Lecture 14
Chapter 11
Relationships Can Be Deceiving

- Illegitimate Correlations
- Reasons for a Relationship
- Establishing Causation in Observational Study

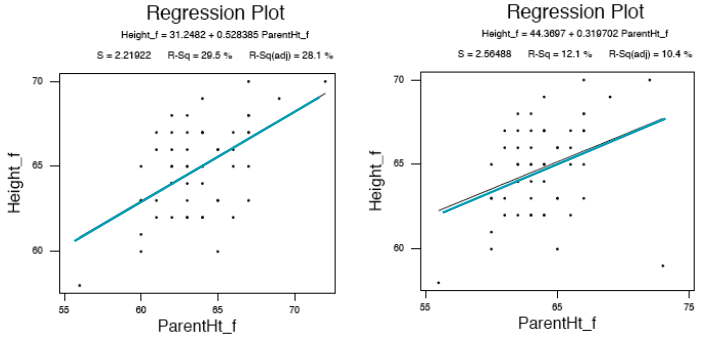
Definitions

- **Outlier** (in regression): point that is unusually far (vertically) from the regression line.
- **Influential observation:** point with high degree of influence on regression line.
  - A point that is inconsistent with the trend of the data can decrease the correlation.
  - A point that is consistent with the trend of the data can inflate the correlation.

An illegitimate correlation is one that fails to reflect the true strength of the relationship.

Example: *An Influential Height Observation*

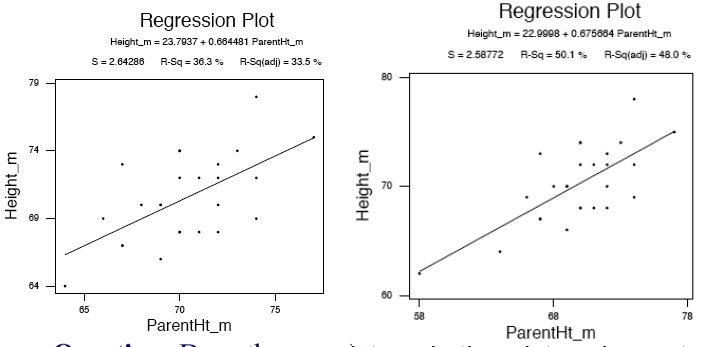
- **Background**:Add data: very short daughter, very tall mother.



- **Question:** Does the correlation do the relationship justice?
- **Response:**

Example: *Another Influential Observation*

- **Background**:Add data: very short son, very short father.



- **Question:** Does the correlation do the relationship justice?
- **Response:**

More about Correlation *r*

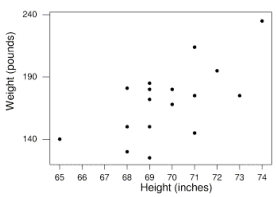
- Correlation is a standardized measure of the direction and strength of the linear relation between 2 quantitative variables
  - A strong curved relationship may have *r* close to 0
  - *r* is unaffected by change of units
  - *r* based on averages overstates strength
  - *r* may be high due to confounding variables

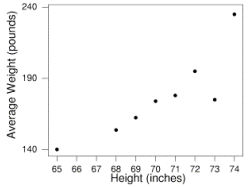
Example: *Correlation Based on Averages*

Ht	65	68			69					70		71			72	73	74
Wt	140	130	150	181	125	150	172	180	185	168	180	145	175	214	195	175	235
AvWt	140	153.7			162.4					174.0		178.0			195	175	235

- **Background**: For male students plot…

Left: wt. vs. ht. or **Right:** *average* wt. vs. ht.





- **Question:** Which one has *r* =+0.87? (other *r* =+0.65)
- **Response:** Plot on _____ has r=+0.87 (stronger)

Example: *Another r Based on Averages*

- **Background**: Richard Doll regressed lung cancer death rates on per capita cigarettes for 11 countries, and found *r* =+0.7 (fairly strong relationship).
- **Question:** Can we predict very well whether or not an individual will die of lung cancer, if we know whether or not he/she smokes?
- **Response:**

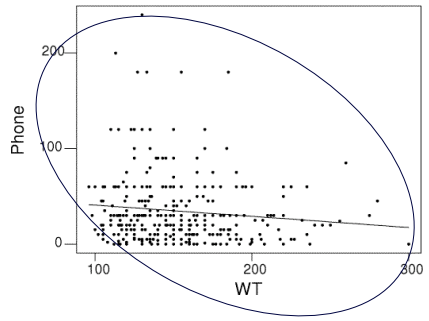
In general, correlation based on averages tends to overstate strength because scatter due to individuals has been reduced.

Confounding Variables in Regression

- Combining two groups that differ with respect to a variable that is related to both explanatory and response variables can affect the nature of their relationship.

