**AP Statistics** **Name:** Key
**Mr Murphy** **Date:**
**Summarizing Bivariate Data Sample Test** **Period:**

## Multiple Choice (1 pt. each)

1. Suppose a data set has a linear regression line of $\hat{y} = 6 - 0.8x$. If the mean of the *x*'s is 5, what is the mean of the *y*'s.

Since $(\bar{x}, \bar{y})$ is on the LSRL

$\bar{y} = 6 - 0.8\bar{x} = 6 - 0.8(5) = 6 - 4 = 2$

(a) 2 (b) 5 (c) 10
(d) 6 (e) -5

2. You have the following regression equation for the effect of streetlights per block (*x*), on the crimes per month (*y*): $\hat{y} = 2.4 - 0.2x$. Calculate the residual for a block with 10 streetlights and 1 crime a month (10, 1).

$\hat{y} = 2.4 - 0.2(10) = 2.4 - 2 = 0.4$

(a) -0.6 (b) 0.6 (c) -0.4
(d) 0.4 (e) -1.2

residual $= y - \hat{y} = 1 - 0.4 = 0.6$

3. The residuals for a complete data set are shown below, and $r^2$ for the LSRL that resulted in these residuals is 88.6%.

| **x** | 25 | 30 | 35 | 40 | 45 | 50 | 55 | 60 | 65 | 70 |
|---|---|---|---|---|---|---|---|---|---|---|
| **Residual** | 0 | -1.2 | 2 | 0.97 | 0.9 | -1.1 | -2.2 | -0.6 | 0.25 | 0.98 |

Which of the following is/are true?
I. The linear model is a good model for the data. → resid plot on calculator is scattered
II. The sum of the squares of the residuals is zero. → sum(resid) list ≠ 0
III. The correlation is either $\pm\sqrt{0.886}$.

(a) I only (b) II only (c) III only
(d) I and III only (e) I, II, and III

4. Given a set of ordered pairs (*x*, *y*) with $s_x = 2.5$, $s_y = 1.9$, $r = 0.63$, what is the slope of the regression line of *y* on *x*?

$b = r \cdot \frac{S_y}{S_x} = 0.63 \frac{1.9}{2.5} = 0.4788$

(a) 0.48
(b) 0.65
(c) 1.32
(d) 1.90
(e) 2.63

5. A study of department chairperson ratings and student ratings of the performance of high school statistics teachers reports a correlation of $r = 1.15$ between the between the two ratings. From this information we can conclude that

(a) chairpersons and students tend to agree on who is a good teacher.
(b) chairpersons and students tend to disagree on who is a good teacher.
(c) there is little relationship between chairperson and students ratings of teachers.
(d) there is a strong association between chairperson and student ratings of teachers, but it would be incorrect to incur causation.
(e) a mistake in arithmetic has been made. $-1 \leq r \leq 1$

6. Consider a data set A: (2, 8), (3, 6), (4, 9) and (5, 9). Which of the following is the proper interpretation of the $r^2$ value? $r^2 = 0.3$

(a) There's a 30% increase in the variation of the data set.
(b) Thirty percent of the variation in the *x* values can be explained by knowledge of the *y* values.
(c) Thirty percent of the variation in the *y* values can be explained by knowledge of the *x* values.
(d) You've reduced the total variation by 70%.
(e) None of the above.

7. The regression equation $\hat{y} = 1278.5 - 0.5x$ shows the relationship between the number of calories consumed in a day (*x*) and the marathon times in minutes (*y*) in a sample of world-class distance runners. Interpret the meaning of the slope in the equation stated above.

(a) A one-calorie increase in consumption per day results in a predicted **increase** of 0.5 minutes in marathon time.
(b) A one-calorie increase in consumption per day results in a predicted **decrease** of 0.5 minutes in marathon time.
(c) An increase of 0.5 calories per day results in a predicted one-minute **decrease** in marathon times.
(d) A decrease of 0.5 calories leads to a predicted 1278.5 minute **increase** in marathon times.
(e) None of the above.

