y = 1.41 + 0.806x$

(c) See figure of part (a).

(d) $S_e = \sqrt{\frac{SS_y - bSS_{xy}}{n-2}}$

$= \sqrt{\frac{150.395 - 0.8062(174.99)}{6-2}}$

$= 1.526$

(e) Use $x = 22.9$.

$y_p = 1.41 + 0.806(22.9) = 19.87$ thousand dollars

(f) $t_{0.80, 4 d.f.} = 1.533$

$E = t_c S_e \sqrt{1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{SS_x}}$

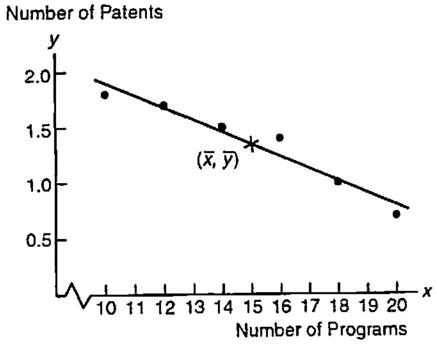
$= 1.533(1.526)\sqrt{1 + \frac{1}{6} + \frac{(22.9 - 24.73)^2}{217.053}}$

$= 2.54$

A 80% confidence interval for y is

$$y_p - E \le y \le y_p + E$$
$$19.87 - 2.54 \le y \le 19.87 + 2.54$$
$$17.33 \le y \le 22.41 \text{ thousand dollars}$$

10. (a) Number of Research Programs and Mean Number of Patents per Program



(b) $\bar{x} = \frac{\sum x}{n} = \frac{90}{6} = 15.0$

$\bar{y} = \frac{\sum y}{n} = \frac{8.1}{6} = 1.35$

$b = \frac{SS_{xy}}{SS_x} = \frac{-7.7}{70} = -0.1100$

$a = \bar{y} - b\bar{x} = 1.35 - (-0.1100)(15.0) = 3.0$

$y = a + bx$ or $y = 3.0 - 0.11x$

(c) See figure of part (a).

(d) $S_e = \sqrt{\frac{SS_y - bSS_{xy}}{n-2}}$

$= \sqrt{\frac{0.895 - (-0.11)(-7.7)}{6-2}}$

$= 0.1095$

(e) Use $x = 15$.

$y_p = 3.0 - 0.11(15) = 1.35$ patents

(f) $t_{0.85,4d.f.} = 1.778$

$$E = t_c S_e \sqrt{1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{SS_x}}$$

$$= 1.778(0.1095)\sqrt{1 + \frac{1}{6} + \frac{(15-15)^2}{70}}$$

$$= 0.21$$

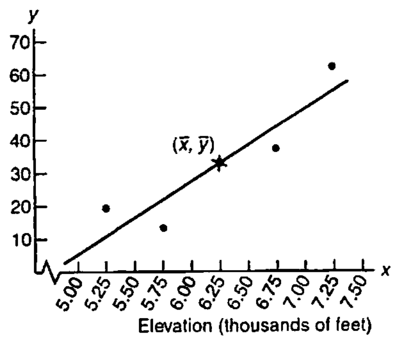
A 85% confidence interval for y is

$$y_p - E \le y \le y_p + E$$

$$1.35 - 0.21 \le y \le 1.35 + 0.21$$

$$1.14 \le y \le 1.56 \text{ patents}$$

11. (a) Cultural Affiliation and Elevation of Achaeological Sites



(b) $\bar{x} = \frac{\sum x}{n} = \frac{31.25}{5} = 6.25$

$\bar{y} = \frac{\sum y}{n} = \frac{164}{5} = 32.8$

$b = \frac{SS_{xy}}{SS_x} = \frac{55}{2.5} = 22.0$

$a = \bar{y} - b\bar{x} = 32.8 - 22.0(6.25) = -104.7$

$y = a + bx$ or $y = -104.7 + 22.0x$

(c) See figure of part (a).

(d) $S_e = \sqrt{\frac{SS_y - bSS_{xy}}{n-2}}$

$= \sqrt{\frac{1452.8 - (22)(55)}{5-2}}$

$= 8.996$

(e) Use $x = 6.5$.

$y_p = -104.7 + 22.0(6.5) = 38.3$ percent

(f) $t_{0.75,3d.f.} = 1.423$

$$E = t_c S_e \sqrt{1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{SS_x}}$$

$$= 1.423(8.996)\sqrt{1 + \frac{1}{5} + \frac{(6.5-6.25)^2}{2.5}}$$

$$= 14.2$$

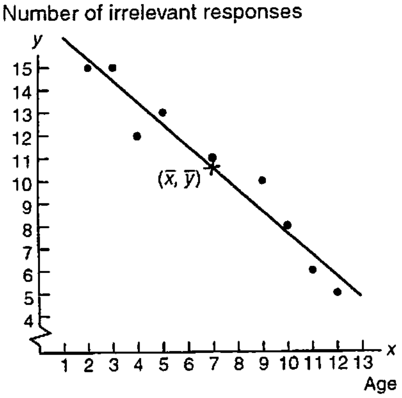
A 75% confidence interval for y is

$$y_p - E \le y \le y_p + E$$

$$38.3 - 14.2 \le y \le 38.3 + 14.2$$

$$24.1 \le y \le 52.5 \quad \text{percent}$$

12. (a) Ages of Children and Their Responses to Questions



(b) $\bar{x} = \frac{\sum x}{n} = \frac{63}{9} = 7.0$

$\bar{y} = \frac{\sum y}{n} = \frac{95}{9} = 10.56$

$b = \frac{SS_{xy}}{SS_x} = \frac{-104}{108} = -0.96296$

$a = \bar{y} - b\bar{x} = 10.56 - (-0.96296)(7.0) = 17.30$

$y = a + bx$ or $y = 17.30 - 0.963x$

(c) See figure of part (a).

(d) $S_e = \sqrt{\frac{SS_y - bSS_{xy}}{n-2}}$

$= \sqrt{\frac{106.\bar{2} - (-0.96296)(-104)}{9-2}}$

$= 0.93152$

(e) Use $x = 9.5$.

$y_p = 17.30 - 0.963(9.5) = 8.15$ irrelevant responses

(f) $t_{0.99, 7 d.f.} = 3.499$

$E = t_c S_e \sqrt{1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{SS_x}}$

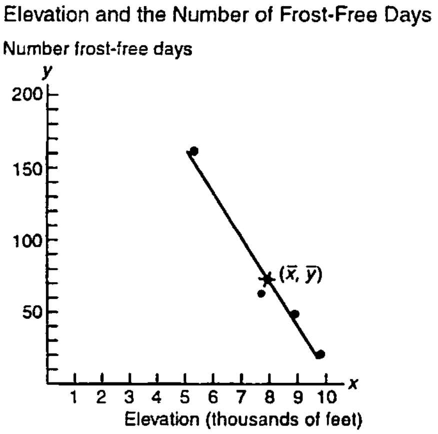
$= 3.499(0.93152)\sqrt{1 + \frac{1}{9} + \frac{(9.5 - 7.0)^2}{108}}$

$= 3.52$

A 99% confidence interval for y is

$$y_p - E \le y \le y_p + E$$
$$8.15 - 3.52 \le y \le 8.15 + 3.52$$
$$4.6 \le y \le 11.7 \text{ irrelevant responses}$$

13. (a)



(b) $\bar{x} = \frac{\sum x}{n} = \frac{39.6}{5} = 7.92$

$\bar{y} = \frac{\sum y}{n} = \frac{368}{5} = 73.6$

$b = \frac{SS_{xy}}{SS_x} = \frac{-352.26}{11.408} = -30.8783$

$a = \bar{y} - b\bar{x} = 73.6 - (-30.8783)(7.92) = 318.16$

$y = a + bx$ or $y = 318.16 - 30.878x$

(c) See figure of part (a).

Note: Compare this figure to that in Problem 5 above, the point (5.3, 162) is an outlier (possibly in x, definitely in y) but it is more or less along the regression line that would be drawn if it were eliminated from the data set. Thus, this is not an "influential" observation.

(d) $S_e = \sqrt{\frac{SS_y - bSS_{xy}}{n-2}}$

$= \sqrt{\frac{11.299.2 - (-30.8783)(-352.26)}{5-2}}$

$= 11.860$

(e) Use $x = 6$.

$y_p = 318.16 - 30.878(6) = 132.89$ days

(f) $t_{0.85, 3d.f.} = 1.924$

$E = t_c S_e \sqrt{1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{SS_x}}$

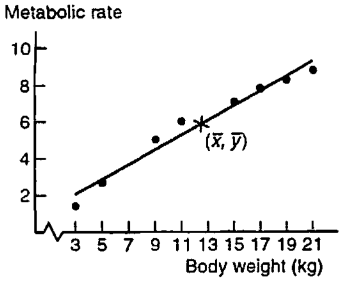
$= 1.924(11.860)\sqrt{1 + \frac{1}{5} + \frac{(6-7.92)^2}{11.408}}$

$= 28.16$

A 85% confidence interval for y is

$$y_p - E \le y \le y_p + E$$
$$132.89 - 28.16 \le y \le 132.89 + 28.16$$
$$104.7 \le y \le 161.1 \text{ days}$$

14. (a) Body Weight (kg) and Metabolic Rate (100 kcal/24 h)



(b) $\bar{x} = \frac{\sum x}{n} = \frac{100}{8} = 12.5$

$\bar{y} = \frac{\sum y}{n} = \frac{47.1}{8} = 5.8875$

$b = \frac{SS_{xy}}{SS_x} = \frac{121.55}{302} = 0.40248$

$a = \bar{y} - b\bar{x} = 5.8875 - 0.40248(12.5) = 0.8565$

$y = a + bx$ or $y = 0.8565 + 0.4025x$

(c) See figure of part (a).

(d)
$$S_e = \sqrt{\frac{SS_y - bSS_{xy}}{n-2}} = \sqrt{\frac{50.52875 - (0.40248)(121.55)}{8-2}} = 0.5176$$

(e) Use $x = 16$.

