ANNOUNCEMENTS:

- Remember that discussion today is not for credit. Go over R Commander. Go to 192 ICS, except at 4pm, go to 192 or 174 ICS.

TODAY: Sections 5.3 to 5.5.

Note this is a change made in the daily outline from what was posted earlier.

HOMEWORK (due Fri, Oct 8):

Chapter 5: #29, 34, 43, 51

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.

Example 5.1 *Height and Handspan*

Data:

Height (in.)	Span (cm)
71	23.5
69	22.0
66	18.5
64	20.5
71	21.0
72	24.0
67	19.5
65	20.5
76	24.5
67	20.0
70	23.0
62	17.0

and so on, for n = 167 observations.

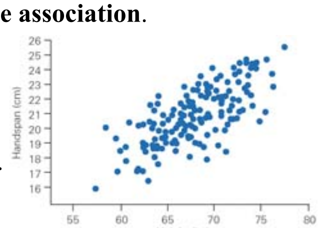
Data shown are the first 12 observations of a data set that includes the heights (in inches) and fully stretched handspans (in centimeters) of 167 college students.

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**.

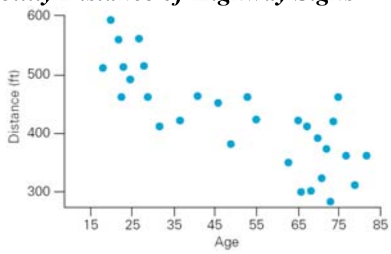


Negative Association:

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 5.2 *Driver Age and Maximum Legibility Distance of Highway Signs*

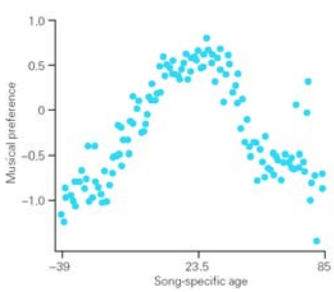


- We see a **negative** association with a **linear** pattern.
- We use a **straight-line equation** to model this relationship.

Neither positive nor negative association: *The Development of Musical Preferences*

- The 108 participants in the study ranged in age from 16 to 86 years old.
- Each rated 28 "top 10 songs" from 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 5.3 *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

5.3 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

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

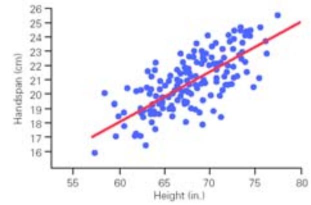
- Easiest to compute using calculator or computer!
- Notice that it is the product of the standardized (z) score for x and for y, multiplied for each point, then added, then (almost) averaged.
- So, if x and y both have big z-scores for the same pairs, correlation will be large.

Example 5.1 *Height and Handspan*

Regression equation: Handspan = –3.0 + 0.35 Height

Correlation r = +0.74,

a somewhat **strong positive linear** relationship.

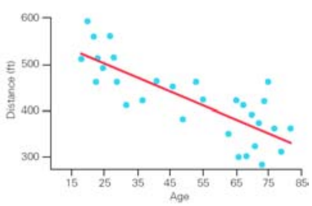


Example 5.2 *Driver Age and Legibility Distance of Highway Signs (again)*

Regression equation: Distance = 577 – 3(Age)

Correlation r = – 0.8,

a fairly **strong negative linear** association.

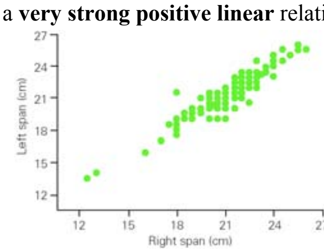


Example 5.11 *Left and Right Handspans*

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.

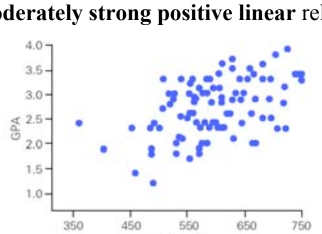


Example 5.12 *Verbal SAT and GPA*

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.

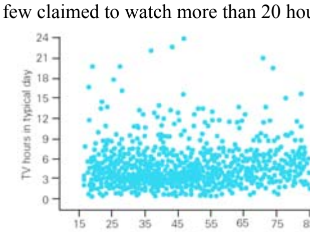


Example 5.13 *Age and Hours of TV Viewing*

Relationship between age and hours of daily television viewing for 1913 survey respondents.

Correlation r = 0.12 => a **weak connection**.