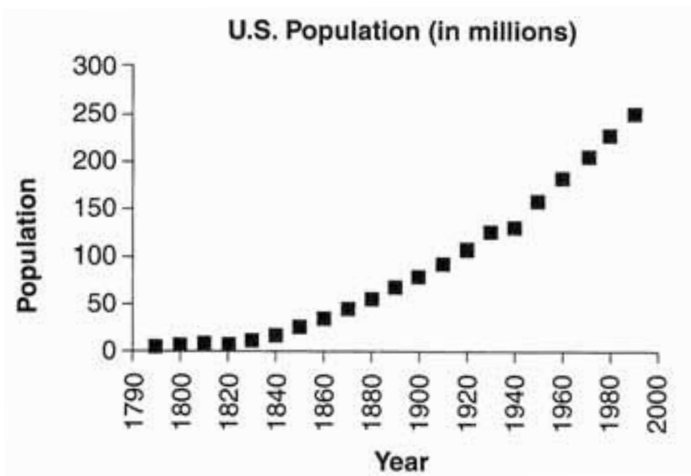
$$\text{slope} = \frac{\Delta y}{\Delta x} = \frac{-0.5}{1} \ \frac{\text{minutes}}{\text{Calories}}$$

8. Which of the following statements about outliers is true?

I. Removing an outlier from a data set can have a major effect on the regression line.
II. If you calculated the residual between the outlier and a regression line, it would probably be large.
III. You will typically find an outlier horizontally distant from the rest of the data along the *x* - axis.

(a) I only (b) II only (c) III only (d) I and II only (e) I and III only

9. The scatterplot below is for the population of the United States every decade from 1780 to 1990.




To transform this relationship you can take the natural logarithms of the response variable to find the regression equation for these data. The resulting regression equation, $r$, and $r^2$ is:

$$\ln y = -35.4173 + 0.02078x,\ r = 0.9835, r^2 = 0.9672$$

This equation would predict the population for 1915 to be:

(a) $10^{4.3764}$ people
(b) $4.3764$ million people
(c) $e^{4.3764}$ million people
(d) $e^{4.3764}$ people
(e) $4.3764 \cdot 10$ million people

$\ln \hat{y} = -35.4173 + 0.02078(1915)$

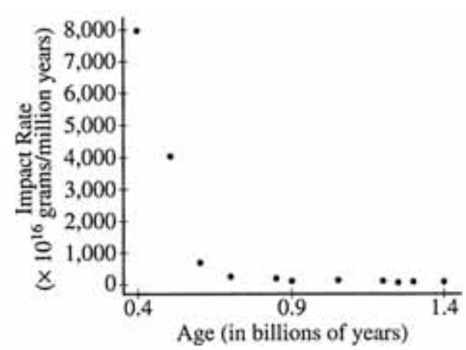
$\ln \hat{y} = 4.3764$

$e^{\ln \hat{y}} = e^{4.3764}$

$\hat{y} = e^{4.3764}$ million people

**Free Response (4pts. each)**

1. (2004B Q1) The Earth's Moon has many impact craters that were created when the inner solar system was subjected to heavy bombardment of small celestial bodies. Scientists studied 11 impact craters on the Moon to determine whether there was any relationship between the age of the craters (based on radioactive dating of lunar rocks) and the impact rate (as deduced from the density of the craters). The data are displayed in the scatterplot below.




(a) Describe the nature of the relationship.

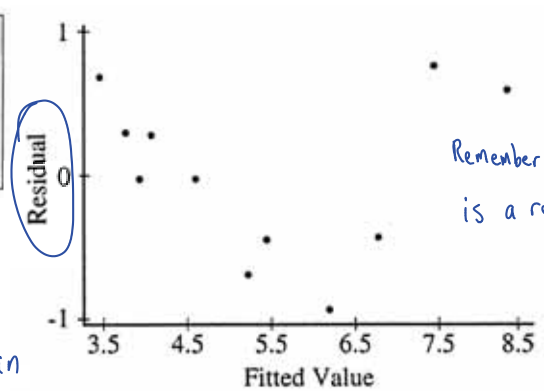
There is a negative non-linear relationship between the age and the impact rate of the craters on the moon

Prior to fitting a linear regression model, the researchers transformed both impact rate and age by using logarithms. The following computer output and residual plot were produced.