$y_p = 0.8565 + 0.4025(16) = 7.3$ (100 kcal/24 h)

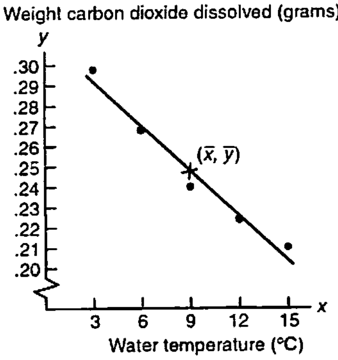
(f) $t_{0.75, 6 d.f.} = 1.273$

$$E = t_c S_e \sqrt{1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{SS_x}} = 1.273(0.5176)\sqrt{1 + \frac{1}{8} + \frac{(16-12.5)^2}{302}} = 0.7$$

A 75% confidence interval for y is

$$y_p - E \le y \le y_p + E$$
$$7.3 - 0.7 \le y \le 7.3 + 0.7$$
$$6.6 \le y \le 8.0 \text{ (100 kcal/24 h)}$$

15. (a) Solubility of Carbon Dioxide in Water



(b) $\bar{x} = \frac{\sum x}{n} = \frac{45}{5} = 9.0$

$\bar{y} = \frac{\sum y}{n} = \frac{1.24}{5} = 0.248$

$b = \frac{SS_{xy}}{SS_x} = \frac{-0.66}{90} = -0.00733$

$a = \bar{y} - b\bar{x} = 0.248 - (-0.00733)(9.0) = 0.314$

$y = a + bx$ or $y = 0.314 - 0.00733x$

(c) See figure of part (a).

(d) $$S_e = \sqrt{\frac{SS_y - bSS_{xy}}{n-2}} = \sqrt{\frac{0.004984 - (-0.00733)(-0.66)}{5-2}} = 0.007$$

(e) Use $x = 10$.

$y_p = 0.314 - 0.00733(10) = 0.241$ grams

(f) $t_{0.90,3d.f.} = 2.353$

$$E = t_c S_e \sqrt{1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{SS_x}} = 2.353(0.007)\sqrt{1 + \frac{1}{5} + \frac{(10-9)^2}{90}} = 0.018$$

A 90% confidence interval for y is

$$y_p - \Sigma \le y \le y_p + E$$
$$0.241 - 0.018 \le y \le 0.241 + 0.018$$
$$0.223 \le y \le 0.259 \text{ grams}$$

16. (a) Results checks.

(b) Results checks.

(c) Yes.

(d) $y = 0.143 + 1.071x$

$$y - 0.143 = 1.071x$$

$$\frac{y - 0.143}{1.071} = x$$

$$\frac{1}{1.071}y - \frac{0.143}{1.071} = x$$

or

$$x = 0.9337y - 0.1335$$

The equation $x = 0.9337y - 0.1335$ does not match part (b), with the symbols x and y exchanged.

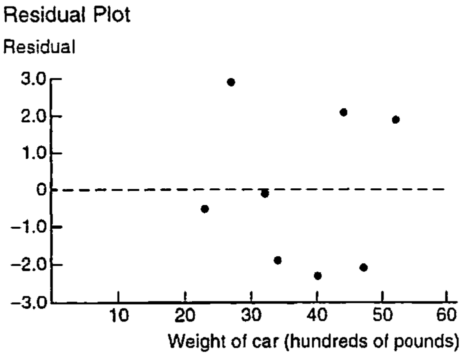
(e) In general, switching x and y values produces a different least-squares equation. It is important that when you perform a linear regression, you know which variable is the explanatory variable and which is the response variable.

17. (a) Yes. The pattern of residuals appears randomly scattered around the horizontal line at 0.

(b) No. There do not appear to be any outliers.

18. (a)

x	y	$y_p = 43.3263 - 0.6007x$	Residual $= y - y_p$
27	30	27.1	2.9
44	19	16.9	2.1
32	24	24.1	–0.1
47	13	15.1	–2.1
23	29	29.5	–0.5
40	17	19.3	–2.3
34	21	22.9	–1.9
52	14	12.1	1.9



(b) The residuals seem to be scattered randomly around the horizontal line at 0. There do not appear to be any outliers.

Section 10.3

1. (a) No, high positive correlation does not mean causation.

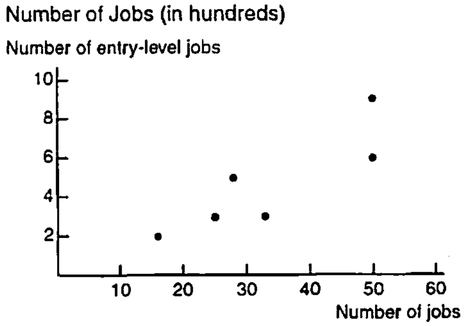
(b) An increase in the population is a third factor that might cause traffic accidents and the number of safety stickers to increase together.

2. (a) No, high positive correlation does not mean causation.
   (b) There is an increase in buying power due to increase in salaries.

3. (a) No, strong negative correlation does not mean causation.
   (b) Better medical treatment is a third factor that might be decreasing infant mortalities and at the same time increasing life span.

4. (a) No, strong positive correlation does not mean causation.
   (b) An increase in population could account for increases both in consumption of soda pop and in number of traffic accidents.

5. (a)

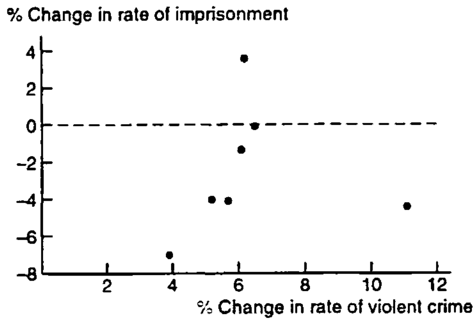


   (b) r should be close to 1 because the points seem to be clustered fairly close to a straight line going up from left to right.

   (c) $$r = \frac{SS_{xy}}{\sqrt{SS_x SS_y}} = \frac{153.\overline{3}}{\sqrt{953.\overline{3}(33.\overline{3})}} = 0.860$$
   $$r^2 = (0.860)^2 = 0.740$$

   This means that 74.0% of the variation in y = number of entry-level jobs can be explained by the corresponding variation in x = total number of jobs using the least squares line. 100% − 74.0% = 26.0% of the variation is unexplained.

6. (a)



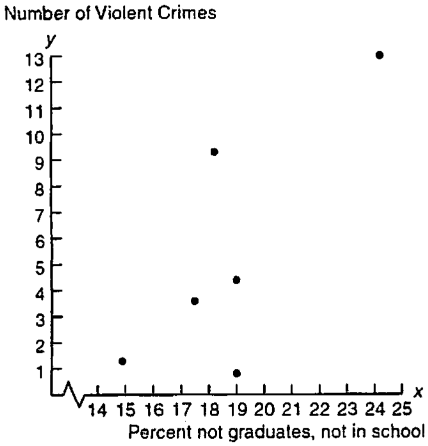
   (b) r should be close to 0 because the points are not all clustered around a straight line, due to (11.1, −4.4) (which is an influential observation).

(c) $r = \frac{SS_{xy}}{\sqrt{SS_x SS_y}} = \frac{3.9314}{\sqrt{30.4086(72.8486)}} = 0.084$

$r^2 = (0.084)^2 = 0.007$

This means that 0.7% of the variation in y = percent change in the rate of imprisonment can be explained by the corresponding variation in x = percent change in the rate of violent crime using the least squares line. 100% – 0.7% = 99.3% of the variation is unexplained.

7. (a) Percentage of 16 to 19-Year-Olds Not in School and Number of Violent Crimes per 1000



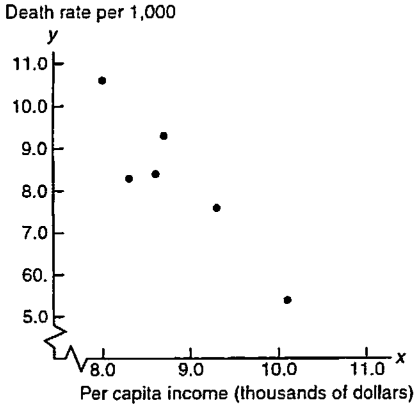
(b) r should be closer to 1 because the points are not somewhat clustered around a straight line going up from left to right.

(c) $r = \frac{SS_{xy}}{\sqrt{SS_x SS_y}} = \frac{55.91}{\sqrt{46.5(115.18)}} = 0.764$

$r^2 = (0.764)^2 = 0.584$

This means that 58.4% of the variation in y = reported violent crimes per 1000 residents can be explained by the corresponding variation in x = percentage of 16-to-19-year-olds not in school and not high school graduates using the least squares line. 100% – 58.4% = 41.6% of the variation is unexplained.

8. (a) Per Capita Income and Death Rates in Small Cities in Oregon



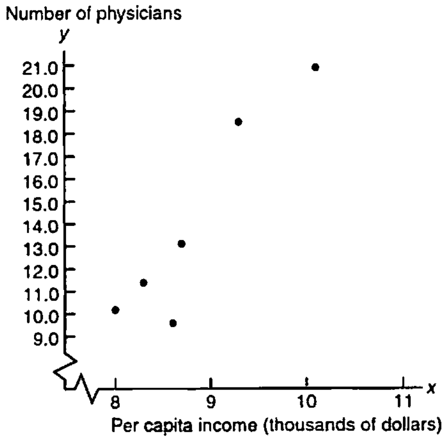
(b) r should be close to –1 because the points are clustered fairly close to a straight line going down from left to right.

(c) $r = \frac{SS_{xy}}{\sqrt{SS_x SS_y}} = \frac{-6.073}{\sqrt{2.873(15.193)}} = -0.919$

$r^2 = (-0.919)^2 = 0.845$

This means that 84.5% of the variation in y = death rate per 1000 residents can be explained by the corresponding variation in x = per capita income in thousands of dollars using the least squares line. 100% – 84.5%= 15.5% of the variation is unexplained.

