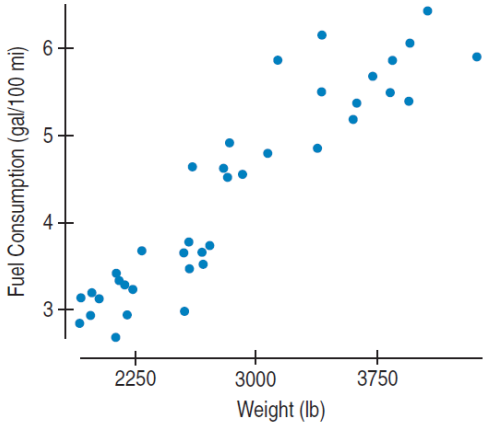
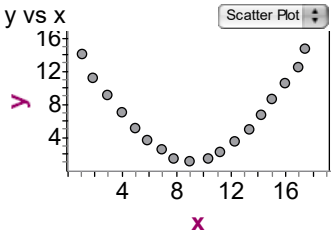
AP STATISTICS **Name:** _____ **Period:** ____
Review Unit IV – Scatterplots & Regressions

Know the definitions of the following terms: correlation coefficient, slope, residual, outlier, coefficient of determination, influential observation, extrapolation, explanatory variable, response variable. Read chapters 7, 8, and 9.
Note: You do **NOT** need to know how to create a **RESIDUAL PLOT** on the calculator (nor by hand). Just make sure you know how to interpret a residual plot.

1. Here is the scatterplot of weight (in pounds) versus fuel consumption (in gallons used per 100 miles driven) for a random sample of 38 automobiles. Describe the association between weight and fuel consumption for these automobiles. *(Remember: Describe strength, form, direction, and use context!)*



2. The scatterplot shows a relationship between x and y that results in a correlation coefficient of r = 0.024. Explain why $r = 0.024$ in this situation even though there appears to be a strong relationship between the x and y variables.

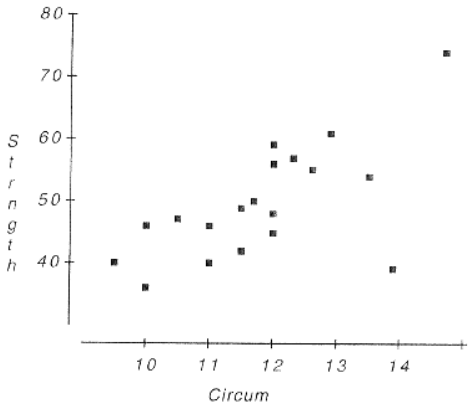


3. The linear model for the association between fat grams and protein grams in Burger King's menu items is

$$\widehat{fat\ grams} = 6.8 + 0.97(protein\ grams)$$

What is the estimated *increase* in fat grams that corresponds to an *increase* of 10 grams of protein?

4. Researchers investigating the association between the size and strength of muscles measured the forearm circumference (in inches) of 20 teenage boys. Then they measured the strength of the boys' grips (in lbs). Their data are plotted at right.



a) Describe the *association* between forearm circumference and strength.

b) If the point at the lower right corner (at about 14" and 38 lbs.) were removed, would the correlation become stronger, weaker, or remain about the same?

c) If the point at the lower right corner (at about 14" and 38 lbs.) were removed, would the slope of the least squares regression line become more negative, more positive, or stay the same?

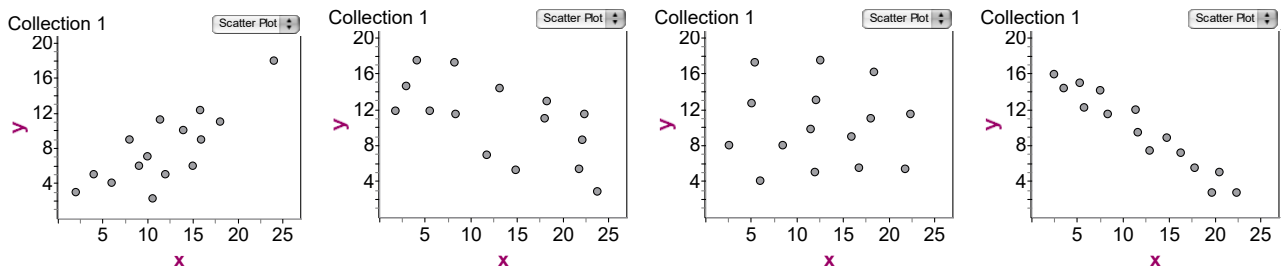
5. **Storks** Data show that there is a positive association between the population of 17 European countries and the number of stork pairs in those countries.

a) Briefly explain what "positive association" means in this context.