Regression Equation: ln(rate) = 4.82 – 3.92 ln(age)

| Predictor | Coef | SE Coef | T | P |
|---|---|---|---|---|
| Constant | 4.8247 | 0.1931 | 24.98 | 0.000 |
| ln(age) | –3.9232 | 0.4514 | –8.69 | 0.000 |

S = 0.5977  R-Sq = 89.4%  R-Sq (adj) = 88.2%




Remember that this is a residual plot

(b) Interpret the value of $r^2$.

89.4% of the variation in ln(rate) can be explained by ln(age)

(c) Comment on the appropriateness of this linear regression for modeling the relationship between the transformed variables.

1) Scatterplot of ln(rate) and ln(age) not given

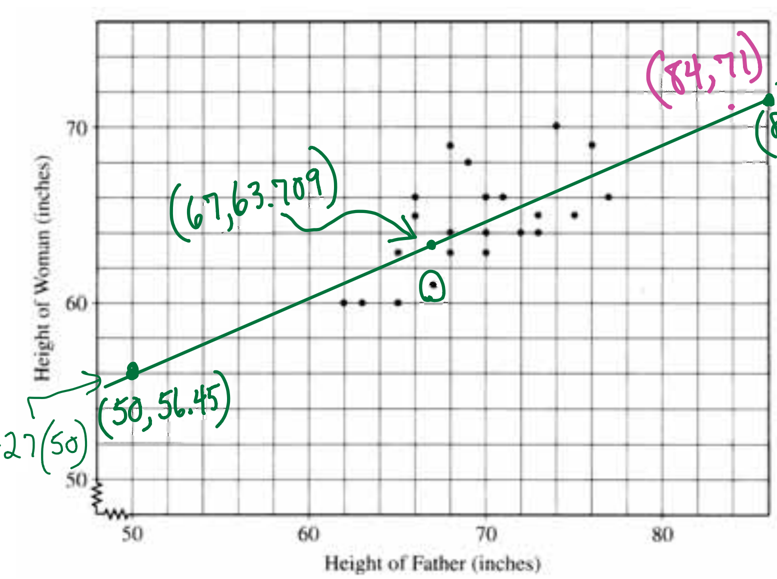
2) $r = -0.946$, strong negative linear relationship between ln(rate) and ln(age)

3) Residual plot above has a curved pattern

4) $S_e = 0.5977$ nothing to compare it to

Curved Residual plot indicates a linear model for the relationship between ln(age) and ln(rate) is not appropriate

2. (2007B Q4) Each of 25 adult women was asked to provide her own height (*y*), in inches, and the height (*x*), in inches, of her father. The scatterplot below displays the results. Only 22 of the 25 pairs are distinguishable because some of the (*x*, *y*) pairs were the same. The equation of the LSRL is $\hat{y} = 35.1 + 0.427x$.




$\hat{y} = 35.1 + 0.427(86)$

$\hat{y} = 35.1 + 0.427(50)$

(a) Draw the LSRL on the scatterplot above.

(b) One father's height was *x* = 67 inches and his daughter's height was *y* = 61 inches. Circle the point on the scatterplot above that represents this pair and draw the segment on the scatterplot that corresponds to the residual for it. Give the numerical value for the residual.

67 inches $\rightarrow$ $\hat{y} = 35.1 + 0.427(67) = 63.709$

residual $= y - \hat{y} = 61 - 63.709 = -2.709$ inches

(c) Suppose the point *x* = 84, *y* = 71 is added to the data set. Would the slope of the LSRL increase, decrease, or remain about the same? Explain.

Would the correlation increase, decrease, or remain about the same? Explain.

(Note: No calculations are necessary to answer this question.)

The slope would remain about the same (or slightly increase because the point is above the line) because the point lands close to the line

The correlation would increase for the same reason plus because it is further away from the rest of the x values making it a likely influential point.