9. (a) Per Capita Income and Number of Physicians per 10,000



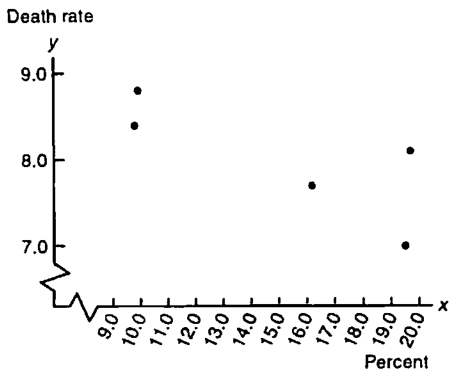
(b) r should be close to 1 because the points are clustered fairly close to a straight line going up from left to right.

(c) $r = \frac{SS_{xy}}{\sqrt{SS_x SS_y}} = \frac{16.54}{\sqrt{2.873(109.215)}} = 0.934$

$r^2 = (0.934)^2 = 0.872$

This means that 87.2% of the variation in y = number of medical doctors per 10,000 residents can be explained by the corresponding variation in x = per capita income in thousands of dollars using the least squares line. 100% – 87.2% = 12.8% of the variation is unexplained.

10. (a) Percentage of 16 to 19-Year-Olds Not in School and Death Rate per 1000 Residents



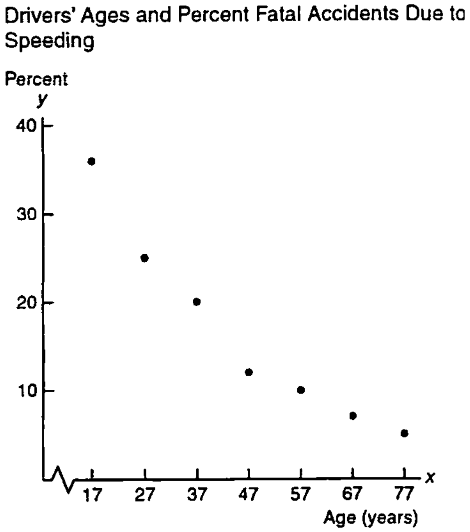
(b) r should be closer to –1 because the points are clustered some what close to a straight line going down from left to right.

(c) $r = \dfrac{SS_{xy}}{\sqrt{SS_x SS_y}} = \dfrac{-10.55}{\sqrt{96.828(1.9)}} = -0.778$

$r^2 = (-0.778)^2 = 0.605$

This means that 60.5% of the variation in y = death rate per 1000 residents can be explained by the corresponding variation in x = percentage of 16-to-19-year-olds not in school and not high school graduates using the least squares line. 100% – 60.5% = 39.5% of the variation is unexplained.

11. (a)



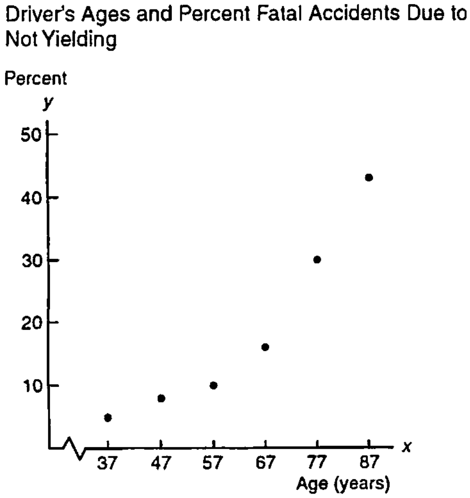
(b) r should be closer to –1 because the points are clustered very close to a straight line going down from left to right. (Note also that the data values fall nicely on a curve.)

(c) $r = \dfrac{SS_{xy}}{\sqrt{SS_x SS_y}} = \dfrac{-1390}{\sqrt{2800(749.714)}} = -0.959$

$r^2 = (-0.959)^2 = 0.920$

This means that 92% of the variation in y = percentage of all fatal accidents due to speeding can be explained by the corresponding variation in x = age in years of a licensed automobile driver using the least squares line. 100% – 92% = 8% of the variation is unexplained.

12. (a)



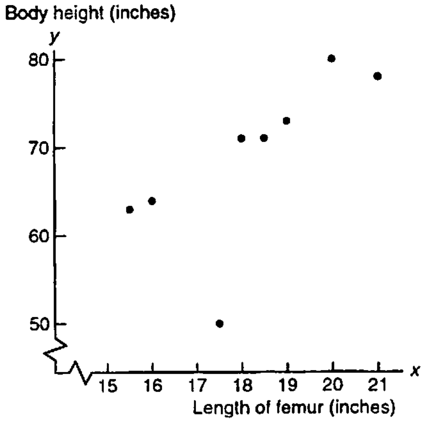
(b) r should be closer to 1 because the points are clustered very close to a straight line going up from left to right. (Note also that the data follow a curve.)

(c) $$r = \frac{SS_{xy}}{\sqrt{SS_x SS_y}} = \frac{1310}{\sqrt{1750(1103.\overline{3})}} = 0.943$$

$r^2 = (0.943)^2 = 0.889$

This means that 88.9% of the variation in y = percentage of fatal accidents due to failure to yield the right of way can be explained by the corresponding variation in x = age of a licensed driver in years using the least squares line. 100% − 88.9% = 11.1% of the variation is unexplained.

13. (a) Body Height and Bone Size



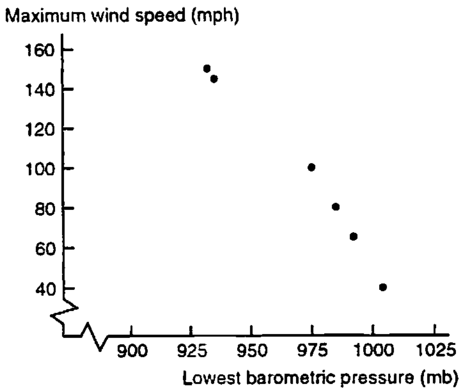
(b) r should be closer to 1 because the points are clustered close to a straight line going up from left to right.

(c) $$r = \frac{SS_{xy}}{\sqrt{SS_x SS_y}} = \frac{88.875}{\sqrt{24.4688(647.5)}} = 0.7061$$

$r^2 = (0.7061)^2 = 0.499$

This means that 49.9% of the variation in y = body height can be explained by the corresponding variation in x = length of femur using the least squares line. 100% − 49.9% = 50.1% of the variation is unexplained.

14. (a) Lowest Barometric Pressure and Maximum Wind Speed for Tropical Cyclones



(b) r should be closer to −1 because the points are clustered very close to a straight line going down from left to right.

(c) $r = \frac{SS_{xy}}{\sqrt{SS_x SS_y}} = \frac{-6575}{\sqrt{4557.5(9683.\overline{3})}} = -0.9897$

$r^2 = (-0.9897)^2 = 0.9795$ or 0.98

This means that 98% of the variation in y = maximum wind speed of the cyclone can be explained by the corresponding variation in x = lowest pressure as a cyclone approaches using the least squares line. 100% – 98% = 2% of the variation is unexplained.

15. (a) We get the same result.

$SS_{xy} = SS_{yx}$

(b) We get the same result.

(c) We get the same result.

(d) First set: $r = \frac{SS_{xy}}{\sqrt{SS_x SS_y}} = \frac{5}{\sqrt{4.\overline{6}(14)}} = 0.618590$

Second set: $r = \frac{SS_{xy}}{\sqrt{SS_x SS_y}} = \frac{5}{\sqrt{14(4.\overline{6})}} = 0.618590$

$r = 0.618590$ in both cases.

The least-squares equations are not necessarily the same.

16. (a) No. Interest rate probably affects both investment returns.

(b) For $w = 0.6x + 0.4y$, $a = 0.6$, $b = 0.4$

$\mu_w = a\mu_x + b\mu_y$

$\mu_w = 0.6(7.32) + 0.4(13.19)$

$\mu_w = 9.67$

$\sigma_w^2 = a^2\sigma_x^2 + b^2\sigma_y^2 + 2ab\sigma_x\sigma_y\rho$

$\sigma_w^2 = (0.6)^2(6.59)^2 + (0.4)^2(18.56)^2 + 2(0.6)(0.4)(6.59)(18.56)(0.424)$

$\sigma_w^2 = 95.64$

$\sigma_w = \sqrt{95.64} = 9.78$

(c) For $w = 0.4x + 0.6y$, $a = 0.4$, $b = 0.6$

$\mu_w = 0.4(7.32) + 0.6(13.19)$

$\mu_w = 10.84$

$\sigma_w^2 = (0.4)^2(6.59)^2 + (0.6)^2(18.56)^2 + 2(0.4)(0.6)(6.59)(18.56)(0.424)$

$\sigma_w^2 = 155.85$

$\sigma_w = \sqrt{155.85} = 12.48$

(d) $w = 0.4x + 0.6y$ produces higher returns with greater risk as measured by σ_w.

Section 10.4

1. (a) Results check.

(b) $H_0: \rho = 0$
$H_1: \rho < 0$
$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{-0.377\sqrt{8-2}}{\sqrt{1-(-0.377)^2}} = -0.997$$
$d.f. = n-2 = 8-2 = 6$

At 5% level of significance, $t_0 = -1.943$.
P value > 0.125
Since $-0.997 > -1.943$ and P value > 0.125, we fail to reject H_0. The sample evidence does not support a negative correlation.

(c) $H_0: \beta = 0$
$H_1: \beta \neq 0$
$$t = \frac{b-\beta}{\frac{S_e}{\sqrt{SS_x}}} = \frac{-0.468-0}{\frac{7.443}{\sqrt{252.40}}} = -0.999$$
$d.f. = n-2 = 8-2 = 6$

At 5% level of significance, $t_0 = \pm 2.447$.
P value > 0.250
Since $-2.447 < -0.999 < 2.447$, we fail to reject H_0. The sample evidence does not support a nonzero slope.

