

## Homework 3

Name \_\_\_\_\_

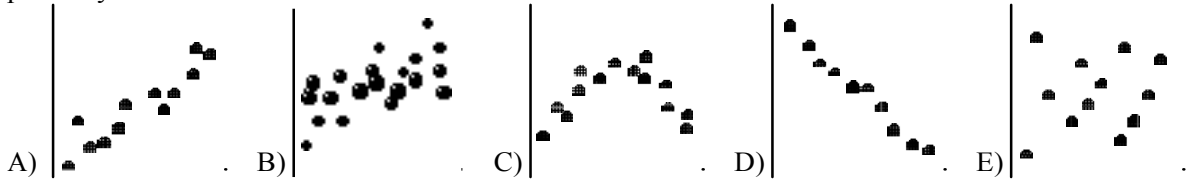
\_\_\_ 1. Scientists have collected data on the heights of mothers and daughters. The correlation between the mothers' heights and the heights of their 16 year-old daughters is most likely to be . . .

- A) near  $-1.0$     B) near 0    C) near  $+0.7$     D) exactly  $+1.0$     E) somewhat greater than 1.0

\_\_\_ 2. The car industry crashed some cars into a cement wall at speeds of 5 to 25 mph to learn about the amount of damage to the cars. They found a correlation of  $r = 0.60$  between speed (MPH) and damage (\$). If the speed at which a car hit the barrier is 1.5 standard deviations above the mean speed, we expect the damage to be \_\_\_\_\_ the mean damage. (hint: consider the slope formula as a function of the standard deviations:  $b = r * s_y/s_x$ ). Show your work.

- A) equal to    B) 0.36 SD above    C) 0.60 SD above    D) 0.90 SD above    E) 1.5 SD above

\_\_\_ 3. Which of the following shows a strong *association* between two variables even though the *correlation* is probably near zero?



\_\_\_ 4. The correlation between  $X$  and  $Y$  is  $r = 0.35$ . If we double each  $X$  value, decrease each  $Y$  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.

\_\_\_ 5. The correlation between a family's weekly income and the amount they spend on restaurant meals is found to be  $r = 0.30$ . Which must be true?

- I. Families tend to spend about 30% of their incomes in restaurants.  
 II. In general, the higher the income, the more the family spends in restaurants.  
 III. The line of best fit passes through 30% of the (*income*, *restaurant*\$) datapoints.
- A) I only    B) II only    C) III only    D) II and III only    E) I, II, and III

\_\_\_ 6. A doctor finds that the more overweight a person is, the higher his pulse rate tends to be. In fact, the model suggests that 12-pound differences in weight are associated with differences in pulse rate of 4 beats per minute. Which is true?

- I. The correlation between pulse rate and weight is 0.33
- II. If you lose 6 pounds, your pulse rate will slow down 2 beats per minute.
- III. A positive residual means a person's pulse rate is higher than the model predicts

A) none      B) I only      C) II only      D) III only      E) II and III only

\_\_\_ 7. Education research consistently shows that students from wealthier families tend to have higher SAT scores. The slope of the line that predicts *SAT score from family income* is 6.25 points per \$1000, and the correlation between the variables is 0.48. Then the slope of the line that predicts *family income from SAT score* (in \$1000 per point) ... *Show your work*

A) is 0.037      B) is 0.16      C) is 3.00      D) is 6.25      E) is 13.02.

\_\_\_ 8. A regression analysis of company profits and the amount of money the company spent on advertising found  $r^2 = 0.72$ . Which of these is true?

- I. This model can correctly predict the profit for 72% of companies.
- II. On average, about 72% of a company's profit results from advertising.
- III. On average, companies spend about 72% of their profits on advertising.

A) none      B) I only      C) II only      D) III only      E) I and

\_\_\_ 9. All but one of these statements contain a mistake. Which could be true?

- A) The correlation between a football player's weight and the position he plays is 0.54.
- B) The correlation between the amount of fertilizer used and the yield of beans is 0.42.
- C) The correlation between a car's length and its fuel efficiency is 0.71 miles per gallon.
- D) There is a high correlation (1.09) between height of a corn stalk and its age in weeks.
- E) There is a correlation of 0.63 between gender and political party.

\_\_\_ 10. 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) none of these.

\_\_\_ 11. Which statement about influential points is true?

- I. Removal of an influential point changes the regression line.
- II. Data points that are outliers in the horizontal direction are more likely to be influential than points that are outliers in the vertical direction.
- III. Influential points have large residuals.

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

\_\_\_ 12. Which is true?

- I. Random scatter in the residuals indicates a model with high predictive power.
- II. If two variables are very strongly associated, then the correlation between them will be near +1.0 or -1.0.
- III. The higher the correlation between two variables the more likely the association is based in cause and effect.

A) none      B) I only      C) II only      D) I and II only      E) I, II, and III

\_\_\_ 13. A company's sales increase by the *same* amount each year. This growth is . . .

