Solutions to Practice Problems

Multiple Choice

1. The correct answer is (d).

$$b = r.\frac{s_y}{s_x} = (0.55)\left(\frac{0.75}{1.6}\right) = 0.26$$

2. The correct answer is (e). The value of a residual = actual value − predicted value = 25 − [2.35 + 0.86(29)] = −2.29.

3. The correct answer is (a). $r^2 = (-0.58)^2 = 0.3364$. This is the *coefficient of determination,* which is the proportion of the variation in the response variable that is explained by the regression on the independent variable. Thus, about one-third (33.3%) of the variation in hours spent exercising can be explained by hours spent watching television. (b) is incorrect since correlation does not imply causation. (c) would be correct if $b = -0.58$, but there is no obvious way to predict the response value from the explanatory value just by knowing r. (d) is incorrect for the same reason (b) is incorrect. (e) is incorrect since r, not r^2, is given. In this case $r^2 = 0.3364$, which makes (a) correct.

4. The correct answer is (c). $ln(y) = 1.64 - 0.88(3.1) = -1.088 \Rightarrow y = e^{-1.088} = 0.337$.

5. The correct answer is (c). The pattern is more or less random about 0, which indicates that a line would be a good model for the data. If the data are linearly related, we would expect them to have a non-zero correlation.

6. The correct answer is (b). I is incorrect—the *predicted* weight of a person 61 inches tall is 104.6 pounds. II is a correct interpretation of the slope of the regression line (you could also say that "For each additional inch of Height, Weight *is predicted to* increase by 3.6 pounds). III is incorrect. It may well be true, but we have no way of knowing that from the information given.

7. The correct answer is (c). The predicted score for the student is 273.5 + (91.2)(3) = 547.1. The residual is the actual score minus the predicted score, which equals 510 − 547.1 = −37.1.

8. The correct answer is (b). Consider the expression for r. $r = \frac{1}{n-1}\sum\left(\frac{x-\bar{x}}{s_x}\right)\left(\frac{y-\bar{y}}{s_y}\right)$.

   Adding 12 to each Y-value would not change s_y. Although the average would be 12 larger, the differences $y - \bar{y}$ would stay the same since each Y-value is also 12 larger. By taking the negative of each X-value, each term $\frac{x-\bar{x}}{s_x}$ would reverse sign (the mean also reverses sign) but the absolute value of each term would be the same. The net effect is to leave unchanged the absolute value of r but to reverse the sign.

9. The correct answer is (e). The question is asking for the coefficient of determination, r^2 (R-sq on many computer printouts). In this case, $r = 0.8877$ and $r^2 = 0.7881$, or 78.8%. This can be found on your calculator by entering the GPA scores in `L1`, the SAT scores in `L2`, and doing `STAT CALC 1-Var Stats L1,L2`.

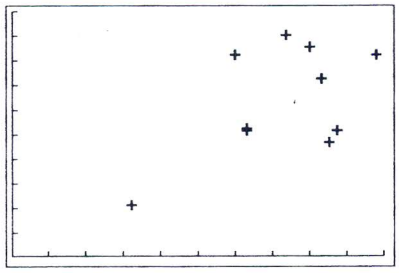
10. The correct answer is (a). The point $(\bar{x}, \bar{y})$ always lies on the LSRL. Hence, $\bar{y}$ can be found by simply substituting $\bar{x}$ into the LSRL and solving for $\bar{y}$. Thus $\bar{y} = 32.5 - 0.45(29.8) = 19.09$ mpg. Be careful: you are told that the equation uses the weights in hundreds of pounds. You must then substitute 29.8 into the regression equation, not 2980, which would get you answer (c).

Free Response

1. $b = r\frac{s_y}{s_x} = (0.80)\left(\frac{11}{4}\right) = 2.2,\ a = \bar{y} - b\bar{x} = 20 - (2.2)(14.5) = -11.9.$

   Thus, $\hat{y} = -11.9 + 2.2x$.