(d) $d.f. = 6, t_c = 1.273, b \approx -0.468$
$$E = t_c \frac{S_e}{\sqrt{SS_x}} = 1.273 \frac{7.443}{\sqrt{252.40}} = 0.596$$

A 75% confidence interval is

$$b - E < \beta < b + E$$
$$-0.468 - 0.596 < \beta < -0.468 + 0.596$$
$$-1.064 < \beta < 0.128$$

Since the confidence interval includes both positive and negative values, we conclude that the slope is zero and so faculty salaries are not tied to tuition. For a $100 change in tuition and fees, there is essentially no change in faculty salaries.

2. (a) Results check.

(b) $H_0: \rho = 0$
$H_1: \rho \neq 0$
$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0.187\sqrt{9-2}}{\sqrt{1-(0.187)^2}} = 0.504$$
$d.f. = n-2 = 9-2 = 7$

At 5% level of significance, $t_0 = \pm 2.365$.
P value > 0.250
Since $-2.365 < 0.504 < 2.365$ and P value > 0.250, we fail to reject H_0. The sample evidence does not support a nonzero correlation.

(c) $H_0: \beta = 0$

$H_1: \beta \neq 0$

$t = \frac{b-\beta}{\frac{S_e}{\sqrt{SS_x}}} = \frac{0.172-0}{\frac{6.369}{\sqrt{350.70}}} = 0.506$

$d.f. = n-2 = 9-2 = 7$

At 5% level of significance. $t_0 = \pm 2.365$.

P value > 0.250

Since $-2.365 < 0.506 < 2.365$ and P value > 0.250. we fail to reject H_0. The sample evidence does not support a nonzero slope. (The t–values in (b) and (c) differ due to roundoff error.)

(d) $d.f. = 7, t_c = 1.415, b \approx 0.172$

$E = t_c \frac{S_e}{\sqrt{SS_x}} = 1.415 \frac{6.369}{\sqrt{350.70}} = 0.481$

An 80% confidence interval is

$$b - E < \beta < b + E$$
$$0.172 - 0.481 < \beta < 0.172 + 0.481$$
$$-0.309 < \beta < 0.653$$

Since the confidence interval includes both positive and negative values, we conclude that the slope is zero and so faculty salaries are not tied to tuition. For a $100 change in tuition and fees, there is essentially no change in faculty salaries.

3. (a) Results check.

(b) $H_0: \rho = 0$

$H_1: \rho < 0$

$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{-0.976\sqrt{7-2}}{\sqrt{1-(-0.976)^2}} = -10.02$

$d.f. = n-2 = 7-2 = 5$

At 1% level of significance, $t_0 = -3.365$.

P value < 0.005

Since $-10.02 < -3.365$ and P value < 0.005, we reject H_0. The sample evidence supports a negative correlation.

(c) $H_0: \beta = 0$

$H_1: \beta < 0$

$t = \frac{b-\beta}{\frac{S_e}{\sqrt{SS_x}}} = \frac{-0.054-0}{\frac{0.166}{\sqrt{940.89}}} = -9.98$

$d.f. = n-2 = 7-2 = 5$

At 1% level of significance, $t_0 = -3.365$.

P value < 0.005

Since $-9.98 < -3.365$ and P value < 0.005, we reject H_0. The sample evidence supports a negative slope.

(d) $d.f. = 5, t_c = 2.015, b \approx -0.054$

$$E = t_c \frac{S_e}{\sqrt{SS_x}} = 2.015 \frac{0.166}{\sqrt{940.89}} = 0.011$$

A 90% confidence interval is

$$\begin{aligned} b - E &< \beta < b + E \\ -0.054 - 0.011 &< \beta < -0.054 + 0.011 \\ -0.065 &< \beta < -0.043 \end{aligned}$$

For every meter more of depth, the optimal time decreases from about 0.04 to 0.07 hour.

4. (a) Results check.

(b) $H_0: \rho = 0$
$H_1: \rho > 0$

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0.984\sqrt{5-2}}{\sqrt{1-(0.984)^2}} = 9.57$$

$d.f. = n - 2 = 5 - 2 = 3$

At 1% level of significance, $t_0 = 4.541$.
P value < 0.005
Since 9.57 > 4.541 and P value < 0.005, we reject H_0. The sample evidence supports a positive correlation.

(c) $H_0: \beta = 0$
$H_1: \beta > 0$

$$t = \frac{b - \beta}{\frac{S_e}{\sqrt{SS_x}}} = \frac{6.876 - 0}{\frac{2.532}{\sqrt{12.25}}} = 9.505$$

$d.f. = n - 2 = 5 - 2 = 3$

At 1% level of significance, $t_0 = 4.541$.
P value < 0.005
Since 9.505 > 4.541 and P value < 0.005, we reject H_0. The sample evidence supports a positive slope.
Note: carrying more decimal places than shown in (b) gives $t = 9.505$, the same as in (c).

(d) $d.f. = 3, t_c = 3.182, b \approx 6.876$

$$E = t_c \frac{S_e}{\sqrt{SS_x}} = 3.182 \frac{2.532}{\sqrt{12.25}} = 2.302$$

A 95% confidence interval is

$$\begin{aligned} b - E &< \beta < b + E \\ 6.876 - 2.302 &< \beta < 6.876 + 2.302 \\ 4.57 &< \beta < 9.18 \end{aligned}$$

For every one unit increase in oxygen pressure breathing only available air, the oxygen pressure breathing pure oxygen increases from about 4.57 units to 9.18 units.

5. (a) Results check.

(b) $H_0: \rho = 0$
$H_1: \rho > 0$
$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0.956\sqrt{6-2}}{\sqrt{1-(0.956)^2}} = 6.517$
$d.f. = n-2 = 6-2 = 4$

At 1% level of significance, $t_0 = 3.747$.
P value < 0.005
Since 6.517 > 3.747 and P value < 0.005, we reject H_0. The sample evidence supports a positive correlation.

(c) $H_0: \beta = 0$
$H_1: \beta > 0$
$t = \frac{b-\beta}{\frac{S_e}{\sqrt{SS_x}}} = \frac{0.758-0}{\frac{0.1527}{\sqrt{1.733}}} = 6.535$
$d.f. = n-2 = 6-2 = 4$

At 1% level of significance, $t_0 = 3.747$.
P value < 0.005
Since 6.535 > 3.747 and P value < 0.005, we reject H_0. The sample evidence supports a positive slope.

(d) $d.f. = 4, t_c = 2.132, b \approx 0.758$
$E = t_c \frac{S_e}{\sqrt{SS_x}} = 2.132 \frac{0.1527}{\sqrt{1.733}} = 0.247$

A 90% confidence interval is

$$b - E < \beta < b + E$$
$$0.758 - 0.247 < \beta < 0.758 + 0.247$$
$$0.511 < \beta < 1.005$$

For every \$1,000 increase in list price, there is an increase in dealer price of between \$511 and \$1005.

6. (a) Results check.

(b) $H_0: \rho = 0$
$H_1: \rho > 0$
$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0.977\sqrt{5-2}}{\sqrt{1-(0.977)^2}} = 7.936$
$d.f. = n-2 = 5-2 = 3$

At 1% level of significance, $t_0 = 4.541$.
P value < 0.005
Since 7.936 > 4.541 and P value < 0.005, we reject H_0. The sample evidence supports a positive correlation.

(c) $H_0: \beta = 0$
$H_1: \beta > 0$
$$t = \frac{b-\beta}{\frac{S_e}{\sqrt{SS_x}}} = \frac{0.879-0}{\frac{0.1522}{\sqrt{188.26}}} = 7.924$$
$d.f. = n-2 = 5-2 = 3$

At 1% level of significance, $t_0 = 4.541$.
P value < 0.005
Since 7.924 > 4.541 and P value < 0.005, we reject H_0. The sample evidence supports a positive slope.

(d) $d.f. = 3, t_c = 2.353, b \approx 0.879$
$$E = t_c \frac{S_e}{\sqrt{SS_x}} = 2.353\frac{1.522}{\sqrt{188.26}} = 0.261$$

A 90% confidence interval is

$$b-E < \beta < b+E$$
$$0.879-0.261 < \beta < 0.879+0.261$$
$$0.618 < \beta < 1.140$$

For every increase of $1,000 in list price, the dealer price is from $618 to $1,140 higher.

7. (a) $H_0: \rho = 0$
$H_1: \rho \neq 0$
$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0.90\sqrt{6-2}}{\sqrt{1-(0.90)^2}} = 4.129$$
$d.f. = n-2 = 6-2 = 4$

At 1% level of significance, $t_0 = \pm 4.604$.
Since $-4.604 < 4.129 < 4.604$ and $0.01 < P$ value, we do not reject H_0. The correlation coefficient ρ is not significantly different from zero at the 0.01 level of significance.

(b) $H_0: \rho = 0$
$H_1: \rho \neq 0$
$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0.90\sqrt{10-2}}{\sqrt{1-(0.90)^2}} = 5.840$$
$d.f. = n-2 = 10-2 = 8$

At 1% level of significance, $t_0 = \pm 3.355$.

Since 5.840 > 3.355 and P value < 0.01, we reject H_0. The correlation coefficient ρ is significantly different from zero at the 0.01 level of significance.

(c) From part (a) to part (b), n increased from 6 to 10, the test statistic t increased from 4.12 to 5.840, and the critical values t_0 decreased (in absolute value) from 4.604 to 3.355. For the same $r = 0.90$ and the same level of significance $\alpha = 0.01$, we rejected H_0 for the larger n but not for the smaller n.
In general, as n increases, the degrees of freedom ($n - 2$) increase and the critical value(s) become(s) closer to zero. Also, as n increases, the test statistic $\left(t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}\right)$ moves farther from zero. The combination of the critical value(s) approaching zero while the test statistic moves farther out into the tail of the t–distribution means we are more likely to reject H_0 for larger n (using the same r and α).

8. (a) Yes. The t values are equal (differences are due to rounding error).

(b) Essay.