Note: a few claimed to watch more than 20 hours/day!

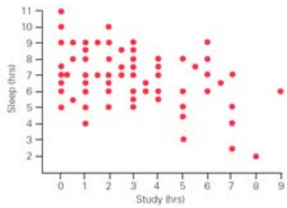


Example 5.14 *Hours of Sleep and Hours of Study*

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 $r = -0.36$

=> a **not too strong negative association**.



Recall: The Equation for the Regression Line: $\hat{y} = b_0 + b_1 x$

- **r^2 = a different interpretation of r.**
- **Prediction Error** = difference between the observed value of y and the predicted value
- **Residual** = $y - \hat{y}$ = *prediction error*
- **Least Squares Regression Line:**

  minimizes **SSE** = the sum of the squared residuals.

Example 5.2 *Driver Age and Legibility Distance of Highway Signs (again)*

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

x = Age	y = Distance	$\hat{y} = 577 - 3x$	Residual
18	510	577 – 3(18)=523	510 – 523 = -13
20	590	577 – 3(20)=517	590 – 517 = 73
22	516	577 – 3(22)=511	516 – 511 = 5

Can compute the residual for all 30 observations.
Positive residual => observed value *higher* than predicted.
Negative residual => observed value *lower* than predicted.

Ex 5.2 in R Commander:

Age and Sign Distance

```
• 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 of r^2

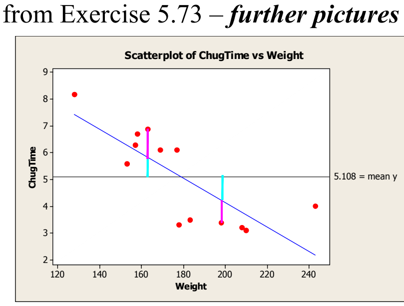
Squared correlation r^2 is between 0 and 1 and indicates the **proportion of variation in the response *(y)* explained by *x*.**

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

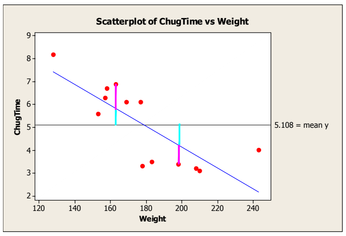
SSE = **sum of squared errors (residuals)** = sum of squared differences between observed y values and predicted values based on least squares line.

$$r^2 = \frac{SSTO - SSE}{SSTO}$$

Total variation for each point = (y − mean y)
Unexplained part = residual = (actual y – predicted y)
Explained by knowing *x* = (predicted y − mean y)
Data from Exercise 5.73 – ***further pictures on board***.



Total variation summed over all points = SSTO = 36.6
Unexplained part summed over all points = SSE = 13.9
Explained by knowing x summed over all points = 22.7
62% of the variability in chug times is explained by knowing the weight of the person



$$r^2 = \frac{SSTO - SSE}{SSTO} = \frac{36.6 - 13.9}{36.6} = \frac{22.7}{36.6} = 62\%$$

Interpretation of r^2 for other examples

Example 5.11: *Left and Right Handspans*
r^2 **= 0.90** => span of one hand is very predictable from span of other hand.

Example 5.13: *TV viewing and Age*
r^2 **= 0.014** => only about 1.4% knowing a person's age doesn't help much in predicting amount of daily TV viewing.

Ex 5.11 in R: *Left and Right Handspans*

```
• Coefficients:
•              Estimate Std. Error t value Pr(>|t|)
• (Intercept)   1.46346    0.47917   3.054  0.00258 **
• RtSpan        0.93830    0.02252  41.670  < 2e-16 ***
• ---
• Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.'
  0.1 ' ' 1

