♦ Sects 2.4, 2.5: Cautions About Correlation and Regression; Causation

♦ Correlation and regression describe _______ relationships!!

♦ Correlation and least-squares regression are not ________ to outliers.

Regression Diagnostics

Question: Is regression the best model for our data?

1) Look at a ___________________________________________ of the data. Ask
   Does the form of the relationship appear to be ____________________________?

2) Consider the correlation r, and the $r^2$ value
   If the correlation does not indicate a strong relationship, or if the $r^2$ value is low, linear
   regression may not be appropriate. There is no set rule for how high an $r^2$ value you need,
   but values above ______ (_______ % or more of the variation in y is explained by the
   regression) indicate that linear regression explains the relationship well.

3) Look at a plot of the residuals

   A Residual plot is a scatterplot of the regression residuals against the explanatory variable.

   A residual plot will:
   • be centered at zero on the y axis
   • show no linear pattern

   To assess our model, we look for patterns in the residual plots. Our regression line is a good
   model if

   So, we hope to see:

   Patterns to look out for:

   Curves                        Spread changing with x

   If these are seen, linear regression is not appropriate for modeling the relationship. Consult your
   local statistician to develop a better model!

Things to watch out for:
1) Using Averaged Data

Example: S&P 500

Scenario 1: The S&P 500 stock index is an average price of 500 stocks. Suppose we
   plot this average value for each day this month.
Scenario 2: In a single graph______________________________

Q: Which scenario will produce a scatterplot with a higher correlation? Why?

2) **Clusters**

When a scatterplot shows distinct clusters, ________________________________

3) **Lurking Variables**

Relationships between variables often is greatly influenced by other variables.

A **lurking variable** is a variable that has______________________________

Example

Children who watch many hours of television get lower grades in school on the average than those who do not watch much TV. Suggest some lurking variables that may explain this relationship because they are related to both heavy TV viewing and poor grades.

4) **Influential Observations**

An observation is **influential** for a statistical calculation if

Typically, points that are outliers ________________________________ are influential for the least-squares regression line.

Example: Sales - Profits Data

5) **Association is not Causation!!**

When analyzing relationships between variables, one has to be careful about claiming that changes in the explanatory variable *cause* changes in the response variable.

There may be a **lurking variable** that can account for the association.

Example

A study of elementary school children, ages 6 to 11, finds a high positive correlation between shoe size and scores on a test of reading comprehension. Does it make sense to claim a bigger shoe size *causes* better reading skills? What explains this correlation?
Types of Association

x = explanatory variable  z = lurking variable  y = response variable

Match the following scenarios with its corresponding type of association. (see also Intro. to Chapter 3)

a) Confounding  b) Causation  c) Common Response

Scenario 1: __________________________

The homerun ball hit by ____________________ hits you in the head. Soon after you develop a
________________________.

Scenario 2: __________________________

It was a sunny day at the ballpark. You didn’t wear sunscreen resulting in an unpleasant sunburn. During the last play of the game ____________________ hit a homerun ball that lands on your head. As you are driving home, you develop a ____________________

Scenario 3: __________________________

The number of flu cases was recorded each week through the year as was the amount of ice cream sold. It was noted that as ice cream sales ____________________ the number of flu cases ____________________. Can we conclude that consuming ice cream ____________________ the flu?
Causation
1. best established via carefully designed randomized, comparative experiments (see Ch. 3)

2. in absence of experimentation look for
   a. strong association
   b. consistent association
   c. higher doses associated with stronger responses
   d. alleged cause precedes effect in time
   e. alleged cause is plausible