A possible proof, working on the left side first, follows:

Prove: $t = \dfrac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \dfrac{b\sqrt{SS_x}}{S_e}$

Proof: Begin with these versions of r, b, S_e, and r^2:

page 582 (9): $r = \dfrac{SS_{xy}}{\sqrt{SS_x\, SS_y}}$

page 561: $b = \dfrac{SS_{xy}}{SS_x}$

page 568 (7): $S_e = \sqrt{\dfrac{\Sigma(y-y_p)^2}{n-2}}$

or $\sqrt{\Sigma(y-y_p)^2} = \sqrt{n-2}\, S_e$

page 589: $r^2 = \dfrac{\Sigma(y_p-\overline{y})^2}{\Sigma(y-\overline{y})^2} = \dfrac{\Sigma(y_p-\overline{y})^2}{SS_y}$ (1)

Then from page 589: $SS_y = \Sigma(y-\overline{y})^2 = \Sigma(y_p-\overline{y})^2 + \Sigma(y-y_p)^2$
or $\Sigma(y_p-\overline{y})^2 = SS_y - \Sigma(y-y_p)^2$ (2)

So $r^2 = \dfrac{\Sigma(y_p-\overline{y})^2}{SS_y} = \dfrac{SS_y - \Sigma(y-y_p)^2}{SS_y}$ using (2)

$= 1 - \dfrac{\Sigma(y-y_p)^2}{SS_y}$

or $1-r^2 = \dfrac{\Sigma(y-y_p)^2}{SS_y}$

$\sqrt{1-r^2} = \dfrac{\sqrt{\Sigma(y-y_p)^2}}{\sqrt{SS_y}}$

$\dfrac{1}{\sqrt{1-r^2}} = \dfrac{\sqrt{SS_y}}{\sqrt{\Sigma(y-y_p)^2}} = \dfrac{\sqrt{SS_y}}{\sqrt{n-2}\, S_e}$ using (1)

Then $t = \dfrac{r\sqrt{n-2}}{\sqrt{1-r^2}} = r\left(\dfrac{1}{\sqrt{1-r^2}}\right)\sqrt{n-2} = \left(\dfrac{SS_{xy}}{\sqrt{SS_x\, SS_y}}\right)\left(\dfrac{\sqrt{SS_y}}{\sqrt{n-2}\, S_e}\right)\sqrt{n-2}$

$= \dfrac{SS_{xy}}{\sqrt{SS_x}\, S_e} = \dfrac{SS_{xy}}{\sqrt{SS_x}\, S_e}\left(\dfrac{\sqrt{SS_x}}{\sqrt{SS_x}}\right) = \left(\dfrac{SS_{xy}}{SS_x}\right)\dfrac{\sqrt{SS_x}}{S_e} = \dfrac{b\sqrt{SS_x}}{S_e}$

as was to be shown.

Section 10.5

1. $x_1 = 1.6 + 3.5x_2 - 7.9x_3 + 2.0x_4$

(a) The response variable is x_1.

(b) The constant term is 1.6.
The coefficient 3.5 goes with corresponding explanatory variable x_2.
The coefficient –7.9 goes with corresponding explanatory variable x_3.
The coefficient 2.0 goes with corresponding explanatory variable x_4.

(c) $x_2 = 2, x_3 = 1, x_4 = 5$
$x_1 = 1.6 + 3.5(2) - 7.9(1) + 2.0(5) = 10.7$
The predicted value is 10.7.

(d) In multiple regression, the coefficients of the explanatory variables can be thought of as "slopes" (the change in the response variables per unit change in the explanatory variable) if we look at one explanatory variable's coefficient at a time, while holding the other explanatory variables as arbitrary and fixed constants.
x_3 and x_4 held constant. x_2 increased by one unit:
The change in x_1 would be an increase of 3.5 units.
x_3 and x_4 held constant, x_2 increased by two units:
The change in x_1 would be an increase of 2(3.5) = 7 units.
x_3 and x_4 held constant, x_2 decreased by four units:
The change in x_1 would be an decrease of 4(3.5) = 14 units.

(e) $d.f. = n - k - 1 = 12 - 3 - 1 = 8$
A 90% confidence interval for the coefficient of x_2 is $b_2 - tS_2 < \beta_2 < b_2 + tS_2$
$3.5 - 1.86(0.419) < \beta_2 < 3.5 + 1.86(0.419)$
$2.72 < \beta_2 < 4.28$

(f) $H_0: \beta_2 = 0$
$H_1: \beta_2 \neq 0$
$t = \frac{b_2 - \beta_2}{S_2} = \frac{3.5 - 0}{0.419} = 8.35$
$d.f. = 8, t_0 = \pm 2.306$ for 5% level of significance.
Since $8.35 > 2.306$, we reject H_0.
We conclude that $\beta_2 \neq 0$ and x_2 should be included as an explanatory variable in the least-squares equation.

2. $x_3 = -16.5 + 4.0x_1 + 9.2x_4 - 1.1x_7$

(a) The response variable is x_3.
The explanatory variables are x_1, x_4, and x_7.

(b) The constant term is -16.5.
The coefficient 4.0 goes with the corresponding explanatory variable x_1.
The coefficient 9.2 goes with the corresponding explanatory variable x_4.
The coefficient -1.1 goes with the corresponding explanatory variable x_7.

(c) $x_1 = 10, x_4 = -1, x_7 = 2$
$x_3 = -16.5 + 4.0(10) + 9.2(-1) - 1.1(2) = 12.1$
The predicted value is 12.1.

(d) In multiple regression, the coefficients of the explanatory variables can be thought of as "slopes" (the change in the response variables per unit change in the explanatory variable) if we look at one explanatory variable's coefficient at a time, while holding the other explanatory variables as arbitrary and fixed constants.

x_1 and x_7 held constant, x_4 increased by one unit:

The change in x_3 would be an increase of 9.2 units.

x_1 and x_7 held constant, x_4 increased by three units:

The change in x_3 would be an increase of 3(9.2) = 27.6 units.

x_1 and x_7 held constant, x_4 decreased by two units:

The change in x_3 would be a decrease of 2(9.2) = 18.4 units.

(e) $d.f. = n - k - 1 = 15 - 3 - 1 = 11$

A 90% confidence interval for the coefficient of x_4 is $b_4 - tS_4 < \beta_4 < b_4 + tS_4$

$9.2 - 1.796(0.921) < \beta_4 < 9.2 + 1.796(0.921)$

$7.55 < \beta_4 < 10.85$

(f) $H_0\colon \beta_4 = 0$

$H_1\colon \beta_4 \neq 0$

$$t = \frac{b_4 - \beta_4}{S_4} = \frac{9.2 - 0}{0.921} = 9.989$$

$d.f. = 11, t_0 = \pm 3.106$ for 1% level of significance.

Since 9.989 > 3.106, we reject H_0.

We conclude that $\beta_4 \neq 0$ and x_4 should be included as an explanatory variable in the least-squares equation.

3. (a)

	$\bar{x}$	s	$CV = \frac{s}{\bar{x}} \cdot 100$
x_1	150.09	13.63	9.08%
x_2	62.45	9.11	14.59%
x_3	195.0	17.31	8.88%

Relative to its mean, x_2 has the greatest spread of data values and x_3 has the smallest spread of data values.

(b) $r^2_{x_1x_2} \approx (0.979)^2 \approx 0.958$

$r^2_{x_1x_3} \approx (0.971)^2 \approx 0.943$

$r^2_{x_2x_3} \approx (0.946)^2 \approx 0.895$

The variable x_2 has the greatest influence on x_1 (0.958 > 0.943).

Yes. Both variables x_2 and x_3 show a strong influence on x_1 because 0.958 and 0.943 are close to 1.

95.8% of the variation of x_1 can be explained by the corresponding variation in x_2.

94.3% of the variation of x_1 can be explained by the corresponding variation in x_3.

(c) $R^2 = 0.977$

97.7% of the variation in x_1 can be explained by the corresponding variation in x_2 and x_3 taken together.

(d) $x_1 = 30.99 + 0.861x_2 + 0.355x_3$

In multiple regression, the coefficients of the explanatory variables can be thought of as "slopes" (the change in the response variables per unit change in the explanatory variable) if we look at one explanatory variable's coefficient at a time, while holding the other explanatory variables as arbitrary and fixed constants.

If age (x_2) were held fixed and x_3 increased by 10 pounds, the systolic blood pressure is expected to increase by 0.335(10) = 3.35.

If weight (x_3) were held fixed and x_2 increased by 10 years, the systolic blood pressure is expected to increase by 0.861(10) = 8.61.

(e) $H_0: \beta_i = 0$

$H_1: \beta_i \neq 0$

$d.f = n - k - 1 = 11 - 2 - 1 = 8$

$t_0 = \pm 2.306$ for 5% level of significance.

For β_2, the sample test statistic is $t = 3.47$.

For β_3, the sample test statistic is $t = 2.56$.

Since 3.47 > 2.306 and 2.56 > 2.306, reject H_0 for each coefficient and conclude that the coefficients of x_2 and x_3 are not zero. Explanatory variables x_i whose coefficients (β_i) are nonzero contribute information in the least squares equation, i.e., without these x_i, the resulting least squares regression equation is not as good a fit to the data as is the regression equation which includes these x_i.

(f) $d.f. = 8, t = 1.86$

A 90% confidence interval for β_i is

$$b_i - tS_i < \beta_i < b_i + tS_i$$
$$0.861 - 1.86(0.2482) < \beta_2 < 0.861 + 1.86(0.2482)$$
$$0.40 < \beta_2 < 1.32$$
$$0.335 - 1.86(0.1307) < \beta_3 < 0.335 + 1.86(0.1307)$$
$$0.09 < \beta_3 < 0.58$$

(g) $x_1 = 30.99 + 0.861(68) + 0.335(192) \approx 153.9$

Michael's predicted systolic blood pressure is 153.9 and a 90% confidence interval for this new observation's value, given these x_i, i.e., the prediction interval, is 148.3 to 159.4.

4. (a)

	$\bar{x}$	s	$CV = \frac{s}{\bar{x}} \cdot 100$
x_1	79.04	12.28	15.53%
x_2	79.48	12.50	15.73%
x_3	81.48	11.77	14.44%
x_4	162.04	24.04	14.83%

Relative to its mean, each exam had about the same spread of scores. Yes; it seems that all of the exams were about the same level of difficulty.

