MORTON

# Chapter 4: More on Two-Variable Data Review

1. Which of the following is true?

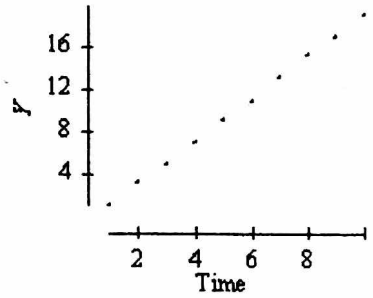
A) log(AB) = logAlogB.
B) log(A+B) = logA + logB.
C) log(A/B) = logA – logB.
D) All of the above.

2. Which of the following would provide evidence that a power law model describes the relationship between a response $y$ and an explanatory variable $x$?

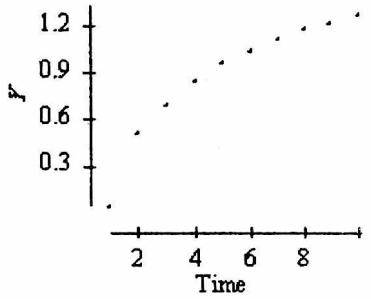
A) A scatterplot of $y$ versus $x$ looks approximately linear.
B) A scatterplot of log $y$ versus $x$ looks approximately linear.
C) A scatterplot of $y$ versus log $x$ looks approximately linear.
D) A scatterplot of log $y$ versus log $x$ looks approximately linear.

3. Which of the following scatterplots would indicate that $Y$ is growing exponentially over time?

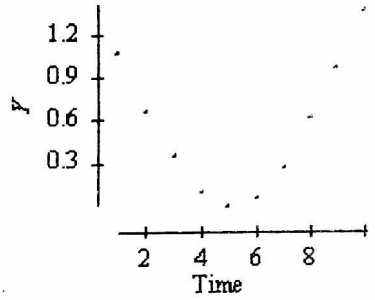
A)



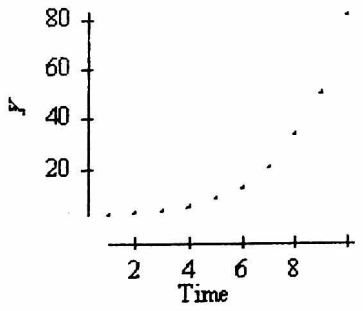

B)



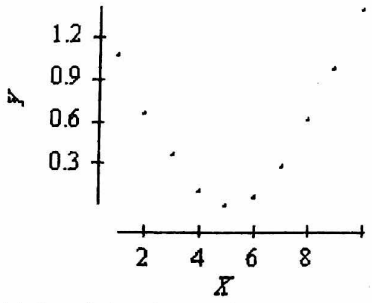

C)




D)



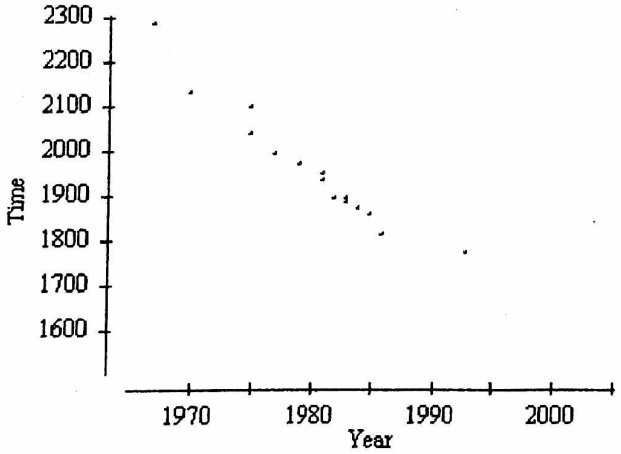
4. A scatterplot of a response variable $Y$ versus an explanatory variable $X$ is given below.




Which of the following is true?

(A) There is a nonlinear relationship between $Y$ and $X$.
B) There is a very strong positive correlation between $Y$ and $X$ because there is an obvious relation between these variables.
C) There is a monotonic relation between $Y$ and $X$.
D) All of the above.

5. A scatterplot of the world record time for women in the 10,000-meter run versus the year in which the record was set appears below. Note that the time is in seconds and the data are for the period 1965–1995.




