ANNOUNCEMENTS:
• Grades available on eee for Week 1 clickers, Quiz and Discussion. If your clicker grade is missing, check next week before contacting me. If any other grades are missing let me know now.
• Quiz 1 answers now available (for your questions)
• If you are on the waiting list, have been doing the work, and still want to add, contact me.

TODAY: Sections 3.3 to 3.5.

HOMEWORK (due Wed, Jan 23):
Chapter 3: #42, 48, 74

Three tools for studying relationships between two quantitative variables:
• Scatterplot, a two-dimensional graph of data values
• Regression equation, an equation that describes the average relationship between a response and explanatory variable
• Correlation, a statistic that measures the strength and direction of a linear relationship

Recall, Positive/Negative Association:
• Two variables have a positive association when the values of one variable tend to increase as the values of the other variable increase.
• Two variables have a negative association when the values of one variable tend to decrease as the values of the other variable increase.

Positive Association: Height and Handspan
Taller people tend to have greater handspan measurements than shorter people do. (Why basketball players can “palm” the ball!) They have a positive association. The handspan and height measurements also seem to have a linear relationship.

Negative Association:
Driver Age and Maximum Legibility Distance of Highway Signs
• A research firm determined the maximum distance at which each of 30 drivers could read a newly designed sign.
• The 30 participants in the study ranged in age from 18 to 82 years old.
• We want to examine the relationship between age and the sign legibility distance.

Example 3.1 Height and Handspan
Data:
<table>
<thead>
<tr>
<th>Height (in.)</th>
<th>Span (cm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>71</td>
<td>23.5</td>
</tr>
<tr>
<td>69</td>
<td>22.0</td>
</tr>
<tr>
<td>66</td>
<td>18.5</td>
</tr>
<tr>
<td>64</td>
<td>20.5</td>
</tr>
<tr>
<td>71</td>
<td>21.0</td>
</tr>
<tr>
<td>72</td>
<td>24.0</td>
</tr>
<tr>
<td>67</td>
<td>19.5</td>
</tr>
<tr>
<td>65</td>
<td>20.5</td>
</tr>
<tr>
<td>76</td>
<td>24.5</td>
</tr>
<tr>
<td>67</td>
<td>20.0</td>
</tr>
<tr>
<td>70</td>
<td>23.0</td>
</tr>
<tr>
<td>62</td>
<td>17.0</td>
</tr>
</tbody>
</table>

Data shown are the first 12 observations of a data set that includes the heights (in inches) and fully stretched handspans (in centimeters) of 167 college students.

and so on, for $n = 167$ observations.
Example 3.2 Driver Age and Maximum Legibility Distance of Highway Signs

- We see a negative association with a linear pattern.
- We use a straight-line equation to model this relationship.

Neither positive nor negative association: The Development of Musical Preferences

- 108 participants in the study, ranged in age from 16 to 86 years old.
- Each rated 28 “top 10 songs” from a 50 year period.
- Song-specific age \( (x) \) = respondent’s age in the year the song was popular. (Negative value means person wasn’t born yet when song was popular.)
- Musical preference score \( (y) \) = amount song was rated above or below that person’s average rating. (Positive score => person liked song, etc.)

Example 3.3 The Development of Musical Preferences

- Popular music preferences acquired in late adolescence and early adulthood.
- The association is nonlinear.

Review of what we do with a regression line

When the best equation for describing the relationship between \( x \) and \( y \) is a straight line, the equation is called the regression line.

Two purposes of the regression line:
- to estimate the average value of \( y \) at any specified value of \( x \)
- to predict the value of \( y \) for an individual, given that individual’s \( x \) value

3.3 Measuring Strength and Direction with Correlation

Correlation \( r \) indicates the strength and the direction of a straight-line relationship.

- The strength of the linear relationship is determined by the closeness of the points to a straight line.
- The direction is determined by whether one variable generally increases or generally decreases when the other variable increases.

Interpretation of \( r \)

- \( r \) is always between –1 and +1
- \( r = -1 \) or \( +1 \) indicates a perfect linear relationship
  - \( r = +1 \) means all points are on a line with positive slope
  - \( r = -1 \) means all points are on a line with negative slope
- Magnitude of \( r \) indicates the strength of the linear relationship
- Sign indicates the direction of the association
- \( r = 0 \) indicates a slope of 0, so knowing \( x \) does not change the predicted value of \( y \)
Formula for $r$

$$r = \frac{1}{n-1} \sum \left( \frac{x_i - \bar{x}}{s_x} \right) \left( \frac{y_i - \bar{y}}{s_y} \right)$$

- Easiest to compute using calculator or computer!
- Notice that it is the product of the “sample” standardized ($z$) score for $x$ and for $y$, multiplied for each point, then added, then (almost) averaged.
- So, if $x$ and $y$ both have big $z$-scores for the same pairs, correlation will be large.