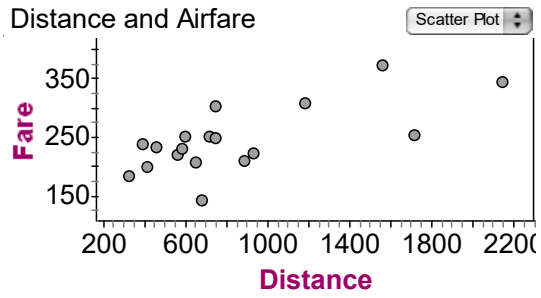
b) Wildlife advocates want the stork population to grow, so they approach the governments of these countries to encourage their citizens to have children. As a statistician, what do you think of this plan? Explain briefly.

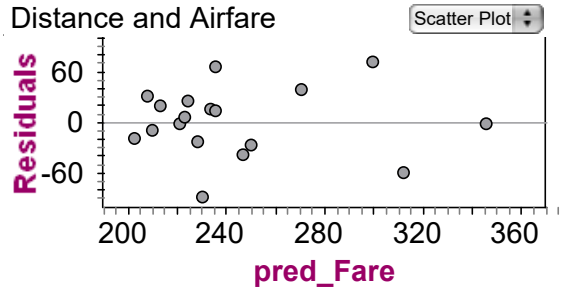
6. For the following scatterplots, write the appropriate correlation coefficient underneath each plot.

-0.6112 0 .7994 -0.9713 0.2005 0.0023 -1.0



7. The distance (in miles) and airfare (in dollars) from Atlanta to 18 major cities was gathered. A scatterplot, a residual plot, and the computer output from a regression analysis are shown:





Variable	Coef	S.E. of Coef	t-ratio	P
Constant	177.215	19.99	8.86	.0001
dist	0.079	0.0204	3.86	.0014

s = 41.9 R-sq = 48% R-sq adjusted = 49.8%

a) Is a linear model appropriate for this data? Explain.

b) State the least squares regression line equation that summarizes the relationship between the distance a plane travels and the airfare. Define any variables used in this equation or state the equation *in context.*

c) Interpret the meaning of the slope of the regression line *in context*.

d) If appropriate, interpret the meaning of the y-intercept. If not appropriate, explain why not.

e) Interpret R^2 in context.

f) Interpret s_e in this context.

g) Predict the airfare if the plane travels 100 miles. Comment on your prediction.

8. **Car commercials** A car dealer investigated the association between the number of TV commercials he ran each week and the number of cars he sold the following weekend. He found the correlation to be $r = 0.56$. During the time he collected the data he ran an average of 12.4 commercials a week with a standard deviation of 1.8, and sold an average of 30.5 cars with a standard deviation of 4.2. Next weekend he is planning a sale, hoping to sell 40 cars.

a) Write an equation of the linear model to estimate the number of cars he might sell on a certain weekend based on the number of TV commercials run that week. Define any variable used in this equation or state the equation *in context*. *(Show your work as well as your formulas…* ***BUT ONLY IF YOU WANT ANY CREDIT!)***

b) If the car dealer decides to pay for 18 TV commercials, how many cars might he expect to sell the following weekend?

c) Let us suppose that in a particular week, the car dealer paid for 10 commercials and sold 22 cars. Calculate the residual for the number of cars sold that weekend, and interpret this value in context.

d) Suppose that the dealer pays for a certain number of commercials, and the resulting value of the residual is negative. Does this mean that the model OVERestimated or UNDERestimated the actual number of cars sold?