Based on this plot, we can expect

A) that by 2005 the world record time for women will be well below 1500 seconds.
B) that about every decade, we can expect the world record time to decrease by at least 100 seconds.
C) that about every decade, we can expect the world record time to decrease by about 50 seconds.
(D) none of the above. EXTRAPOLATION!

6. Two variables, $x$ and $y$, are measured on each of several individuals. The correlation between these variables is found to be 0.88. To help us interpret this correlation we should do which of the following?

A) Compute the least-squares regression line of $y$ on $x$ and consider whether the slope is positive or negative.
B) Interchange the roles of $x$ and $y$ (i.e., treat $x$ as the response and $y$ as the predictor variable) and recompute the correlation.
C) Plot the data.
D) All of the above.

7. Which of the following would be necessary to establish a cause-and-effect relation between two variables?

A) strong association between the variables.
B) an association between the variables observed in many different settings.
C) plausibility of the alleged cause.
D) all of the above.

* Do an experiment *

8. A researcher computed the average Math SAT score of all high school seniors who took the SAT exam for each of the 50 states. The researcher also computed the average salary of high school teachers in each of these states and plotted these average salaries against the average Math SAT scores for each state. The plot showed a distinct negative association between average Math SAT scores and teacher salaries. The researcher may legitimately conclude which of the following?

A) Increasing the average salary of teachers will cause the average of Math SAT scores to decrease, but it is not correct to conclude that increasing the salaries of individual teachers causes the Math SAT scores of individual students to increase.
B) States that pay teachers high salaries tend to do a poor job of teaching mathematics, on average.
C) States whose students tend to perform poorly in mathematics probably have a higher proportion of problem students and thus need to pay teachers higher salaries in order to attract them to teach in those states.
D) The data used by the researcher do not provide evidence that increasing the salary of teachers will cause the performance of students on the Math SAT to get worse.

Best option! ↘

9. A study of the salaries of full professors at Upper Wabash Tech shows that the median salary for female professors is considerably less than the median male salary. However, further investigation shows that the median salaries for male and female full professors are about the same in every department (English, physics, etc.) of the university. This apparent contradiction is an example of

A) extrapolation.
B) Simpson's paradox.
C) causation.
D) correlation.

10. $X$ and $Y$ are two categorical variables. The best way to determine whether there is a relation between them is to

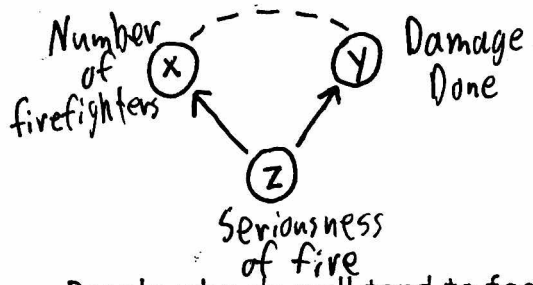
A) calculate the correlation between $X$ and $Y$.
B) draw a scatterplot of the $X$ and $Y$ values.
C) make a two-way table of the $X$ and $Y$ values.
D) do all of the above.

AP Statistics
Chapter 4.2 - Cautions about Correlation

Name Mr. Morton

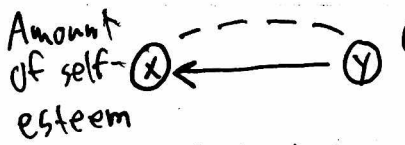
For each question, state whether the relationship between the two variables involves causation common response, or confounding. Identify possible lurking variables. Draw a diagram of the relationship in which each circle represents a variable. Write a brief description of the variable by each circle.

1. There is a strong positive correlation between the number of firefighters at a fire and the amount of damage the fire does. Therefore, sending lots of firefighters causes more damage.




Common Response:
Lurking Variable is Seriousness of the fire.

2. People who do well tend to feel good about themselves. Perhaps helping people feel good about themselves will help them do better in school and life. Raising self-esteem became for a time a goal in many schools. California created a state commission to advance the cause of self-esteem. Consider the statement "Self-esteem causes better work in school".




Causation:
Relationship goes both ways.

3. The correlation between the average SAT math score and the average SAT verbal score for each of the 50 states and District of Columbia is $r = 0.962$.