2. (a)

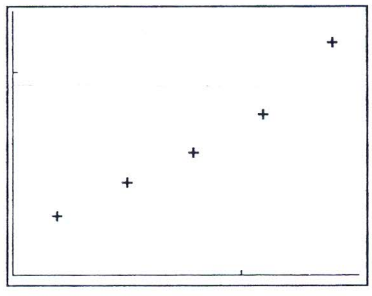


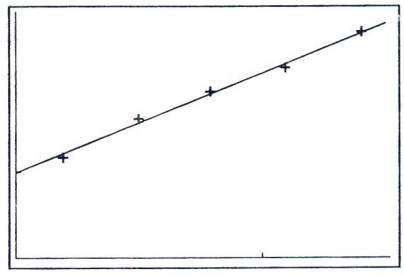
   (b) There seems to be a moderate positive relationship between the scores: students who did better on the first test tend to do better on the second, but the relationship isn't very strong; $r = 0.55$.

3. A line is not a good model for the data because the residual plot shows a definite pattern: the first 8 points have negative residuals and the last 8 points have positive residuals. The box is in a cluster of points with positive residuals. We know that, for any given point, the residual equals actual value minus predicted value. Because actual – predicted > 0, we have actual > predicted, so that the regression equation is likely to underestimate the actual value.

4. The regression equation for predicting time from year is *time* = 79.21 – 0.61(*year*). We need *time* = 60. Solving 60 = 79.1 – 0.61(*year*), we get *year* = 31.3. So, we would predict that times will drop under one minute in about 31 or 32 years. The problem with this is that we are extrapolating far beyond the data. Extrapolation is dangerous in any circumstance, and especially so 24 years beyond the last known time. It's likely that the rate of improvement will decrease over time.

5. A scatterplot of the data (graph on the left) appears to be exponential. Taking the natural logarithm of each *y*-value, the scatterplot (graph on the right) appears to be more linear.





Taking the natural logarithm of each y-value and finding the LSRL, we have ln(#Roaches) = 0.914 + 0.108 (Days) = 0.914 + 0.108(9) = 1.89. Then #Roaches = $e^{1.89}$ = 6.62.

6. The correlation between walking more and better health may or may not be causal. It may be that people who are healthier walk more. It may be that some other variable, such as general health consciousness, results in walking more and in better health. There may be a causal association, but in general, correlation is not causation.

7. Carla has reported the value of r^2, the coefficient of determination. If she had predicted each girl's grade based on the average grade only, there would have been a large amount of variability. But, by considering the regression of grades on socioeconomic status, she has reduced the total amount of variability by 72%. Because $r^2 = 0.72$, $r = 0.85$, which is indicative of a strong positive linear relationship between grades and socioeconomic status. Carla has reason to be happy.

8. (a) is false. $\Sigma(y - \hat{y}) = 0$ for the LSRL, but there is no unique line for which this is true.
   (b) is true.
   (c) is true. In fact, this is the definition of the LSRL—it is the line that minimizes the sum of the squared residuals.
   (d) is true since $b = r\frac{s_y}{s_x}$ and $\frac{s_y}{s_x}$ is constant.
   (e) is false. The slope of the regression lines tell you by how much the response variable changes *on average* for each unit change in the explanatory variable.

9. $\hat{y} = 26.211 - 0.25x = 26.211 - 0.25(73) = 7.961$. The residual for $x = 73$ is the actual value at 73 minus the predicted value at 73, or $y - \hat{y} = 7.9 - 7.961 = -0.061$. (73, 7.9) is below the LSRL since $y - \hat{y} < 0 \Rightarrow y < \hat{y}$.

10. (a) $r = +0.75$; the slope is positive and is the opposite of the original slope.
    (b) $r = -0.75$. It doesn't matter which variable is called x and which is called y.
    (c) $r = -0.75$; the slope is the same as the original slope.