9. The following table gives information on the temperature in a city and the volume of ice cream (in pounds) sold at an ice cream parlor for a random sample of eight days during the summer of 1993.

Temperature	93	86	77	89	98	102	87	79
Ice Cream Sold	187	169	123	198	232	267	158	117

a) Create a scatterplot. Find the correlation coefficient in context, and describe the association between the two variables.

b) Write the least squares regression equation in context.

c) Identify and interpret the coefficient of determination in context.

d) Predict how much ice cream will be sold if the temperature is 83 degrees.

e) On a 95 degree day, the residual was 6.214. How many pounds of ice cream were sold?

Multiple Choice Questions

10. Which of the following associations is likely to have a negative correlation?

    A) Number of hours devoted to studying for a final exam and a student's grade on the final
    B) A teacher's salary and the number of years teaching experience that the teacher has
    C) The age of an automobile and the number of miles an automobile has
    D) The number of children in a family and the weekly amount of money spent on food.
    E) The speed a car travels and the time required to travel a given distance on a flat deserted road

11. Residuals (or errors) in a regression analysis are defined as

    A) $\hat{y}$ B) $y - \hat{y}$ C) $\hat{y} - y$ D) $(y - \hat{y})^2$

12. The value of the coefficient of determination is always in the range

    A) 0 to 1 B) –1 to 1 C) –1 to 0 D) $-\infty, \infty$

13. The value of the correlation coefficient is always in the range

    A) 0 to 1 B) –1 to 1 C) –1 to 0 D) $-\infty, \infty$

14. Which of the following statements about correlation is a valid statement?

    A) The correlation between height and weight is 0.568 inches per pound.
    B) The correlation between weight and length of foot is 0.488.
    C) The correlation between the breed of a dog and its weight is 0.435.
    D) The correlation between gender and age is -0.171.

15. Which of the following statements concerning residuals in a least square regression line is true?

    A) The sum of the residuals is 0.
    B) A plot of the residuals is useful for assessing the fit of the least-squares regression line.
    C) The value of a residual is the observed value of the response minus the value of the response that one would predict from the least-squares regression line.
    D) All of the above.

16. Which of the following is <u>not true</u> of the correlation coefficient of a set of bivariate data (or are all the statements true)?

    A) The higher the correlation coefficient, the steeper the line of best fit.
    B) A correlation coefficient close to zero does not necessarily indicate a weak relationship between the variables.
    C) The correlation coefficient is not a resistant measure of association.
    D) Two sets of bivariate data can have approximately equal correlation coefficients but very different scatterplots.
    E) All of these are true.

17. Which of the following is true of the least-squares regression line?

    A) The slope is the change in the response variable that would be predicted by a one unit increase in the explanatory variable.
    B) It rarely passes through the point $(\overline{x}, \overline{y})$ where $\overline{x}$ and $\overline{y}$ are the means of the explanatory and response variables, respectively.
    C) It will sometimes pass through all the data points if r = -1 or +1
    D) All of the above

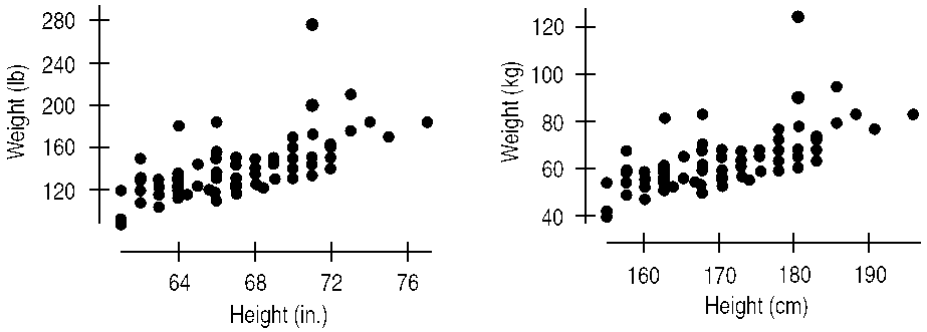
18. When using midterm exam scores to predict a student's final grade in a class, one would prefer to have a

A) positive residual, because that means the student's final grade is higher than we would predict with the model.
B) positive residual, because that means the student's final grade is lower than we would predict with the model.
C) residual equal to zero because that means the student's final grade is exactly what we would predict with the model.
D) negative residual, because that means the student's final grade is higher than we would predict with the model.

