Exploring Data

Numerical Summaries for Relationships between Variables

How two variables vary together

- We have already looked at single-variable measures of spread:
  \[ \text{variance} = s^2 = \frac{1}{n-1} \sum (x_i - \bar{x})^2 \]
- We now have pairs of two variables:
  \[(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\]
- We need a statistic that summarizes how they are spread out "together"

Correlation

- Correlation of two variables:
  \[ r = \frac{1}{n - 1} \sum \frac{(x_i - \bar{x})(y_i - \bar{y})}{s_x \cdot s_y} \]
- We divide by standard deviation of both X and Y, so correlation has no units

Measure of Strength

- Correlation is a measure of the strength of linear relationship between variables X and Y
- Correlation has a range between -1 and 1
  - \( r = 1 \) means the relationship between X and Y is exactly positive linear
  - \( r = -1 \) means the relationship between X and Y is exactly negative linear
  - \( r = 0 \) means that there is no linear relationship between X and Y
**Examples**

Education and Mortality  
$r = -0.51$

Draft Order and Birthday  
$r = -0.22$

http://guessthecorrelation.com

---

**Cautions about Correlation**

- Correlation is only a good statistic to use if the relationship is roughly linear
- Correlation can not be used to measure non-linear relationships
- Always plot your data to make sure that the relationship is roughly linear!

---

**Correlation with Non-linear Data**

- A high correlation does not always mean a strong linear relationship
- A low correlation does not always mean no relationship (just no linear relationship)

---

**Linear Regression**

- If our X and Y variables do show a linear relationship, we can calculate a best fit line in addition to the correlation

\[ Y = a + b \cdot X \]

- The values \( a \) and \( b \) together are called the regression coefficients
  - \( a \) is called the intercept
  - \( b \) is called the slope

---

**Example: US Cities**

Mortality = \( a + b \cdot \)Education

- How to determine our "best" line ?
  - ie. best regression coefficients \( a \) and \( b \) ?

---

**Finding the "Best" Line**

- Want line that has smallest total "Y-residuals"

\[ \text{total residuals} = \sum (y_i - (a + b \cdot x_i))^2 \]
Best values for slope and intercept

- Line with smallest total residuals is called the “least-squares” line
- need calculus to find \( a \) and \( b \) that give “least-squares” line

\[
b = r \cdot \frac{s_y}{s_x} \quad a = \bar{y} - b \cdot \bar{x}
\]

Example: Education and Mortality

- Negative association means negative slope \( b \)

\[
\text{Mortality} = 1353.16 - 37.62 \cdot \text{Education}
\]

Interpreting the regression line

- The slope coefficient \( b \) is the average change you get in the \( Y \) variable if you increased the \( X \) variable by one
  - Eg. increasing education by one year means an average decrease of \( \approx 38 \) deaths per 100,000
- The intercept coefficient \( a \) is the average value of the \( Y \) variable when the \( X \) variable is equal to zero
  - Eg. an uneducated city (Education = 0) would have an average mortality of 1353 deaths per 100,000

Example: Crack Cocaine Sentences

- Extra gram of crack means extra 0.27 months (about 8 days) sentence on average
- Intercept is not always interpretable: having no crack gets you an 90 month sentence!

Prediction

- The regression equation can be used to predict our \( Y \) variable for a specific value of our \( X \) variable
  - Eg: \( \text{Sentence} = 90.4 + 0.269 \cdot \text{Quantity} \)
    - Person caught with 175 grams has a predicted sentence of 90.4 + 0.269 \cdot 175 = 137.5 months
  - Eg: \( \text{Mortality} = 1353.16 - 37.62 \cdot \text{Education} \)
    - City with only 6 median years of education has a predicted mortality of 1353.16 - 37.62 \cdot 6 = 1127 deaths per 100,000
Problems with Prediction

- Quality of predictions depends on the truth of the assumption of a linear relationship
  - Crack sentences aren’t really linear
- We shouldn’t extrapolate predictions beyond the range of data for our X variable
  - Eg: Sentence = 90.4 + 0.269 · Quantity
    - Having no crack still has a predicted sentence of 90 months!
- Eg: Mortality = 1353.16 - 37.62 · Education
  - A city with median education of 35 years is predicted to have no deaths at all!

The “Clemens Report”

- Done by Clemens’ agents: Hendricks Sports Mgmt
- Very long and involved report: 45 pages, 18,000 words
- Compared Clemens to three other pitchers
- Claim that Clemens does not have an unusual late career

Selection Bias!

- However, only looked at the 3 pitchers that are most similar to him in their late careers
- This set of comparison pitchers minimizes the chances that Clemens will look unique
- Want to look at a larger comparison set of pitchers: all starting pitchers with long careers – 32 pitchers (including Clemens) have had at least 15 years as starting pitcher with at least 3000 total innings pitched

Better Analysis

- Best fit line can be used to make comparisons between players easier
- In this case, curves are even better (Stat 112) because careers are non-linear

Conclusions

- Expanding set of comparison pitchers makes Clemens’ career path look more unusual
- Clemens has atypical trajectory relative to other pitchers
- Of course, these curves tell us nothing about why Clemens’ career path is unusual!
Next Class - Lecture 8

• Introduction to Probability
• Moore and McCabe: Section 4.1-4.2