Example 3.1 Height and Handspan

Regression equation: Handspan = $-3.0 + 0.35$ Height

Correlation $r = +0.74$, a somewhat strong positive linear relationship.

Example 3.2 Driver Age and Legibility Distance of Highway Signs (again)

Regression equation: Distance = $577 - 3$(Age)

Correlation $r = -0.8$, a fairly strong negative linear association.

Example 3.12 Left and Right Handspans

If you know the span of a person’s right hand, can you accurately predict his/her left handspan?

Correlation $r = +0.95 =>$ a very strong positive linear relationship.

Example 3.13 Verbal SAT and GPA

Grade point averages (GPAs) and verbal SAT scores for a sample of 100 university students.

Correlation $r = +0.485 =>$ a moderately strong positive linear relationship.

Example 3.14 Age and Hours of TV Viewing

Relationship between age and hours of daily television viewing for 1299 survey respondents in the 2008 “General Social Survey.”

Correlation $r = +0.136 =>$ a weak connection.

Note: a few claimed to watch TV 24 hours/day!
Example 3.15 *Hours of Sleep and Hours of Study*

Relationship between reported hours of sleep the previous 24 hours and the reported hours of study during the same period for a sample of 116 college students.

Correlation $r = -0.36$

$\Rightarrow$ a not too strong negative association.

A different interpretation of $r$, or actually, $r^2$

- Recall the equation for the regression line:
  \[
  \hat{y} = b_0 + b_1 x
  \]
- Prediction Error or Residual:
  \[
  y - \hat{y} = \text{Difference between the observed value of } y \text{ and the predicted value.}
  \]
- Least Squares Regression Line: minimizes $SSE = \text{the sum of the squared residuals.}$

Example 3.2 *Driver Age and Legibility Distance of Highway Signs (again)*

Regression equation: $\hat{y} = 577 - 3x$

<table>
<thead>
<tr>
<th>$x$</th>
<th>$y$</th>
<th>$\hat{y}$</th>
<th>Residual</th>
</tr>
</thead>
<tbody>
<tr>
<td>18</td>
<td>510</td>
<td>577 - 3(18)=523</td>
<td>510 - 523 = -13</td>
</tr>
<tr>
<td>20</td>
<td>590</td>
<td>577 - 3(20)=517</td>
<td>590 - 517 = 73</td>
</tr>
<tr>
<td>22</td>
<td>516</td>
<td>577 - 3(22)=511</td>
<td>516 - 511 = 5</td>
</tr>
</tbody>
</table>

Can compute the residual for all 30 observations. Positive residual $\Rightarrow$ observed value higher than predicted. Negative residual $\Rightarrow$ observed value lower than predicted.

Ex 3.2 in R Commander:

* Coefficients:  
  * Estimate Std. Error t value Pr(>|t|)  
    * (Intercept) 576.6819 23.4709 24.570  < 2e-16 ***  
    * Age -3.0068 0.4243 -7.086 1.04e-07 ***  
  * Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
  * Residual standard error: 49.76 on 28 degrees of freedom  
  * Multiple R-squared: 0.642, Adjusted R-squared: 0.6292  

We will learn about this “multiple R-squared” next.

New interpretation, $r^2$

Squared correlation $r^2$ is between 0 and 1 and indicates the proportion of variation in the response ($y$) “explained” by knowing $x$.

$SSTO = \text{sum of squares total} = \text{sum of squared differences between observed } y \text{ values and } \bar{y}.$

We will break $SSTO$ into two pieces, $SSE + SSR$:

- $SSE = \text{sum of squared residuals, unexplained}$
- $SSR = \text{sum of squares due to regression or explained.}$

Sum of squared differences $(\bar{y} - \hat{y})$

New interpretation of $r^2$

$SSTO = SSR + SSE$

**Question:** How much of the total variability in the $y$ values (SSTO) is in the “explained” part (SSR)?

How much better can we predict $y$ when we know $x$ than when we don’t?

\[
 r^2 = \frac{SSR}{SSR + SSE}
\]
Data from Exercise 3.92

**Total variation** for each point = (actual y − mean y)

**Unexplained part** = residual = (actual y − predicted y)

Explained by knowing x = (predicted y − mean y)

<table>
<thead>
<tr>
<th>Weight</th>
<th>ChugTime</th>
</tr>
</thead>
<tbody>
<tr>
<td>59.06</td>
<td>5.42</td>
</tr>
<tr>
<td>61.67</td>
<td>5.68</td>
</tr>
<tr>
<td>63.28</td>
<td>5.94</td>
</tr>
<tr>
<td>64.89</td>
<td>6.20</td>
</tr>
</tbody>
</table>

**Total variation** summed over all points = SSTO = 36.6

**Unexplained part** summed over all points = SSE = 13.9

Explained by knowing x summed = SSR = 22.7

62% of the variability in chug times is explained by knowing the weight of the person.

\[ r^2 = \frac{SSR}{SSTO} = \frac{22.7}{36.6} = 62\% \]

---

**Example: Height and Weight of 43 males**

The regression equation is

\[ \text{Weight} = 390 + 1.33 \text{Height} \]

R-Sq = 32.3% =>

The variable height explains 32.3% of the variation in the weights of college men.

