Lesson 2.2.2

2-55. a. Find mean and s.d. of both variables, standardize the points, multiply each standardized x by its standardized y, add up all the products, then divide by $n - 1$.

b. No. Since the standardized values are multiplied, which is commutative, this formula does not care. Thus, r will stay the same if you flip the variables.

c. You have to standardize the original data to calculate r, so of course they are the same.

d. Since the data is standardized before calculating r, units are divided out and do not matter.

e. Only the strength of the association, using whatever is the most appropriate scale on the graph. The order of variables and units do not matter. If one or both variables are linearly transformed, that also does not matter because the scatterplot stays the same so the correlation does as well. Linear transformations (with a positive slope) of one or both variables will not change the visual strength, direction, or form of the relationship.

2-56. a. The equation of a regression line is $y = a + bx$. The line being considered is standardized, so substitution gives $z_y = a + bz_x$. r is given as the slope and the line passes through $(0, 0)$, making the y-intercept 0. Substitution gives $z_y = 0 + rz_x$ or $z_y = rz_x$.

b. $\frac{y-\overline{y}}{s_y} = r\left(\frac{x-\overline{x}}{s_x}\right) \rightarrow y - \overline{y} = r\left(\frac{s_y}{s_x}\right)(x - \overline{x}) \rightarrow y = r\frac{s_y}{s_x}x + \overline{y} - r\frac{s_y}{s_x}\overline{x}$. Therefore the slope is $b = r\frac{s_y}{s_x}$.

c. $(\overline{x}, \overline{y})$

d. y-intercept $a = \overline{y} - b\overline{x}$ (where b is the slope).

2-57. a. positive, moderately weak: 0.567 is not a strong correlation.

b. slope $= r \cdot \frac{s_y}{s_x} = 0.567 \cdot \frac{8.684}{0.676} = 7.284$ points/hour.

y-intercept $= 79.33 - 7.284 \cdot 2.423 = 61.7$. The slope means that for each additional hour preparing, the LSRL predicts a score increase of 7.28 points. The y-intercept means that a student who prepares 0 hours is expected to score 61.7% by the LSRL.

c. Her expected score is a 90.8 and she earns an 85, so her residual is –5.8 points. In other words she scores about 5.8 points lower than expected by the LSRL model based on her preparation time alone.

2-58. a. $z_y = mz_x$

b. $\text{SSR} = \sum (z_{yi} - mz_{xi})^2$

c. $\text{SSR} = \sum ({z_{yi}}^2 - 2mz_{xi}z_{yi} + m^2{z_{xi}}^2)$

d. $\text{SSR} = \sum {z_{yi}}^2 - 2m\sum z_{xi}z_{yi} + m^2\sum {z_{xi}}^2$

e. 1: in a standardized data set the variance and standard deviation are 1.

f. $\frac{\text{SSR}}{n-1} = \frac{\sum {z_{yi}}^2}{n-1} - \frac{2m\sum z_{xi}z_{yi}}{n-1} + \frac{m^2\sum {z_{xi}}^2}{n-1} = 1 - \frac{2m\sum x_{zi}y_{zi}}{n-1} + m^2$

g. r, the correlation coefficient.

h. $\frac{\text{SSR}}{n-1} = m^2 - 2mr + 1$

i. A parabola.

j. Completing the square creates the formula $(m - r)^2 + (1 - r^2) = \frac{\text{SSR}}{n-1}$, so the vertex is $(r, 1 - r^2)$. You could also use the fact that the vertex occurs at $m = -\frac{b}{2a}$ (in standard form) $= \frac{2r}{r} = r$ and substitute to find the point: $(r, 1 - r^2)$.

k. Slope = r, the correlation coefficient.

2-59. a. There is a very strong positive linear association between the differences in ratings and the corresponding difference of game scores, as evidenced by the r very near 1. There are no apparent outliers. The slope is about 1.6. An increase of 1 in the difference of ratings is expected to increase the score difference by 1.6 points.

b. The LSRL is about $\hat{y} = 1.6x$. The expected difference is –3.74. Marin Catholic is expected to win by about 4 goals.

c. The model predicts the difference in scores to be $\hat{y} = 1.6(5) = 8$ points. Since residual = actual – predicted, 7 = actual – 8. The actual difference in scores was 15 points.

d. $b = r\,\frac{s_y}{s_x}$; $1.6 = 0.9\left(\frac{5.2}{S_x}\right)$, so $S_x \approx 2.93$

2-60. a. Increase; slope represents "change in height per shoe size change" so if height changes from inches to cm, there will be a more significant height change with each shoe size value increase.

b. increase, due to the unit change

c. stay the same; correlation is unitless

d. Increase, since residuals are measured in the unit of the response variable.

e. Decrease, since it will be measuring change-in-shoe-size per inch. A reasonable estimate $\approx \frac{1}{1.5}$, or 0.67, but the actual value is 0.482 because the slopes of the two lines must have a product of r^2.

f. Decrease; since the y-intercept will mean the "shoe size of a 0-inch person", which will likely be a negative number.

g. stay the same

2-61. a. never

b. Sometimes. If the point fits the existing pattern, it increases. A highly influential point (often with an extreme x-coordinate) can also drastically change r.

c. Sometimes. If the new point is exactly on the existing LSRL, it will stay the same. Otherwise, it will change.

d. Sometimes, if the datasets are measuring the same thing in the same units.

e. Always. The correlation coefficient is calculated from standardized data.

f. Sometimes. This is a trick question; *if the forms of both associations are linear* then this is always true, but it is possible to form datasets with $r = 0.8$ that are curved. In those cases, r should be treated as fairly meaningless, and so a dataset with $r = 0.6$ and a linear form would have a stronger linear association.

2-62. a. Answers may vary but the histogram at right is a reasonable representation based on shape and scale.

b. Continuous. Like most measurements, the number of significant digits is probably only limited by the instrument used to measure the carbon content.