AP Stats Name____

Cautions about Correlation and Regression

Correlation and regression describe only linear relationships. Remember that the correlation, "r," and LSRL's are not resistant.

Not Resistant:

Extrapolation:

Example:

A rat's mass is measured at time = 0 for a study. At this point, the rat's mass is 40 grams and it gains 2 grams per week throughout the study. The variables measured are time versus weight for a total of 6 weeks. The regression line would represent the weight gain in the rat for those six weeks, but it would not make sense to extend the regression line for time past those 6 weeks. At some point, it would predict values that are incredibly unrealistic.

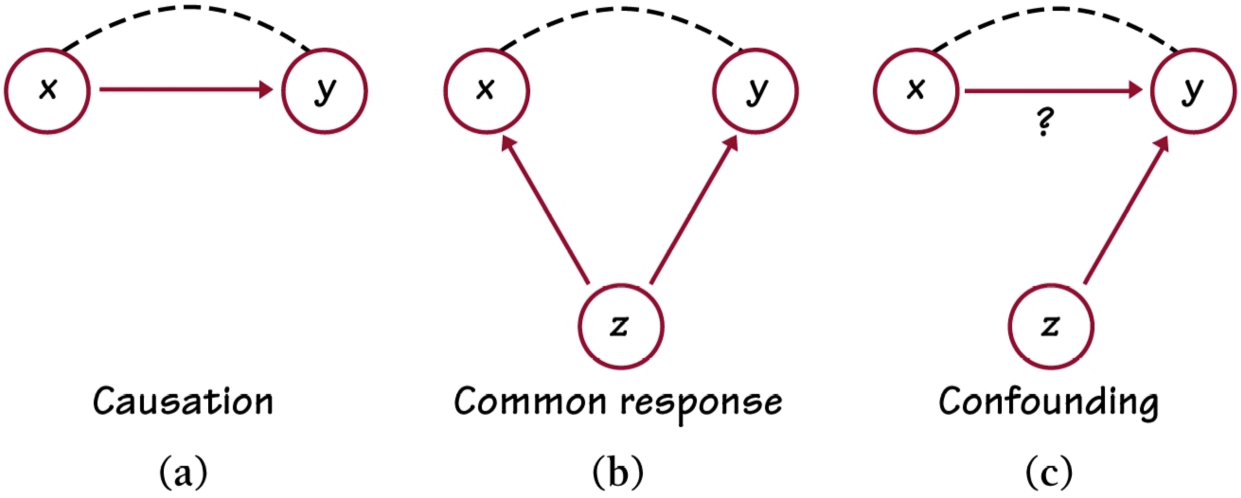
Example: There is data that compares the age of 15 people to their heights. If you created a regression line for the data (perhaps captured between the ages of 4 and 10) and plugged in in 25 years, you would predict the person to be 8 feet tall. This regression model only works for the younger part of the person's life.

Using Averaged Data:

- **Correlations based on averages are usually too high when applied to individuals**
- **Example:** Data is comparing the outside temperature and a household's natural gas consumption per month. The averaged data was the natural gas per month. If the data would have been graphed for the individual days then it would show more scatter about the regression line and lower the correlation. Averaging per month smooths out the other possible variables like day-to-day variations due to doors left open, houseguests, etc.

The goal for studying relationships between two variables is that the changes in the explanatory variable should CAUSE changes in the response variable.

Here are three ways that relationships between variables can be explained.



Examples:

1. **Causation**:

Example 1: x = mother's BMI
y = daughter's BMI

- Body type is partly determined by heredity
- **Even when direct causation is present, it is rarely a complete explanation of an association between two variables**

Example 2: x = amount of artificial sweetener in a rat's diet
y = count of tumors in the rat's bladder

- Experiments show this is true
- **Even well-established causal relations may not have the same general causes in other settings**

2. **Common Response**:

Example 3: x = a high school seniors SAT score
y = the students first-year college grade point average

- The student's ability and knowledge are outside variables that would affect both SAT scores and GPA

Example 4: x = # of firefighters at a fire
y = the amount of damage done

- More firefighters are present at fires where more damage occurs. Presence of firefighters does not increase damage. Both the number of firefighters and the amount of damage are showing a common response to the severity of the fire.

3. **Confounded**:

Example 5: x = activity level
y = weight gain

- These two have a positive strong correlation when graphed but there is another variable that could cause weight gain, which is z = age (slowing down of metabolism as you get older). We can't necessarily distinguish between which variable (x or z) causes weight gain (y).

Example 6: x = the number of years of education a worker has
y = the worker's income

- The confounded variable would be "given" wealth. If you have a prosperous home then you are more likely to go to school and start out with higher earnings without much education. So number of years in school does not determine someone's income necessarily.

The only way that you can determine causation is to have a carefully designed experiment in which the effects of possible outside variables are controlled.

<u>**CRITERIA FOR ESTABLISHING CAUSATION WHEN WE CANNOT DO AN EXPERIMENT**</u>

1.

2.

3.

4.

5.

The problem is that **correlation** is different from **causation**. **Correlation** is when two or more things or events tend to occur at about the same time and might be associated with each other, but aren't necessarily connected by a cause/effect relationship. For instance, in sick people, a runny nose and a sore throat correlate to each other--they tend to show up in the same patients. However, that doesn't mean runny noses cause sore throats, or that sore throats cause runny noses. Forgetting that leads to sloppy thinking.

For the following problems, decided which type of relationship the situations have.

1. When I'm stressed, I get muscle cramps. However, when I'm stressed, I also drink lots of coffee and lose sleep. So it's hard to tell whether my cramps are actually caused by coffee, lack of sleep, stress, or some combination of the above. "Lots of coffee" and "lack of sleep" are examples of:
   a) direct causation
   b) common response
   c) confounding

2. A pro baseball player has an exceptionally poor batting performance during a night game. The very next day the batting coach spends several hours working with the player. The next game the same baseball player has a much better batting performance. The change in performance most likely can be explained by:
   a) direct causation
   b) common response
   c) confounding

3. Chris runs home from school every day. Chris finds that when he runs faster, he gets home sooner. The change in travel time most likely can be explained by:
   a) direct causation
   b) common response
   c) confounding

4. Surf board sales rise when lemonade sales rise. A conclusion about the causality in the preceding example is flawed because it is most likely a result of:
   a) direct causation
   b) common response
   c) confounding