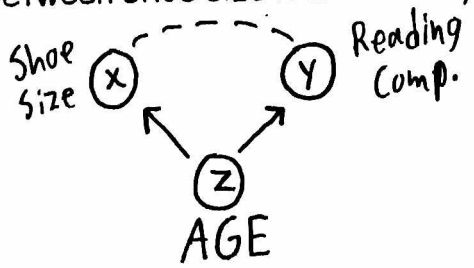
a. Find $r^2$ and explain in simple language what this number tells us.

$r^2 = 0.9254$ | 92.54% of the variation of SAT verbal scores can be explained by average SAT math scores.

b. If you calculated the correlation between the SAT math and verbal scores of a large number of individual students, would you expect the correlation to be about 0.96 or quite different? Explain your answer.

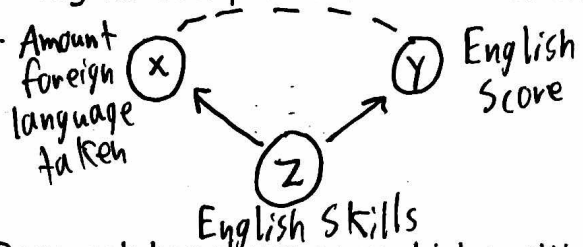
Individual Scores would cause r to drop.

4. A study of elementary school children, ages 6 to 11, finds a high positive correlation between shoe size $x$ and score $y$ on a test of reading comprehension.



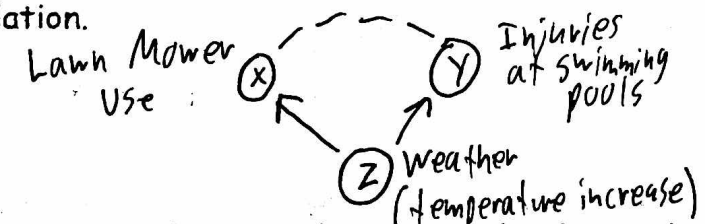

Common Response:
Lurking Variable is AGE.

5. Members of a high school language club believe that study of a foreign language improves student's command of English. From school records, they obtain the scores on an English achievement test given to all seniors. The mean score of seniors who studied a foreign language for at least two years is indeed much higher than the mean score of seniors who studied no foreign language. Identify the explanatory and response variables in this study. Then explain what lurking variable prevents the conclusion that language study improves students' English scores.



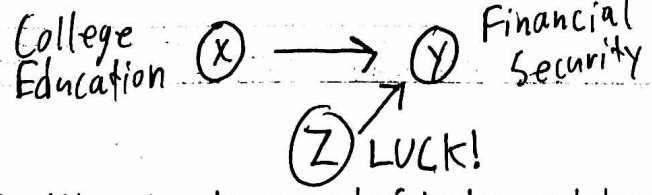

Common Response:
Lurking Variable is English Skills.

6. Research has shown a very high positive correlation between the number of accidents due to lawn mower use and the number of injuries at neighborhood outdoor swimming pools. Explain the correlation.




Common Response:
Lurking Variable is weather.

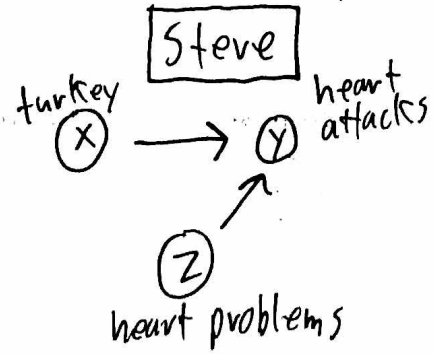
7. Sam's parents have tried to stress that he needs to get a college education in order to achieve financial security in later life. They point out that his Uncle Ed, who dropped out of high school, is barely scraping by, washing dishes at Angus Barn; Aunt Edith, with a Master's in Business Administration, is the business manager for Angus Barn and 4 other affiliated restaurants. Sam disputes their assertion, reminding them that his favorite musician "Big Willie" makes mega-bucks, as does basketball great LeBron James, while his cousin Brandon is a graduate of the NC Justice Academy and only makes $40,000 a year as a police officer.

