**AP Statistics – Regression Test**

**PART I: TRUE OR FALSE.**

<p>| | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>A high correlation between (x) and (y) proves that (x) causes (y).</td>
<td>2.</td>
<td>The two coefficients (a and b) for the line of best fit have the same sign.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3.</td>
<td>The correlation coefficient has the same sign as the slope of the least squares line fitted to the same data.</td>
<td>4.</td>
<td>An (r)-value greater than zero indicates that ordered pairs with high (x)-values will have low (y)-values.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5.</td>
<td>A correlation of (-0.41) and (+0.41) show the same degree of clustering around the regression line.</td>
<td>6.</td>
<td>A correlation of (0.75) indicates a relationship that is 3 times as linear as one for which the correlation is only (0.25).</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7.</td>
<td>The mean of the residuals is always zero.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**PART II: MULTIPLE CHOICE.**

**9.** Suppose one has collected data on \(X\) the diameter of tree trunk (in inches) and \(Y\) tree height (in feet). If the Regression equation is \(\hat{y} = -3.6 + 3.1x\), what is your estimate of the average height of all trees having a trunk diameter of 7 inches?

(A) 18.1 ft    (B) 19.1 ft    (C) 20.1 ft    (D) 21.1 ft    (E) 22.1 ft

**10.** In a random sample of older patients at a large medical practice, the age of a patient and a measure of that patient’s hearing loss were recorded. The correlation between age and hearing loss of the patients in the sample was found to be 0.7. Which one of the following would be a correct statement if the age of a patient were used to predict the amount of hearing loss for a patient?

(A) Forty-nine percent of the time, the least squares regression line accurately predicts hearing loss.
(B) Forty-nine percent of the variation in hearing loss can be explained by the variation in the age of a patient.
(C) About 70% of a person’s hearing loss can be explained by age, according to the regression line relating hearing loss and age.
(D) About 70% of the time, age will correctly predict the amount of hearing loss.
(E) The least squares regression line relating hearing loss to age will have a slope of approximately 0.7.

**11.** A correlation between college entrance exam grades and scholastic achievement was found to be \(-1.08\). On the basis of this you would tell the university that:

(A) The entrance exam is a good predictor of academic success.
(B) The exam is a poor predictor of academic success.
(C) Students who do best on this exam will make the worst students.
(D) Students at this school are underachieving.
(E) They should hire a new statistician.

**12.** A study of the effects of television measured how many hours of television each of 125 grade school children watched per week during a school year and their reading scores. Which variable would you put on the horizontal axis of a scatterplot of the data?

(A) Reading score, because it is the response variable.
(B) Reading score, because it is the explanatory variable.
(C) Hours of television, because it is the response variable.
(D) Hours of television, because it is the explanatory variable.
(E) It makes no difference, because there is no explanatory-response distinction in this study.

**13.** The study described in the previous question found that children who watch more television tend to have lower reading scores than children who watch fewer hours of television. The study report says that, “Hours of television watched explained 9% of the observed variation in the reading scores of the 125 subjects.” The correlation between hours of TV and reading score must be

(A) \(r = -0.3\)
(B) \(r = 0.3\)
(C) \(r = -0.09\)
(D) \(r = 0.09\)
(E) Can’t tell from the information given.
14. A study of child development measures the age (in months) at which a child begins to talk and also the child’s score on an ability test given several years later. The study asks whether the age at which a child talks helps predict the later test score. The least-squares regression line of test score \( y \) on age \( x \) is \( \hat{y} = 110 - 1.3x \). According to this regression line, what happens (on average) each month a child delays speech?

(A) The test score goes down 110 points.
(B) The test score goes down 1.3 points.
(C) The test score goes up 110 points.
(D) The test score goes up 1.3 points.
(E) The test score is 108.7.

15. Biologists have collected data relating the age (in years) of one variety of oak tree to its height (in feet). A scatterplot for 20 of these trees is given below.

If the point labeled A in the above scatterplot were changed to have a height of 40 feet, which one of the following statements would be true?

(A) The slope of the least squares regression line would decrease and the correlation would increase.
(B) The slope of the least squares regression line would decrease and the correlation would decrease.
(C) The slope of the least squares regression line would increase and the correlation would increase.
(D) The slope of the least squares regression line would increase and the correlation would decrease.
(E) The slope of the least squares regression line and the correlation would remain the same.

16. If the correlation between body weight and annual income were high and positive, we could conclude that:

(A) High incomes cause people to eat more food.
(B) Low incomes cause people to eat less food.
(C) High-income people tend to spend a greater proportion of their income on food than low-income people, on average.
(D) High-income people tend to be heavier than low income people, on average.
(E) High incomes cause people to gain weight.

