Lesson 2.2.5

2-84. a. LSRL: $h = 131.54 + 4.01s$. $r = 0.763$. mean height = 162.4 cm

b. 162.4 cm, the mean

c. 167.6 cm

d. Kerin's guess using the LSRL was about 59% better than using the mean. Logic: she was off by 3.7 cm using the mean, but only 1.5 centimeters using the LSRL. Thus, she improved by 2.2 cm by using the LSRL, and $\frac{2.2\text{ cm}}{3.7\text{ cm}} \approx 59\%$

2-85. a. You could use the standard deviation of the residuals (s) or you could use the sum of the absolute residuals or the sum of the squared residuals.

b. You could use the variance, or the standard deviation, or the sum of the square differences.

c. SSR = 94.916

d. SST = 226.9

e. 58.16%. Using the same technique as part (d) of problem 2-74: $\frac{226.9-94.916}{226.9} = 0.5816$.

f. 58.16% of the variation in height can be explained by the linear association between height and shoe size.

g. $0.763^2 = 0.5816$

2-86. a. $r = 1$; There is a perfectly linear positive association between shoe size and height.

b. $R^2 = 100\%$. 100% of the variability in height can be explained by a linear relationship with shoe size. There is no scatter and no variability in the y-values away from the linear relationship. If she knows the shoe size, Alyse can predict the height perfectly.

c. $S = 0$, meaning the average difference between the actual values and predicted values is 0 cm. This is simply another way of stating the prediction is perfect in the units of the problem.

2-87. a. $r \approx 0$; $R^2 \approx 0\%$; None (0%) of the variability in height can be explained by a linear relationship with test score. There is completely random scatter, so there is no way to predict height based on a relationship with test score.

b. $h = 162$. Our predicted height is 162 cm (the mean) no matter what the test score is!

2-88. a. $y = 7.74 + 1.36x$; For zero toppings the predicted cost is $7.74. That is the cost of crust, cheese, sauce, the restaurant equipment, labor, and profit.

b. $r = 0.86$; $R^2 = 74\%$

c. In problem 2-66 students checked the residual plot and it confirmed a linear *form*. The *direction* is positive with a slope of 1.36: an increase in one topping is predicted to increase the cost of the pizza by $1.36. The association is moderately *strong*: 74% of the variability in cost can be explained by a linear relationship with the number of toppings. As an aside, the other 26% of the variability can be explained by variables such as rent, level of service (white tablecloths or take-out), quality of ingredients, and profit margin. There are no apparent outliers.

2-89. a. $y = 7 + 1.5x$; the slope indicates that an increase of one topping is predicted to increase the cost by $1.50; the *y*-intercept indicated that a pizza with no toppings is expected to cost $7.00.

b. $r = 1$; $R^2 = 100\%$; 100% of the variability in the cost of pizza can be explained by a linear relationship with number of toppings. At a single pizza parlor, parlor-to-parlor variability (rent, service, quality) is no longer present. Note: Since these data points are clearly not independent, an essential condition for calculating the correlation coefficient has not been met; *r* should not have been calculated in the first place. Nonetheless, this problem makes a valuable point about the interpretation of *r* and R^2.

2-90. There is a moderately strong negative association, meaning that the more a student watches TV, the lower his/her grade point average is predicted to be. 52% of the variability in GPAs can be explained by a linear relationship with hours of TV. However be careful not to imply a cause. Watching less TV will not necessarily cause a rise in GPA.

2-91. From the previous lesson, association does not mean that one variable caused the other. One possible confounding variable is per capita wealth: wealthier nations may tend to have more TVs and also better health care.

2-92. $r \approx 0$; Answers will vary for the LSRL, but the average number of pairs appears to be about 3.8 which is an LSRL of $y = 3.8$.

2-93. a. With a car readily available these teens might simply be <u>driving more</u> and the extra time on the road is causing them to be in more crashes.

b. Families which can afford the considerable expense of bottled water can also afford better nutrition and better health care.

2-94. a. The *y*-intercept of 18.71 means that a dough with no cost would take 18.7 minutes to prepare. This is nonsensical; all dough has a cost. The slope of about –4 means that for each additional dollar spent per pound of dough on the recipe, the preparation time is reduced by about 4 minutes.

b. *r* is –0.9707: make sure you got the negative from the slope! The *S* value of 0.033 means that the typical preparation time is about 0.033 minutes (about 2 seconds!) away from the time predicted by the model. The R^2 of 94.24% means that 94.24% of the variation in the preparation time is predicted by the linear association with the cost of the dough. The value of *r* represents a strong, negative correlation between preparation time and cost.

c. The residual plot is fairly randomly scattered, which implies the form of this association is approximately linear.

2-95. a. A histogram makes it easy to see when many of the students have nearly the same score and gives a good estimate of the shape and range of the distribution. It does not show exact data values making it harder to find the median value and the quartiles. A histogram has an advantage if the number of values in the data set is very large, which does not occur here with a typical class size.

b. An advantage of a stem-and-leaf plot is that the range can easily be found and exact data values can be seen. A disadvantage is that a stem-and-leaf plot can become unreasonable if too many numbers are represented or if the numbers have too much variability. Neither of these are a concern in this setting.

c. A boxplot specifically shows the range, median, highest and lowest values, and quartiles. It does not show individual scores or possible outliers if there is more than one. It does not show the exact data values and the shape of the distribution can only be estimated. It also does not indicate how many scores are represented.