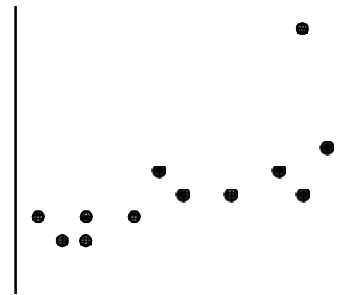
A) linear      B) exponential      C) logarithmic      D) power      E) quadratic

\_\_\_ 14. Two variables that are actually not related to each other may nonetheless have a very high correlation because they both result from some other, possibly hidden, factor. This is an example of

A) leverage.      B) a lurking variable      C) extrapolation.      D) regression.      E) an outlier.

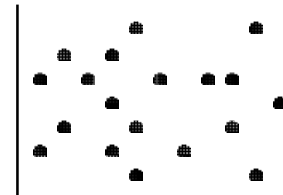
\_\_\_ 15. If the point in the upper right corner of this scatterplot is removed from the data set, then what will happen to the slope of the line of best fit ( $b$ ) and to the correlation ( $r$ ) ?

- A) both will increase.
- B) both will decrease.
- C)  $b$  will increase, and  $r$  will decrease.
- D)  $b$  will decrease, and  $r$  will increase.
- E) both will remain the same.

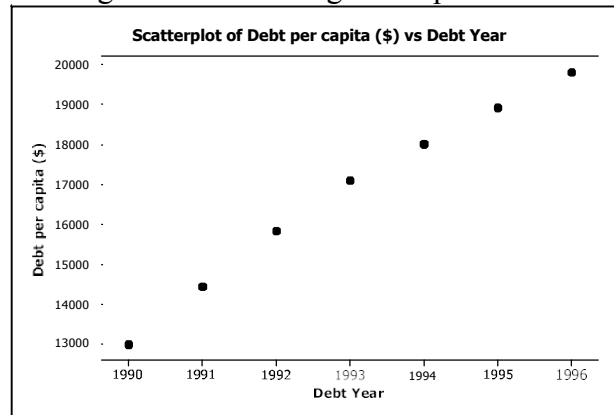


\_\_\_ 16. A least squares line of regression has been fitted to a scatterplot; the model's residuals plot is shown. Which is true?

- A) The linear model is appropriate.
- B) The linear model is poor because some residuals are large.
- C) The linear model is poor because the correlation is near 0.
- D) A curved model would be better.
- E) None of the above



17. **Personal debt** According to *The World Almanac and Book of Facts 2004*, the debt per capita for the years 1990-2001 gives the following scatterplot:

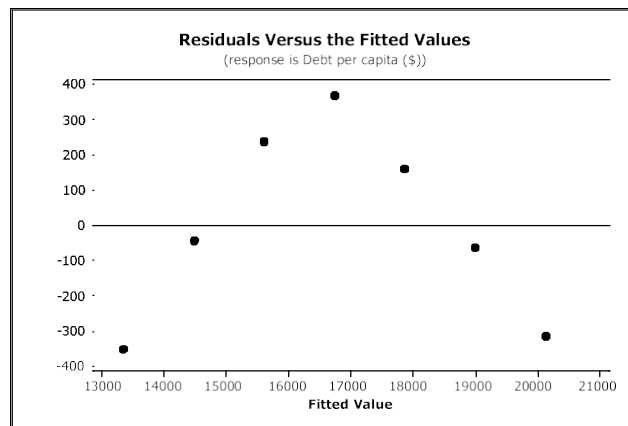


Regression output gives the equation of the regression line as

$$\text{Debt} = -2,231,226 + 1128(\text{Year}) \text{ with } R^2 = 98.8\%.$$

- a. What is the response variable?
  1. -2,231
  2. +1128
  3. 98.8
  4. Debt per capita
- b. Is the correlation coefficient positive or negative?
  1. Positive
  2. Negative
- c. What is the correlation coefficient  $r$ ?
  1. 0.994
  2. -0.998
  3. -0.994
  4. 0.998
- d. Explain in context what the slope of the line means:
  1. On average, debt per capita increases \$1128 per year
  2. On average, debt per capita decreases \$1128 per year
  3. On average, debt per capita increases every year by 98.8%
  4. On average, debt per capita increases every year by the 98.8% of \$1128
- e. Explain in context what  $R^2 = 98.8\%$  means
  1. The debt is almost at the maximum level
  2. The debt is 98.8% of the predictor variable
  3. About 98.8% of the variability of debt per capita is explained by the model
  4. The model is very good and has only 0.2% variability

- f. You decide to take a look at a residuals plot before making any predictions. Based on the following residuals plot, does linear regression seem appropriate for these data?



1. Yes, because the distribution is symmetric and unimodal
2. No, because there a bend in the curve
3. Yes, because there are not outliers
4. No, because we don't have enough datapoints to come to a conclusion