(b) $r^2_{x_1x_2} \approx (0.901)^2 \approx 0.812$
$r^2_{x_1x_3} \approx (0.893)^2 \approx 0.797$
$r^2_{x_1x_4} \approx (0.946)^2 \approx 0.895$
$r^2_{x_2x_3} \approx (0.846)^2 \approx 0.716$
$r^2_{x_2x_4} \approx (0.929)^2 \approx 0.863$
$r^2_{x_3x_4} \approx (0.972)^2 \approx 0.945$

Exam 3 had the most influence on the final exam 4 (0.945 > 0.895 > 0.863). Even though exam 3 had more influence, the other two exams still had a lot of influence on the final because 0.895 and 0.863 are close to 1.

(c) $R^2 = 0.990$

99.0% of the variation in x_4 can be explained by the corresponding variation in x_1, x_2, and x_3 taken together.

(d) $x_4 = -4.34 + 0.356x_1 + 0.543x_2 + 1.17x_3$

In multiple regression, the coefficients of the explanatory variables can be thought of as "slopes" (the change in the response variables per unit change in the explanatory variable) if we look at one explanatory variable's coefficient at a time, while holding the other explanatory variables as arbitrary and fixed constants.

If age x_1 and x_2 are held fixed and x_3 is increased by 10 points, the final exam score is expected to increase by $10(1.17) = 11.7 \approx 12$ points.

(e) H_0: $\beta_i = 0$
H_1: $\beta_i \neq 0$
$d.f = n - k - 1 = 25 - 3 - 1 = 21$
$t_0 = \pm 2.080$ for 5% level of significance.
For β_1, the sample test statistic is $t = 2.93$.
For β_2, the sample test statistic is $t = 5.38$.
For β_3, the sample test statistic is $t = 11.33$.
Since 2.93 > 2.08, 5.38 > 2.08, and 11.33 > 2.08, reject H_0 for each coefficient and conclude that the coefficients of x_1, x_2 and x_3 are not zero. Explanatory variables x_i whose coefficients (β_i) are nonzero contribute information in the least squares equation, i.e., without these x_i, the resulting least squares regression equation is not as good a fit to the data as is the regression equation which includes these x_i.

(f) $d.f. = 21, t = 1.721$

A 90% confidence interval for β_i is

$$b_i - tS_i < \beta_i < b_i + tS_i$$

$$0.356 - 1.721(0.1214) < \beta_1 < 0.356 + 1.721(0.1214)$$
$$0.147 < \beta_1 < 0.565$$

$$0.543 - 1.721(0.1008) < \beta_2 < 0.543 + 1.721(0.1008)$$
$$0.370 < \beta_2 < 0.716$$

$$1.167 - 1.721(0.1030) < \beta_3 < 1.167 + 1.721(0.1030)$$
$$0.990 < \beta_3 < 1.344$$

(g) $x_4 = -4.34 + 0.356(68) + 0.543(72) + 1.17(75) \approx 147$

Susan's predicted score on the final exam is 147 and a 90% confidence interval for this new observation's value, given these x_i, i.e., the prediction interval, is 142 to 151, all rounded to whole numbers.

5. (a)

	$\bar{x}$	s	$CV = \frac{s}{\bar{x}} \cdot 100$
x_1	85.24	33.79	39.64%
x_2	8.74	3.89	44.51%
x_3	4.90	2.48	50.61%
x_4	9.92	5.17	52.12%

Relative to its mean, x_4 has the largest spread of data values. The larger the CV, the more we expect the variable to change relative to its average value, because a variable with a large CV has a large standard deviation, s, relative to $\bar{x}$, and s measures "spread," or variability, in the data. x_1 has a small CV because we divide by a large mean.

(b) $r^2_{x_1x_2} \approx (0.917)^2 \approx 0.841$
$r^2_{x_1x_3} \approx (0.930)^2 \approx 0.865$
$r^2_{x_1x_4} \approx (0.475)^2 \approx 0.226$
$r^2_{x_2x_3} \approx (0.790)^2 \approx 0.624$
$r^2_{x_2x_4} \approx (0.429)^2 \approx 0.184$
$r^2_{x_3x_4} \approx (0.299)^2 \approx 0.089$

The variable x_4 has the least influence on box office receipts x_1 $(0.226 < 0.841 < 0.865)$.

x_2 = production costs, $r^2_{x_1x_2} = 0.841$.
84.1% of the variation of box office receipts can be attributed to the corresponding variation in production costs.

(c) $R^2 = 0.967$
96.7% of the variation in x_1 can be explained by the corresponding variation in x_2, x_3, and x_4 taken together.

(d) $x_1 = 7.68 + 3.66x_2 + 7.62x_3 + 0.83x_4$

In multiple regression, the coefficients of the explanatory variables can be thought of as "slopes" (the change in the response variables per unit change in the explanatory variable) if we look at one explanatory variable's coefficient at a time, while holding the other explanatory variables as arbitrary and fixed constants.
If x_2 and x_4 were held fixed and x_3 were increased by 1 ($1 million), the corresponding change in x_1 (box office receipts) would be an increase of 7.62 or 7.62 million dollars.

(e) $H_0: \beta_i = 0$
$H_1: \beta_i \neq 0$
$d.f = n-k-1 = 10-3-1 = 6$
$t_0 = \pm 2.447$ for 5% level of significance.
For β_2, the sample test statistic is $t = 3.28$.
For β_3, the sample test statistic is $t = 4.60$.
For β_4, the sample test statistic is $t = 1.54$
Since $3.28 > 2.447$ and $4.60 > 2.447$, reject H_0 for coefficients β_2 and β_3 and conclude that the coefficients of x_2 and x_3 are not zero.
Since $-2.447 < 1.54 < 2.447$, do not reject H_0 for the coefficient β_4 and conclude that the coefficient of x_4 could be zero. If $\beta_4 = 0$, then x_4 contributes nothing to the (population) regression line. We can eliminate the variable x_4 and fit the (estimated) regression line without it, and probably see little, if any, difference between the predicted values of x_1 based on x_2 and x_3 only and the predicted values of x_1 based on x_2, x_3, and x_4.

(f) $d.f. = 6$, $t = 1.943$
A 90% confidence interval for β_i is

$$b_i - tS_i < \beta_i < b_i + tS_i$$

$$3.662 - 1.943(1.118) < \beta_2 < 3.662 + 1.943(1.118)$$
$$1.49 < \beta_2 < 5.83$$

$$7.621 - 1.943(1.657) < \beta_3 < 7.621 + 1.943(1.657)$$
$$4.40 < \beta_3 < 10.84$$

$$0.8285 - 1.943(0.5394) < \beta_4 < 0.8285 + 1.943(0.5394)$$
$$-0.22 < \beta_4 < 1.88$$

(g) $x_1 = 7.68 + 3.66(11.4) + 7.62(4.7) + 0.83(8.1) = 91.94$
The prediction is 91.94 million dollars and a 85% confidence interval for this new observation's value, given these x_i, i.e., the prediction interval, is \$77.6 million to \$106.3 million.

(h) $x_3 = -0.650 + 0.102x_1 - 0.260x_2 - 0.0899x_4$
$x_3 = -0.650 + 0.102(100) - 0.260(12) - 0.0899(9.2)$
$x_3 = 5.63$
The prediction is 5.63 million dollars and a 80% confidence interval for this new observation's value, given these x_i, i.e., the prediction interval, is \$4.21 million to \$7.04 million.

6. (a)

	$\bar{x}$	s	$CV = \frac{s}{\bar{x}} \cdot 100$
x_1	286.574	192.062	67.02%
x_2	3.326	2.011	60.46%
x_3	387.481	191.168	49.34%
x_4	8.100	3.775	46.60%
x_5	9.693	5.140	53.03%
x_6	7.741	4.896	63.25%

Relative to its mean, x_1 has the largest spread of data values and x_4 has a smallest spread of data values.

(b) $r_{23} = 0.844$
$r_{24} = 0.749$
$r_{25} = 0.838$
$r_{26} = -0.766$
$r_{34} = 0.906$
$r_{35} = 0.864$
$r_{36} = -0.807$
$r_{45} = 0.795$
$r_{46} = -0.841$
$r_{56} = -0.870$
$r^2_{x_1x_2} \approx (0.894)^2 \approx 0.799$
$r^2_{x_1x_3} \approx (0.946)^2 \approx 0.895$
$r^2_{x_1x_4} \approx (0.914)^2 \approx 0.835$
$r^2_{x_1x_5} \approx (0.954)^2 \approx 0.910$
$r^2_{x_1x_6} \approx (-0.912)^2 \approx 0.832$

The variable x_5 has the greatest influence on annual net sales (0.910 is the largest). The variable x_2 has the least influence on annual net sales (0.799 is the smallest).

(c) $R^2 = 0.993$
99.3% of the variation in x_1 can be explained by the corresponding variation in x_2, x_3, x_4, x_5, and x_6 taken together.

(d) $x_1 = -18.9 + 16.2x_2 + 0.175x_3 + 11.5x_4 + 13.6x_5 - 5.31x_6$

In multiple regression, the coefficients of the explanatory variables can be thought of as "slopes" (the change in the response variables per unit change in the explanatory variable) if we look at one explanatory variable's coefficient at a time, while holding the other explanatory variables as arbitrary and fixed constants.
If all explanatory variables but x_6 remained fixed and x_6 increased by 2, then the annual net sales are expected to decrease by 2(5.31) = 10.62 or $10,620.
If all explanatory variables but x_4 remained fixed and x_4 increased by 1($1000), then the annual net sales are expected to increase by 11.5 or $11,500.

(e) $H_0\colon \beta_i = 0$
$H_1\colon \beta_i \neq 0$
$d.f = 27 - 5 - 1 = 21$
$t_0 = \pm 2.080$ for 5% level of significance.
For β_2, the sample test statistic is $t = 4.57$.
For β_3, the sample test statistic is $t = 3.03$.
For β_4, the sample test statistic is $t = 4.55$
For β_5, the sample test statistic is $t = 7.67$.
For β_6, the sample test statistic is $t = -3.11$.