11. We know that $b = r\frac{s_y}{s_x}$, so that $2.7 = r(3.33) \rightarrow r = \frac{2.7}{3.33} = 0.81 \rightarrow r^2 = 0.66$. The proportion of the variability that is *not* explained by the regression of y on x is $1 - r^2 = 1 - 0.66 = 0.34$.

12. Because the linear pattern will be stronger, the correlation coefficient will increase. The influential point pulls up on the regression line so that its removal would cause the slope of the regression line to decrease.

13. (a) *rate* = −0.3980 + 0.1183 (*number*).
    (b) $r = \sqrt{0.974} = 0.987$ (r is positive since the slope is positive).
    (c) *rate* = −0.3980 + 0.1183(20) = 1.97 crimes per thousand employees. Be sure to use 20, not 200.

14. (a) *Percentage appreciation* = 1.897 + 0.115 (*number*)
    (b) *Percentage appreciation* = 1.897 + 0.115(85) = 11.67%.
    (c) $r = 0.82$, which indicates a strong linear relationship between the number of new homes built and percent appreciation.
    (d) If the number of new homes built was unknown, your best estimate would be the average percentage appreciation for the 5 years. In this case, the average percentage appreciation is 11.3%. [For what it's worth, the average error (absolute value) using the mean to estimate appreciation is 2.3; for the regression line, it's 1.3.]

15. (a) If $r^2 = 0.81$, then $r = \pm 0.9$. The slope of the regression line for the standardized data is either 0.9 or −0.9.

(b) If $r = +0.9$, the scatterplot shows a strong positive linear pattern between the variables. Values above the mean on one variable tend to be above the mean on the other, and values below the mean on one variable tend to be below the mean on the other. If $r = -0.9$, there is a strong negative linear pattern to the data. Values above the mean on one variable are associated with values below the mean on the other.

16. (a) $r = 0.8$
(b) $r = 0.0$
(c) $r = -1.0$
(d) $r = -0.5$

17. Each of the points lies on the regression line → every residual is 0 → the sum of the squared residuals is 0.

18. (a) $r = 0.90$ for these data, indicating that there is a strong positive linear relationship between student averages and evaluations of Prof. Socrates. Furthermore, $r^2 = 0.82$, which means that most of the variability in student evaluations can be explained by the regression of student evaluations on student average.
(b) If y is the evaluation score of Prof. Socrates and x is the corresponding average for the student who gave the evaluation, then $\hat{y} = -29.3 + 1.34x$. If $x = 80$, then $\hat{y} = -29.3 + 1.34(80) = 77.9$, or 78.

19. (a) True, because
$b = r\frac{s_y}{s_x}$ and $\frac{s_y}{s_x}$ is positive.
(b) True. r is the same if explanatory and response variables are reversed. This is not true, however, for the slope of the regression line.
(c) False. Because r is defined in terms of the means of the x and y variables, it is not resistant.
(d) False. r does not depend on the units of measurement.
(e) True. The definition of r, $r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$,
necessitates that the variables be numerical, not categorical.

20. (a) *Left-hand strength* = 7.1 + 0.35(12) = 11.3 kg.
(b) **Intercept:** The predicted left-hand strength of a person who has zero right-hand strength is 7.1 kg.
Slope: On average, left-hand strength increases by 0.35 kg for each 1 kg increase in right-hand strength. Or left-hand strength is predicted to increase by 0.35 kg for each 1 kg increase in right-hand strength.

Solutions to Cumulative Review Problems

1. A *statistic* is a measurement that describes a sample. A *parameter* is a value that describes a population.

2. FALSE. For an interval of fixed length, there will be a greater proportion of the area under the normal curve if the interval is closer to the center than if it is removed from the center. This is because the normal distribution is mound shaped, which implies that the terms tend to group more in the center of the distribution than away from the center.