---

**Ex 3.12 in R: Left and Right Handspans**

- **Interpretation of** \( r^2 \) **for other examples**

**Example 3.12: Left and Right Handspans**

\( r^2 = 0.90 \Rightarrow \) Span of one hand is very predictable from span of other hand.

**Example 3.14: TV viewing and Age**

\( r^2 = 0.018 \Rightarrow \) only about 1.8%

Knowing a person’s age doesn’t help much in predicting amount of daily TV viewing.

---

**3.4 Difficulties and Disasters in interpreting correlation**

- Extrapolation beyond the range where x was measured
- Allowing outliers to overly influence the results
- Combining groups inappropriately
- Using correlation and a straight-line equation to describe curvilinear data
Extrapolation

- Usually a bad idea to use a regression equation to **predict** values far outside the range where the original data fell.

- No guarantee that the relationship will **continue** beyond the range for which we have observed data.

---

Exercise 3.9: 20 cities in US

\( x= \text{latitude, } y= \text{average Aug temp} \)

Intercept = 114
Slope = -1.00
For instance, Irvine latitude = 33.4, so predict average August temp to be:
\[ 114 - 33.4 = 80.6 \] degrees
(Actual = 74)

---

Groups and Outliers

- Can use different plotting symbols or colors to represent different subgroups.

- Look for **outliers**: points that have an unusual combination of data values.

---

Example 3.4 Height and Foot Length Outliers

Regression equation
\begin{align*}
\text{uncorrected data:} & \quad 15.4 + 0.13 \text{ height} \\
\text{corrected data:} & \quad -3.2 + 0.42 \text{ height}
\end{align*}

Correlation
\begin{align*}
\text{uncorrected data:} & \quad r = 0.28 \\
\text{corrected data:} & \quad r = 0.69
\end{align*}

Three outliers were data entry errors.

---

Example 3.18 Earthquakes in US 1850 to 2009 with magnitude > 7.0 and/or > 20 deaths

SF 1906 was an outlier. Other earthquakes were later and/or in more remote areas.

Correlation: all data, \( r = 0.26 \)
w/o SF, \( r \approx 0.824 \)
**Example 3.19 Height and Lead Feet**

Scatterplot of all data:
College student heights and responses to the question “What is the fastest you have ever driven a car?” \( r = .39 \)

Scatterplot by gender:
Combining two groups led to misleading correlation
\( r = .04; \sim .01 \)

**Example 3.20 Don’t Predict without a Plot**

Population of US (in millions) for each census year between 1790 and 2000.

Correlation: \( r = .96 \)
Regression Line: population = \(-2348 + 1.289(\text{Year})\)
Poor Prediction for Year 2030 = \(-2348 + 1.289(2030)\)
or about 269 million, current is already over 311 million!

### 3.5 Correlation Does Not Prove Causation

**Possible explanations for correlation:**

1. There really is causation (explanatory causes response).
   Ex: \( x = \% \text{ fat calories per day}; y = \% \text{ body fat} \)
   Higher fat intake does cause higher \% body fat.

2. Change in \( x \) may cause change in \( y \), but confounding variables make it hard to separate effects of each.
   Ex: \( x = \text{parents’ IQs}; y = \text{child’s IQ} \)
   Confounded by diet, environment, parents’ educational levels, quality of child’s education, etc.

### Additional reasons for observed correlation (other than \( x \) causes \( y \)):

3. No causation, but explanatory and response variables are both similarly affected by other variables
   Ex: \( x = \text{Verbal SAT}; y = \text{College GPA} \)
   Common cause for both being high or low are IQ, good study habits, good memory, etc.

4. Response variable is causing a change in the explanatory variable (opposite direction)
   Ex: Case study 1.7, \( x = \text{time on internet}; y = \text{depression} \)
   Maybe more depressed people spend more time on the internet, not the other way around.

### Additional examples and notes

- **Examples of “no causation, but explanatory and response variables are both affected by other variables”** is when both variables change over time, or both are related to population size.
- Correlation between total ice cream sales and total number of births in the US each year, 1960 to 2000.
- Correlation between number of ministers and number of bars for cities in California.
- Note: Sometimes correlation is just coincidence!

### Nonstatistical Considerations to Assess Cause and Effect (see page 653)

Here are some hints that may suggest cause and effect from observational studies:

- There is a reasonable explanation for how the cause and effect could occur.
- The relationship occurs under varying conditions in a number of studies.
- There is a “dose-response” relationship.
- Potential confounding variables are ruled out by measuring and analyzing them.
Applets to illustrate concepts

Links removed so you can read the text

HOMEWORK (due Wed, Jan 23):
3.42
3.48
3.74