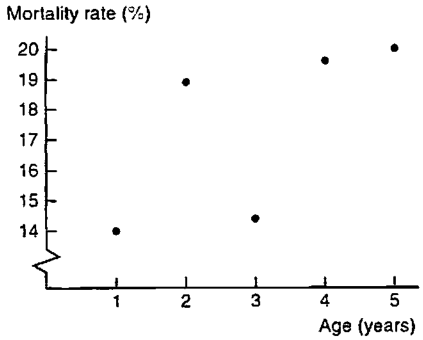
Since all of these t values are larger than 2.080 or less than -2.080, we reject H_0 for each coefficient and conclude that the coefficients of x_2, x_3, x_4, x_5, and x_6 are not zero.

(f) The predicted annual net sales are 194.41 (or $194.41 thousand) and the 80% confidence interval for this new observation's value, given these x_i, i.e., the prediction interval, is $160.76 thousand to $228.06 thousand.

(g) $x_4 = 4.14 + 0.0431x_1 - 0.800x_2 + 0.00059x_3 - 0.661x_5 + 0.057x_6$
The predicted amount spent on local advertising is \$5.571 thousand and the 80% confidence interval for this new observation's value, given these x_i, i.e., the prediction interval, is \$4.048 thousand to \$7.094 thousand. Advertising costs for this store should be between \$4,048 and \$7,094.

Chapter 10 Review Problems

1. (a) Age and Mortality Rate for Bighorn Sheep



(b) $\bar{x} = \frac{\sum x}{n} = \frac{15}{5} = 3$

$\bar{y} = \frac{\sum y}{n} = \frac{86.9}{5} = 17.38$

$b = \frac{SS_{xy}}{SS_x} = \frac{12.7}{10} = 1.27$

$a = \bar{y} - b\bar{x} = 17.38 - 1.27(3) = 13.57$

$y = a + bx$ or $y = 13.57 + 1.27x$

(c) $r = \frac{SS_{xy}}{\sqrt{SS_x SS_y}} = \frac{12.7}{\sqrt{10(34.408)}} = 0.685$

$r^2 = (0.685)^2 = 0.469$

The correlation coefficient r measures the strength of the linear relationship between a bighorn sheep's age and the mortality rate. The coefficient of determination, r^2, means that 46.9% of the variation in y = mortality rate in this age groups can be explained by the corresponding variation in x = age of a bighorn sheep using the least squares line.

(d) $H_0: \rho = 0$

$H_1: \rho > 0$

$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0.685\sqrt{5-2}}{\sqrt{1-(0.685)^2}} = 1.629$

At 1% level of significance, $t_0 = 4.541$.

Since $1.629 < 4.541$, we do not reject H_0.

There does not seem to be a positive correlation between age and mortality rate of bighorn sheep.

(e) $H_0: \beta = 0$

$H_1: \beta > 0$

$$t = \frac{b-\beta}{\frac{S_e}{\sqrt{SS_x}}} = \frac{1.27-0}{\frac{2.468}{\sqrt{10}}} = 1.627$$

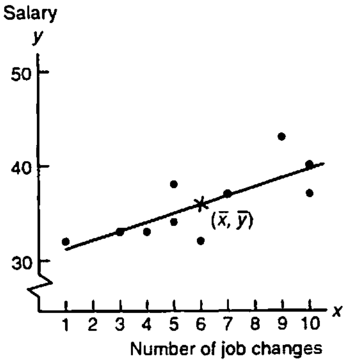
$d.f. = n-2 = 5-2 = 3$

At 1% level of significance, $t_0 = 4.541$.

Since $1.627 < 4.541$, we do not reject H_0.

The sample evidence does not support a positive slope.

2. (a) Annual Salary (thousands) and Number of Job Changes



(b) $\bar{x} = \frac{\sum x}{n} = \frac{60}{10} = 6.0$

$\bar{y} = \frac{\sum y}{n} = \frac{359}{10} = 35.9$

$b = \frac{SS_{xy}}{SS_x} = \frac{77}{82} = 0.939024$

$a = \bar{y} - b\bar{x} = 35.9 - 0.939024(6.0) = 30.266$

$y = a + bx$ or $y = 30.266 + 0.939x$

(c) See the figure in part (a).

(d) Let $x = 2$.

$y_p = 30.266 + 0.939(2) = 32.14$

The predicted salary is $32,140.

(e) $$S_e = \sqrt{\frac{SS_y - bSS_{xy}}{n-2}} = \sqrt{\frac{124.9 - 0.939024(77)}{10-2}} = 2.564058$$

(f) $$E = t_c S_e \sqrt{1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{SS_x}}$$

$$= 1.860(2.564058)\sqrt{1 + \frac{1}{10} + \frac{(2-6.0)^2}{82}}$$

$$= 5.43$$

A 90% confidence interval for y is

$$y_p - E \le y \le y_p + E$$
$$32.14 - 5.43 \le y \le 32.14 + 5.43$$
$$26.71 \le y \le 37.57$$

(g) The correlation coefficient will be positive because the points are clustered around a straight line going up from left to right.

(h) $r = \dfrac{SS_{xy}}{\sqrt{SS_x SS_y}} = \dfrac{77}{\sqrt{82(124.9)}} = 0.761$

$r^2 = (0.761)^2 = 0.579$

This means that 57.9% of the variation in y = salary can be explained by the corresponding variation in x = number of job changes using the least squares line.

(i) $H_0: \rho = 0$

$H_1: \rho > 0$

$t = \dfrac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \dfrac{0.761\sqrt{10-2}}{\sqrt{1-(0.761)^2}} = 3.318$

$d.f. = n - 2 = 10 - 2 = 8$

At 5% level of significance, $t_0 = 1.860$.

Since 3.318 > 1.860, reject H_0 and conclude that the sample evidence supports a positive correlation.

(j) $H_0: \beta = 0$

$H_1: \beta > 0$

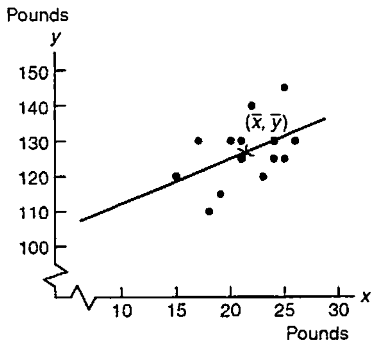
$t = \dfrac{b-\beta}{\frac{S_e}{\sqrt{SS_x}}} = \dfrac{0.939024-0}{\frac{2.564058}{\sqrt{82}}} = 3.316$

$d.f. = n - 2 = 10 - 2 = 8$

At 5% level of significance, $t_0 = 1.860$.

Since 3.16 > 1.860, reject H_0 and conclude that the sample evidence supports a positive slope.

3. (a) Weight of One-Year-Old versus Weight of Adult



(b) $\bar{x} = \dfrac{\Sigma x}{n} = \dfrac{300}{14} = 21.43$

$\bar{y} = \dfrac{\Sigma y}{n} = \dfrac{1775}{14} = 126.79$

$b = \dfrac{SS_{xy}}{SS_x} = \dfrac{184.2857}{143.4286} = 1.285$

$a = \bar{y} - b\bar{x} = 126.79 - (1.285)(21.43) = 99.25$

$y = a + bx$ or $y = 99.25 + 1.285x$

(c) See the figure in part (a).

(d) Let $x = 20$.

$y_p = 99.25 + 1.285(20) = 124.95$

The predicted weight is 124.95 pounds.

(e) $S_e = \sqrt{\frac{SS_y - bSS_{xy}}{n-2}} = \sqrt{\frac{1080.36 - 1.285(184.2857)}{14-2}} = 8.38$

(f) $E = t_c S_e \sqrt{1 + \frac{1}{n} + \frac{(x-\overline{x})^2}{SS_x}}$

$= 2.179(8.38)\sqrt{1 + \frac{1}{14} + \frac{(20-21.43)^2}{143.4286}}$

$= 19.03$

A 95% confidence interval for y is

$$y_p - E \le y \le y_p + E$$
$$124.95 - 19.03 \le y \le 124.95 + 19.03$$
$$105.92 \le y \le 143.98$$

(g) The correlation coefficient will be positive because the points are clustered around a straight line going up from left to right.

(h) $r = \frac{SS_{xy}}{\sqrt{SS_x SS_y}} = \frac{184.2857}{\sqrt{143.4286(1080.36)}} = 0.468$

$r^2 = (0.468)^2 = 0.219$

The correlation coefficient r measures the strength of the linear relationship between a woman's weight at age 1 and at age 30. The coefficient of determination r^2 means that 21.9% of the variation in y = weight of a mature adult (30 years old) can be explained by the corresponding variation in x = weight of a 1-year-old baby using the least squares line.

(i) $H_0: \rho = 0$

$H_1: \rho > 0$

$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0.468\sqrt{14-2}}{\sqrt{1-(0.468)^2}} = 1.834$

$d.f. = n - 2 = 14 - 2 = 12$

At 1% level of significance, $t_0 = 2.681$.

Since $1.834 < 2.681$, do not reject H_0. There does not seem to be any significant positive correlation at the 1% level.

(j) $H_0: \beta = 0$

$H_1: \beta > 0$

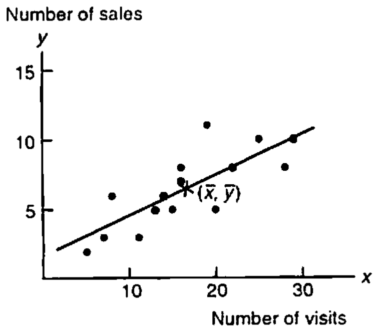
$t = \frac{b - \beta}{\frac{S_e}{\sqrt{SS_x}}} = \frac{1.285 - 0}{\frac{8.38}{\sqrt{143.4286}}} = 1.84$

$d.f. = n - 2 = 14 - 2 = 12$

At 1% level of significance, $t_0 = 2.681$.