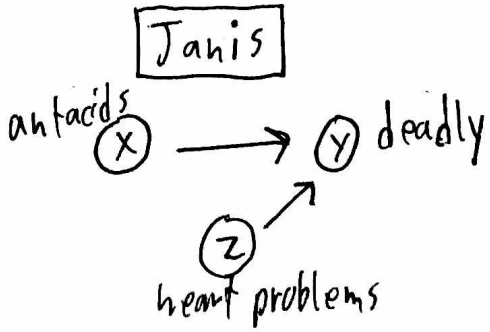



Confounding:
Lurking Variable is LUCK.

8. Uncle PeeWee ate a huge meal of turkey and dressing for Thanksgiving dinner. Afterwards, he took a long nap on the couch. Feeling discomfort in his chest upon waking, he took repeated doses of antacids to quell the intensifying heartburn. Later that evening he died of a massive heart attack. Ever since this episode his son Steve will not eat turkey, because it causes heart attacks. His daughter Janis will still eat turkey in moderate portions, but refuses to take antacids because they are deadly!

Both Confounding! LV: Heart Problems

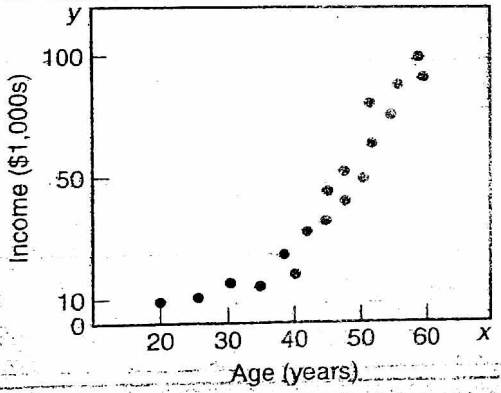


## *MULTIPLE-CHOICE QUESTIONS*

1. A simple random sample of years and earnings was organized into pairs (time in years, earnings in \$1,000's). The scatterplot appears exponential and the transformation $(x_i, y_i) \rightarrow (x_i, \ln y_i)$ is applied to the data. A TI calculator yields the linear regression equation $y = a + bx$ where $a = .3079$, $b = .464$, and $r^2 = .922$.
Which of the following is a valid conclusion?
(A) The earnings gained after 12 years are approximately 5.8759.
(B) The earnings gained after 12 years are approximately 356.345.
(C) The earnings will increase by .464 thousand dollars each year.
(D) The original investment was \$307.90.
(E) None of these is valid. $\hat{y} = 1.36(1.59)^x$

2. Which of the following is *not* a reasonable choice of techniques to attempt to achieve linearity for this scatterplot of data?




(A) $(x_i, y_i) \rightarrow (x_i, \ln y_i)$; linear regression
(B) $(x_i, y_i) \rightarrow (x_i, \sqrt{y_i})$; linear regression
(C) $(x_i, y_i) \rightarrow (\ln x_i, \ln y_i)$; linear regression
(D) Perform two linear regressions: (1) domain [20, 40] and (2) domain [40, 60].
(E) All of these are reasonable techniques.

## *FREE-RESPONSE QUESTIONS*

### *Open-Ended Questions*

Complete a regression analysis for the following age and income data as indicated.

| Age (years) | 20 | 25 | 30 | 35 | 40 | 45 | 50 | 55 | 60 |
|---|---|---|---|---|---|---|---|---|---|
| Income ($1,000) | 18.5 | 23.6 | 29.8 | 38.5 | 49.0 | 64.1 | 78.5 | 102.0 | 130.8 |

1. Construct and label a scatterplot of the data.

2. Perform a linear regression on the data; plot the regression line on the scatterplot.
$\hat{y} = -48.2311 + 2.6913x$

3. Discuss the goodness of fit of the linear regression referencing the correlation coefficient and its residual plot.
Residual curved, $r = 0.9647$

4. Perform the following transformations: exponential and power. $(x, y) \rightarrow (x, \log y)$

5. Perform the linear regression on both sets of transformed data.
$\log \hat{y} = 0.8418 + 0.0212x$

6. Discuss the goodness of fit of these linear regressions referencing the correlation coefficients and each of their residual plots.
Great Fit
$r = 0.9999$, $r^2 = 0.9997$

7. Transform the linear models into the exponential and power models and plot each on the original scatterplot. $\hat{y} = 6.947(1.05)^x$

8. Comment on which of the three regression models fits the data the best. Explain your answer.
When you linearize the exponential function, you can see the r-value at its greatest and the residual plot clear of any visible pattern. Therefore, the exponential function is the correct pick.