17. Residuals are…

(A) possible models not explored by the researcher.
(B) variation in the data that is explained by the model.
(C) the difference between observed responses and values predicted by the model.
(D) data collected from individuals that is not consistent with the rest of the group.
(E) the sum of the “error squares” of all the data points.

18. Which of the following statements about influential scores are true?

I. Influential scores have large residuals.
II. Removal of an influential score sharply affects the regression line.
III. An \( x \)-value that is an outlier in the \( x \)-variable is more indicative that a point is influential than a \( y \)-value that is an outlier in the \( y \)-variable.

(A) I and II
(B) I and III
(C) II and III
(D) I, II, and III
(E) None of the above
19. An efficiency expert wanted to see if there is a relationship between the number of people attending a meeting and the number of minutes late that the meeting started. The table shows the results with the accompanying scatter plot (Figure 1).

<table>
<thead>
<tr>
<th>Number of people attending the meeting</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of minutes late the meeting started</td>
<td>3</td>
<td>6</td>
<td>8</td>
<td>10</td>
<td>14</td>
<td>7</td>
</tr>
</tbody>
</table>

Figure 2 represents the scatter plot with the point (10, 7) removed from the data.

Which one of the following is TRUE about the point (10, 7)?

(A) It has the largest residual.
(B) It is an influential observation.
(C) It is an outlier in the x-direction.
(D) The slope changes little when the point is removed.
(E) The correlation is lower with the removal of the point.

20. Suppose the scatterplot of logX and logY shows a strong positive correlation close to 1. Which of the following must be true?

I. The variables X and Y also have a correlation close to 1.
II. A scatterplot of the variables X and Y shows a strong nonlinear pattern.
III. The residual plot of the variables X and Y shows a random pattern.

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

21. The model \( \sqrt{\text{distance}} = 3.30 + 0.235(\text{speed}) \) can be used to predict the stopping distance (in feet) for a car traveling at a specific speed (in mph). According to this model, about how much distance will a car going 65 mph need to stop?

(A) 4.3 ft  (B) 18.6 ft  (C) 27 ft  (D) 345 ft  (E) 729 ft

**PART III: FREE RESPONSE**

22. There is a relationship between the amount of money in a person’s retirement account and the amount of time that money has been invested. Each of the following strategies has been used to straighten the data. Find the predicted value of \( \text{Money} \) using each model for \( \text{Time} = 30 \) years.

| \( \log(\text{Money}) = 0.92(\text{Time}) - 21 \) | \( \hat{\text{Money}} = 447 + 2,528,951[\log(\text{Time})] \) | \( \sqrt{\text{Money}} = 122 + 65(\text{Time}) \) |
23. According to the article, "First-Year Academic Success: A Prediction for Caucasian and African-American Students" (Journal of College Student Development 1999) there is a mild correlation between high school GPA (HS) and first-year college GPA (Coll). The data, which was collected at a "Southeastern public research university", can be summarized as follows:

\[
\begin{align*}
\text{mean high school GPA} &= 3.7 \\
\text{mean college GPA} &= 2.86 \\
\text{standard dev of high school GPA} &= .47 \\
\text{standard dev of college GPA} &= .85 \\
\text{\( r^2 \) } &= .64
\end{align*}
\]

(a) Find the least square regression line that could be used to predict college GPA. (SHOW ALL WORK)

(b) The correlation indicates a fairly strong relationship between high school and college GPA. Is this relationship causal or associative? Justify your answer.

24. A researcher believes that there is a relationship between the length of your pinky finger (in inches) and your height (in feet). After measuring many of these attributes of people, she came up with the LSRL of \( \hat{\text{height}} = 4.3 + 1.2(\text{pinky}) \).

(a) Adam has a pinky finger that measures 1.35 inches long. His residual from this LSRL is 0.08 feet. How tall is Adam?

(b) Hailey is 5.75 ft tall and has a 1.25 inch pinky finger. What is Hailey’s residual?
25. In a study of the application of a certain type of weed killer, 14 fields containing large numbers of weeds were treated. The weed killer was prepared at seven different strengths by adding 1, 1.5, 2, 2.5, 3, 3.5, or 4 teaspoons to a gallon of water. Two randomly selected fields were treated with all the various strengths of weed killer. After a few days, the percentage of weeds killed on each field was measured. The computer output obtained from fitting a least squares regression line to the data is shown below. A plot of the residuals is provided as well.

(a) What is the equation of the least squares regression line given by this analysis?

(b) Interpret the slope of the regression line in part (a).

(c) Interpret the coefficient of determination \( R^2 \).

(d) Is a linear model appropriate for this data? Explain your reasoning.

(e) Use the regression equation from part (a) to predict the percentage of weeds killed when 2.5 teaspoons of weed killer are used AND decide which of the following you would expect to be a reasonable conclusion. Explain your reasoning.

- The prediction will be too large.
- The prediction will be too small.
- The prediction would have a residual of zero.