19. Which one of the following statements is true?

A) Values of *r* near zero indicate a strong linear relationship.
B) The correlation can be strongly affected by a few outlying observations.
C) Changing the measurement units of x and y may affect the correlation between x and y.
D) Strong correlation means that there is a definite cause-and-effect relationship between x and y
E) Correlation changes when the x and y variables are reversed.

20. Data collected from students in Statistics classes included their height and weight. Originally, the weight of each student was measured in pounds, and height was measured in inches. However, as shown in the scatterplots below, each student's weight is converted from pounds to kilograms, and each student's height is also converted from inches to centimeters:



Which of the following statements are true? *(You may wish to refer to page 147 in your textbook)*

I. The value of "*r*" will stay the same.
II. The z-score for each student's weight and height measurements will stay the same.
III. The slope ("rise over run") of the regression line (for weight vs. height) will stay the same.

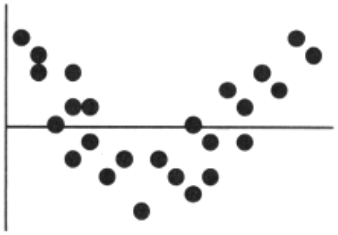
A) I only B) I and II only C) I and III only D) II and III only E) I, II, and III

21. The correlation between *X* and *Y* is $r = 0.35$. If we double each *X* value, increase each *Y* value by 0.20, and interchange the variables (put *X* on the *Y-axis* and vice versa), the new correlation

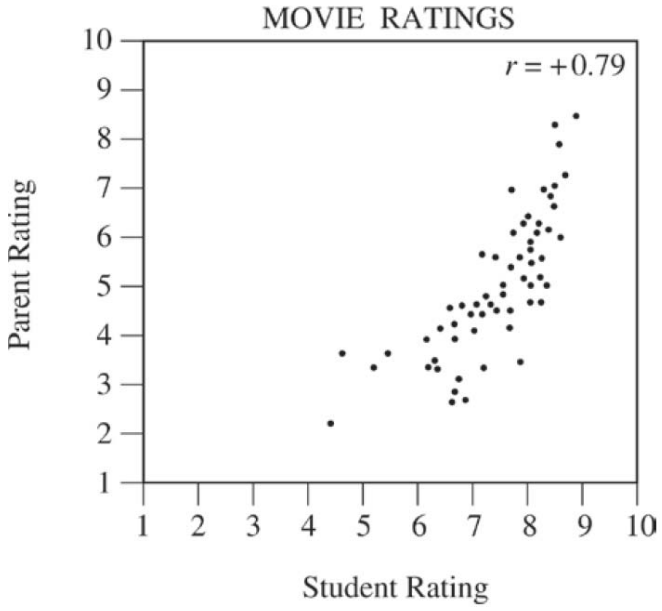
A) is 0.35 B) is 0.50 C) is 0.70 D) is 0.90 E) cannot be determined

22. The residuals plot for a linear model is shown. Which is true?

A) The linear model is okay because approximately the same number of points are above the line as below it.
B) The linear model is okay because the association between the two variables is fairly strong.
C) The linear model is no good because the correlation is near 0.
D) The linear model is no good because some residuals are large.
E) The linear model is no good because of the curve in the residuals.



23. In a recent survey, high school students and their parents were asked to rate 60 recently released movies. The ratings were on a scale from 1 to 9, where 1 was "horrible" and 9 was "excellent". For each movie, the average rating by the students and the average rating by their parents was calculated and the scatterplot below was constructed. The horizontal axis represents the student rating, and the vertical axis represents the parent rating. Thus, an individual data point would represent the rating of a single movie.



Which of the following statements is justified by the scatterplot?

A) The movies that the students liked the best also tended to be the movies that the parents liked the best, but the students tended to give lower scores.