Since $1.84 < 2.681$, we do not reject H_0. The sample evidence does not support a positive slope.

4. (a) Number of Insurance Sales and Number of Visits



(b) $\bar{x} = \frac{\sum x}{n} = \frac{248}{15} = 16.5\bar{3} \approx 16.53$

$\bar{y} = \frac{\sum y}{n} = \frac{97}{15} = 6.4\bar{6} \approx 6.47$

$b = \frac{SS_{xy}}{SS_x} = \frac{221.2\bar{6}}{755.7\bar{3}} = 0.292784$

$a = \bar{y} - b\bar{x} = 6.4\bar{6} - 0.292784(16.5\bar{3}) = 1.626$

$y = a + bx$ or $y = 1.626 + 0.293x$

(c) See the figure in part (a).

(d) Let $x = 18$.

$y_p = 1.626 + 0.293(18) = 6.9$

The predicted number of sales is 6.9 or 7.

(e) $S_e = \sqrt{\frac{SS_y - bSS_{xy}}{n-2}} = \sqrt{\frac{103.7\bar{3} - 0.292784(221.2\bar{6})}{15-2}} = 1.730940$

(f) $E = t_c S_e \sqrt{1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{SS_x}}$

$= 1.771(1.730940)\sqrt{1 + \frac{1}{15} + \frac{(18 - 16.5\bar{3})^2}{755.7\bar{3}}}$

$= 3.17$

A 90% confidence interval for y is

$$y_p - E \le y \le y_p + E$$
$$6.9 - 3.17 \le y \le 6.9 + 3.17$$
$$3.73 \le y \le 10.07$$

(g) $r = \frac{SS_{xy}}{\sqrt{SS_x SS_y}} = \frac{221.2\bar{6}}{\sqrt{755.7\bar{3}(103.7\bar{3})}} = 0.790$

$r^2 = (0.790)^2 = 0.624$

This means that 62.4% of the variation in y = number of people who bought insurance that week can be explained by the corresponding variation in x = number of visits made each week using the least squares line.

(h) $H_0: \rho = 0$
$H_1: \rho > 0$
$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0.790\sqrt{15-2}}{\sqrt{1-(0.790)^2}} = 4.65$$
$d.f. = n - 2 = 15 - 2 = 13$

At 1% level of significance, $t_0 = 2.650$.

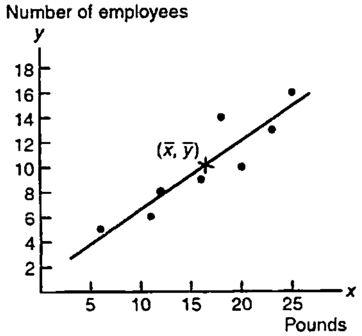
Since $4.65 > 2.650$, we reject H_0 and conclude that the sample evidence supports a positive correlation.

(i) $H_0: \beta = 0$
$H_1: \beta > 0$
$$t = \frac{b-\beta}{\frac{S_e}{\sqrt{SS_x}}} = \frac{0.292784 - 0}{\frac{1.730940}{\sqrt{755.73}}} = 4.65$$
$d.f. = n - 2 = 15 - 2 = 13$

At 1% level of significance, $t_0 = 2.650$.

Since $4.65 > 2.650$, we reject H_0 and conclude that the sample evidence supports a positive slope.

5. (a)



(b) $\bar{x} = \frac{\sum x}{n} = \frac{131}{8} = 16.375 \approx 16.38$
$\bar{y} = \frac{\sum y}{n} = \frac{81}{8} = 10.125 \approx 10.13$
$b = \frac{SS_{xy}}{SS_x} = \frac{160.625}{289.875} = 0.554118 \approx 0.554$
$a = \bar{y} - b\bar{x} = 10.125 - 0.554118(16.375) = 1.051$
$y = a + bx$ or $y = 1.051 + 0.544x$

(c) See the figure in part (a).

(d) Use $x = 15$.
$y_p = 1.051 + 0.544(15) = 9.36$
About 9 or 10 employees should be assigned mail duty.

(e) $$S_e = \sqrt{\frac{SS_y - bSS_{xy}}{n-2}} = \sqrt{\frac{106.875 - (0.554118)(160.625)}{8-2}} = 1.73$$

(f) $E = t_c S_e \sqrt{1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{SS_x}}$

$= 2.447(1.73)\sqrt{1 + \frac{1}{8} + \frac{(15 - 16.375)^2}{289.875}}$

$= 4.5$

A 95% confidence interval for y is

$$y_p - E \le y \le y_p + E$$
$$9.36 - 4.5 \le y \le 9.36 + 4.5$$
$$4.86 \le y \le 13.86$$

(g) $r = \frac{SS_{xy}}{\sqrt{SS_x SS_y}} = \frac{160.625}{\sqrt{289.875(106.875)}} = 0.913$

$r^2 = (0.913)^2 = 0.834$

The correlation coefficient r measures the strength of the linear association between weight of incoming mail and number of employees assigned to answer it. The coefficient of determination, r^2, means that 83.4% of the variation in y = number of employees can be explained by the corresponding variation in x = weight of incoming mail using the least squares line.

(h) $H_0: \rho = 0$

$H_1: \rho > 0$

$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0.913\sqrt{8-2}}{\sqrt{1-(0.913)^2}} = 5.48$

$d.f. = n - 2 = 8 - 2 = 6$

At 1% level of significance, $t_0 = 3.143$.

Since 5.48 > 3143, we reject H_0 and conclude that the sample evidence supports a positive correlation coefficient.

(i) $H_0: \beta = 0$

$H_1: \beta > 0$

$t = \frac{b - \beta}{\frac{S_e}{\sqrt{SS_x}}} = \frac{0.554118 - 0}{\frac{1.73}{\sqrt{289.875}}} = 5.45$

$d.f. = n - 2 = 8 - 2 = 6$

At 1% level of significance, $t_0 = 3.143$.

Since 5.45 > 3.143, we reject H_0 and conclude that the sample evidence supports a positive slope.

(j) $d.f. = 6, t_c = 1.440, b \approx 0.554$

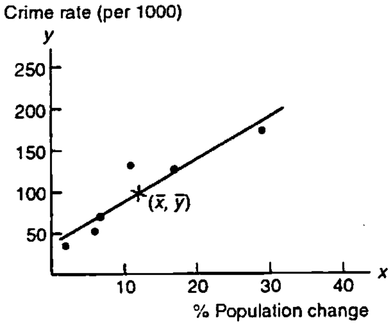
$E = t_c \frac{S_e}{\sqrt{SS_x}} = 1.440 \frac{1.73}{\sqrt{289.875}} = 0.146$

An 80% confidence interval is

$$b - E < \beta < b + E$$
$$0.554 - 0.146 < \beta < 0.554 + 0.146$$
$$0.41 < \beta < 0.70$$

For each additional pound of mail, assign 1 employee to work from 41% to 70% of a work day on mail.

6. (a) Percent Population Change



(b) $\bar{x} = \frac{\sum x}{n} = \frac{72}{6} = 12.0$

$\bar{y} = \frac{\sum y}{n} = \frac{589}{6} = 98.1\bar{6} \approx 98.17$

$b = \frac{SS_{xy}}{SS_x} = \frac{2431}{476} = 5.1071 \approx 5.11$

$a = \bar{y} - b\bar{x} = 98.1\bar{6} - 5.1071(12.0) = 36.881 \approx 36.9$

$y = a + bx$ or $y = 36.9 + 5.11x$

See the figure in part (a).

(c) Let $x = 12$

$y = 36.9 + 5.11(12) = 98.2$

The predicted crime rate is 98.2 crimes per thousand.

(d) $S_e = \sqrt{\frac{SS_y - bSS_{xy}}{n-2}} = \sqrt{\frac{14456.8\bar{3} - 5.1071(2431)}{6-2}} = 22.59$

$$E = t_c S_e \sqrt{1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{SS_x}}$$

$$= 1.533(22.59)\sqrt{1 + \frac{1}{6} + \frac{(12-12)^2}{476}} = 37.41$$

A 80% confidence interval for y is

$$y_p - E \le y \le y_p + E$$
$$98.2 - 37.41 \le y \le 98.2 + 37.41$$
$$60.8 \le y \le 135.6$$

or about 61 to 136 crimes per thousand.

(e) $r = \frac{SS_{xy}}{\sqrt{SS_x SS_y}} = \frac{2431}{\sqrt{476(14456.8\bar{3})}} = 0.927$

$r^2 = (0.927)^2 = 0.859$

$H_0: \rho = 0$

$H_1: \rho \ne 0$

$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0.927\sqrt{6-2}}{\sqrt{1-(0.927)^2}} = 4.94$

$d.f. = n - 2 = 6 - 2 = 4$

At 1% level of significance, $t_0 = \pm 4.604$.

Since $4.94 > 4.604$, we reject H_0 and conclude that the sample evidence supports a significant correlation coefficient.

(f) High correlation does not guarantee a "cause-and-effect" situation. Before causation is estabilished, more work needs to be done taking other variables into account.
High correlation is simply an indication of a mathematical relationship between variables.

(g) H_0: $\beta = 0$
H_1: $\beta > 0$

$$t = \frac{b-\beta}{\frac{S_e}{\sqrt{SS_x}}} = \frac{5.1071-0}{\frac{22.59}{\sqrt{476}}} = 4.93$$

$d.f. = n-2 = 6-2 = 4$

At 1% level of significance, $t_0 = 3.747$.

Since 4.93 > 3.747, we reject H_0 and conclude that the sample evidence supports a positive slope.

(h) $d.f. = 4, t_c = 1.533, b \approx 5.11$

$$E = t_c \frac{S_e}{\sqrt{SS_x}} = 1.533 \frac{22.59}{\sqrt{476}} = 1.59$$

An 80% confidence interval is

$$b - E < \beta < b + E$$
$$5.11 - 1.59 < \beta < 5.11 + 1.59$$
$$3.52 < \beta < 6.70$$

For every percentage point increase in population, expect the crime rate per 1000 to increase from between 3.52 to 6.70 crimes per thousand.