Example: *Additional Variables*

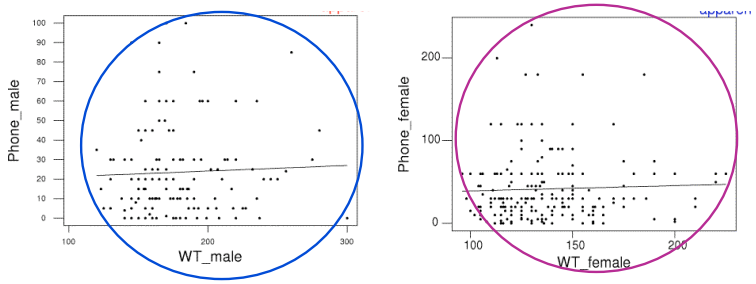
- **Background**: A regression of phone time (in minutes the day before) and weight shows a negative relationship.



- **Questions:** Do heavy people talk on the phone less? Do light people talk more?

Example: *Confounding Variables*

- **Background**: A regression of phone time (in minutes the day before) and weight shows a negative relationship.



- **Response:** _____ is confounding variable→regress separately for males and females→ no relationship

Reasons for a Relationship

1. Explanatory variable is direct cause of responses.
2. Response var. causes changes in explanatory var.
3. Explanatory var. is contributing cause of responses.
4. Confounding variable(s) relate the two variables.
5. Both variables result from a common cause.
6. Both variables are changing over time.
7. The association is just a coincidence.

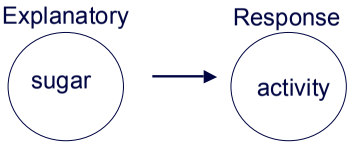
Note: We'll group reasons 1&3 together; 4&5 together.

Example:*Reasons for Sugar/Activity Relationship*

Background: Say an observational study shows that for kids, sugar intake x and activity level y have positive r.

Question: Tell how each reason could be the case.

Response: 1&3

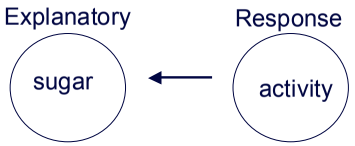


Example: *Reasons for Sugar/Activity Relationship*

Background: Say an observational study shows that for kids, sugar intake *x* and activity level *y* have positive *r*.

Question: Tell how each reason could be the case.

Response: #2

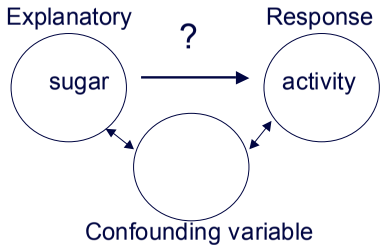


Example: *Reasons for Sugar/Activity Relationship*

Background: Say an observational study shows that for kids, sugar intake *x* and activity level *y* have positive *r*.

Question: Tell how each reason could be the case.

Response: 4&5

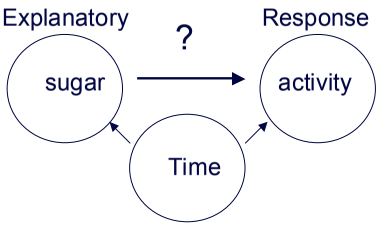


Example: *Reasons for Sugar/Activity Relationship*

Background: Say an observational study shows that for kids, sugar intake *x* and activity level *y* have positive *r*.

Question: Tell how each reason could be the case.

Response: #6

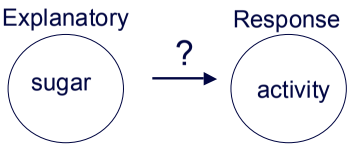


Example: *Reasons for Sugar/Activity Relationship*

Background: Say an observational study shows that for kids, sugar intake *x* and activity level *y* have positive *r*.

Question: Tell how each reason could be the case.

Response: #7



Results may be due to _____. [Small sample size? Or, many studies were conducted simultaneously, and one happened to turn out to have significant results?]

Making a Case for Causation in Obs. Study

Given all the possible ways two variables may be related, can we ever hope to claim causation? Yes, in a well-designed experiment, or an observational study that follows these guidelines:

1. There is a reasonable explanation for cause & effect.
2. The connection happens under varying conditions.
3. Potential confounding variables are ruled out.

Note: The third of these is the most difficult to achieve.

Example: *Guidelines for Evidence of Causation*

- **Background**: Consider article about kids wt & TV.
- **Question:** Does it meet the recommended guidelines?
- **Response:**
  - Reasonable explanation?

_____

Connection happens under varying conditions?

_____

Potential confounding variables ruled out?

_____

Note Dr. Robinson's comment. Not coincidentally, the same Dr. Robinson soon published results of his own study...

Example: *Comparing Evidence of Causation*

- **Background**:Compare 2 articles about kids' wt & TV.
- **Question:** Why are the 2nd article's claims more convincing?
- **Response:**

EXTRA CREDIT (Max. 5 pts.) Display with a scatterplot and describe (mention form, direction, and strength) the relationship between 2 quantitative variables from the survey [but not shoesize vs. height]. Is there an obvious choice for explanatory and response variables? Are there outliers or influential observations? Access the link 800surveyf06.txt at www.pitt.edu/~nancyp/stat-0800/index.html and see instructions to highlight, copy, and paste into MINITAB.