B) The movies that the students liked the best also tended to be the movies that the parents liked the best, but the students tended to give higher scores.

C) The movies that the students liked the best also tended to be the movies that the parents liked the best, but each group tended to give the same scores.

D) The movies that the students liked the best tended to be the movies that the parents liked the least, but the students tended to give lower scores.

E) The movies that the students liked the best tended to be the movies that the parents liked the least, but the students tended to give higher scores.

24. Based on the graph in the previous problem, describe the association between student rating and parent rating for the 60 movies in this sample.

Review ANSEWRS – Unit IV – Scatterplots & Regressions

FREE RESPONSE

1. The association between weight and fuel consumption for these automobiles is...
   - strong (or moderately strong),
   - linear (or fairly linear), and
   - positive – as weight of the automobile increases, so does fuel consumption.

   **make sure to write in complete sentences and use context!!!*
   Grading note: Including the word "correlation" in your response does NOT count for "linear" – you must specify that the relationship between the variables is "linear" (or "fairly linear").

2. Although "x" and "y" have a strong ASSOCIATION, the relationship is curved (nonlinear).
   CORRELATION is only useful for describing a LINEAR association.

3. 9.7

4. a) The association between forearm circumference and strength is moderately weak (due to the low outlier), positive, and mostly linear.

   b) The correlation would become stronger (closer to 1.0)

   c) Currently the right side of the least squares line would be "pulled" down by the point at (14, 38).
   Since removing the point would cause the right side of the line to go UP, the slope of the line would become MORE POSITIVE.

5. a) As the number of stork pairs INCREASES, the populations of these countries also INCREASE.

   b) Having children may not (probably will not?) CAUSE the stork population to increase. Even if there is a strong association between these two variables, **correlation does not imply causation**.

6. From left to right: 0.7994, -0.6112, 0.0023, -0.9713

7. a) Yes: The scatterplot of fare vs distance has a linear pattern, and the plot of residuals vs predicted fare* shows no clear curved pattern.
   (Note: If the horizontal axis of the residual plot is confusing to you… the plot of residuals vs "x" will have the same form as the plot of residuals vs "y-hat". This is because "y-hat" is a direct linear transformation of "x")

   b) $\widehat{airfare} = 177.215 + 0.079(distance)$

   c) For each increase of 1 mile in distance, the model predicts an increase of \$0.079 in airfare.
   (or "For each increase of 1 mile in distance, there is a MEAN increase of \$0.079 in airfare")

   d) A zero-mile long flight is predicted to cost \$177.215. This is perhaps due to overhead costs (also a possible example of extrapolation).

   e) 48% of the variability in airfare is accounted for by the model relating distance and airfare.

   f) s_e = \$41.9. The typical (average-ish) amount that observed airfare differs from the predicted airfare for these flights is about \$41.90. *(don't worry, this is NOT on this test… but we'll see it in the Spring!)*

   g) \$185.115. However 100 miles is outside the range of x-values in this set of data. This is an example of extrapolation, and this prediction might not be very reliable since we don't know if the linear trend continues as we go below 300 miles.

8. a) $\widehat{car\ sold} = 14.2973 + 1.3067(\#\ of\ commercials)$

   b) 37.82 *(expected values should NEVER be rounded to whole numbers!)*

   c) –5.364. This means the dealer sold 5.364 FEWER cars than predicted by the model based on the 10 commercials that the dealer paid for.

   d) OVERestimated

9. a) $r = 0.972$. This indicates a strong, positive, linear association between temperature and ice cream sales.

b) $\widehat{ice\,cream\,sales} = -331.839 + 5.77(temperature)$

c) 0.945; 94.5% of the variability in ice cream sales is accounted for by the model with temperature.

d) 147.486 pounds

e) 222.958 pounds

MULTIPLE CHOICE

10. E
11. B
12. A
13. B
14. B
15. D
16. A
17. A

18. A
19. B
20. B
21. A
22. E

23. B

24. The association between student rating and parent rating for these movies is moderately strong, curved (or non-linear), and positive (generally as student rating increases, so does parent rating).