• Residual standard error: 0.6386 on 188 degrees of
  freedom
• Multiple R-squared: 0.9023,     Adjusted R-squared:
  0.9018
```


5.4 Difficulties and Disasters in interpreting correlation

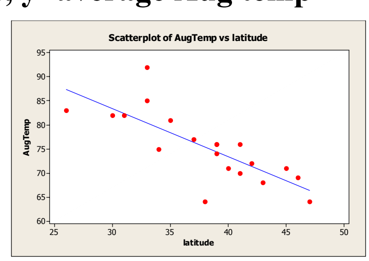
- Extrapolation beyond the range where x was measured
- Allowing outliers to overly influence the results
- Combining groups inappropriately
- Using correlation and a straight-line equation to describe curvilinear data

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.

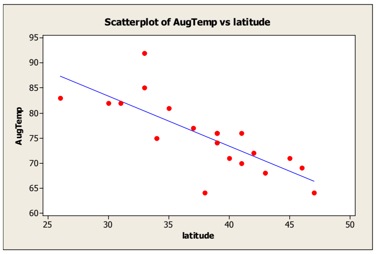
Exercise 5.6: 20 cities in US x=latitude, y=average Aug temp

Intercept = 114
Slope = −1.00
For instance, Irvine latitude = 33.4, so predict average August temp to be:
114 – 33.4 = 80.6 degrees



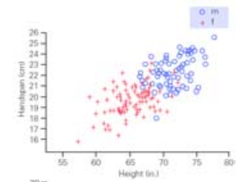
Extrapolation

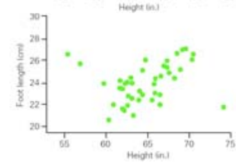
Range of latitudes is from 26 to 47. Would equation hold at the equator, latitude = 0? Predicted *average* temp = 114 degrees! Even worse for Jan. temperatures; intercept = 126.



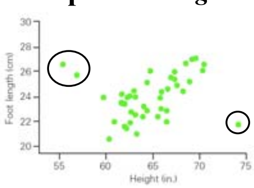
Groups and Outliers

- Can use different plotting symbols or colors to represent **different subgroups**.
- Look for **outliers**: points that have an usual combination of data values.





Example 5.4 *Height and Foot Length Outliers*



Three outliers were data entry errors.

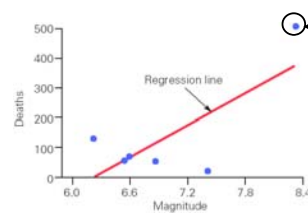
Regression equation

uncorrected data:	15.4 + 0.13 height
corrected data:	-3.2 + 0.42 height

Correlation

uncorrected data:	$r = 0.28$
corrected data:	$r = 0.69$

Example 5.16 *Earthquakes in US – an outlier*



San Francisco earthquake of 1906.

Other earthquakes were later and/or in more remote areas.

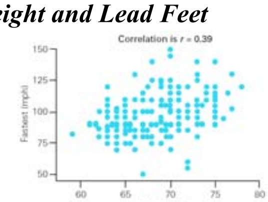
Correlation

all data: $r = 0.73$

w/o SF: $r = -0.96$

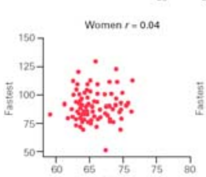
Example 5.17 *Height and Lead Feet*

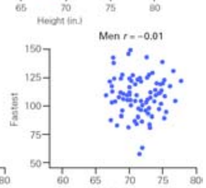
Scatterplot of all data: College student heights and responses to the question "What is the fastest you have ever driven a car?" $r = .39$



Scatterplot by gender: Combining two groups led to misleading correlation

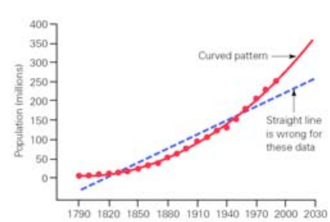
$r = .04; -.01$





Example 5.18 *Don't Predict without a Plot*

Population of US (in millions) for each census year between 1790 and 1990.



Correlation: $r = 0.96$

Regression Line: population = –2218 + 1.218(Year)

Poor Prediction for Year 2030 = –2218 + 1.218(2030) or about 255 million, current is already 308 million!

5.5 Correlation Does Not Prove Causation

Possible explanations for correlation:

1. There really is causation (explanatory causes response).
   Ex: x = % fat calories per day; y = % 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 = parents' IQs; y = 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. Explanatory and response variables are both affected by other variables
   Ex: x = Verbal SAT; y = 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 = time on internet, y = depression. Maybe more depressed people spend more time on the internet, not the other way around.

Additional examples and notes

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

Nonstatistical Considerations to Assess Cause and Effect (see page 707)

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

http://onlinestatbook.com/stat_sim/reg_by_eye/index.html

http://illuminations.nctm.org/LessonDetail.aspx?ID=L455

http://istics.net/stat/Correlations/

http://stat-www.berkeley.edu/~stark/Java/Html/Correlation.htm

What to notice

Outliers that *do not* fit the pattern of the rest of the data:
- Pull the regression line toward them
- Deflate the correlation

Outliers that *do* fit the pattern of the rest of the data, but are far away:
- Don't change the regression line much
- Inflate the correlation, sometimes by a lot