Thinking Straight Friday , May 9

Morning Session

- Review of Assignment and Sampling
- Lecture/discussion on correlation and causation

Afternoon Session beginning at 1 pm

- **Continuation** of Lecture/discussion on correlation and causation
- Workshop on Theories

Be Sure to pick up handout on Virtue Ethics to read along with Rachels Ch. 12 for next Tuesday, May 13

Week 6-1 Friday Session May 9, 2008

Am Statistics Session

Review of Assignment/Sampling Arguments

What is correlation?

How is it related to causation?

Pm Session

Critical Reasoning Workshop on Reconstructing and Criticizing Conceptual Theories

Be Sure to pick up handout on Virtue Ethics to read along with Rachels Ch. 12 for next Tuesday, May 13

Causal Arguments and Statistics

Form of Argument	*Example*
A is correlated with B	**Smoking is correlated with Heart Disease**
(likely) A causes B	**(likely) Smoking causes heart disease**

What makes makes for a good causal argument –Next Tuesday

What makes for a bad one. Today

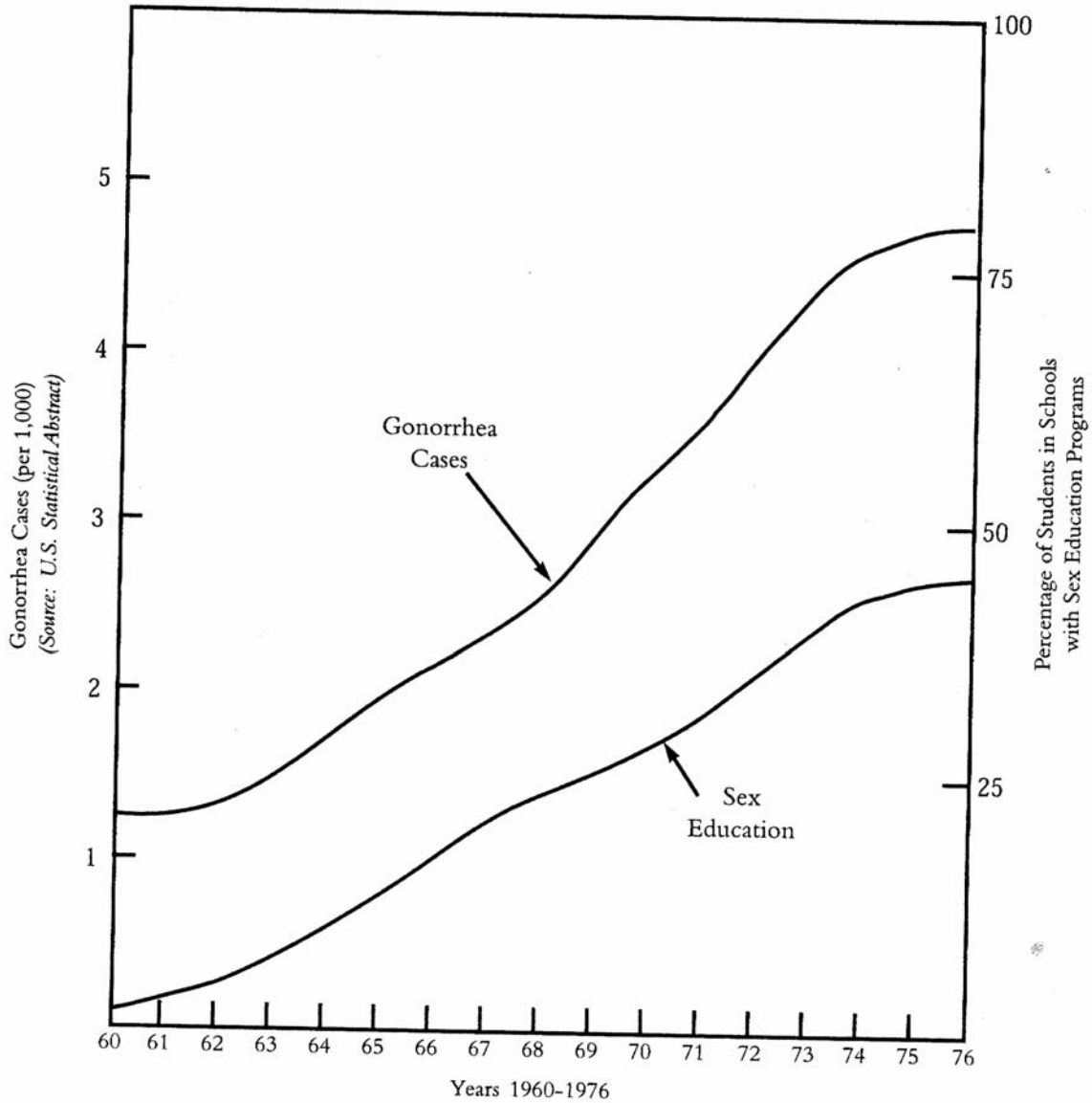
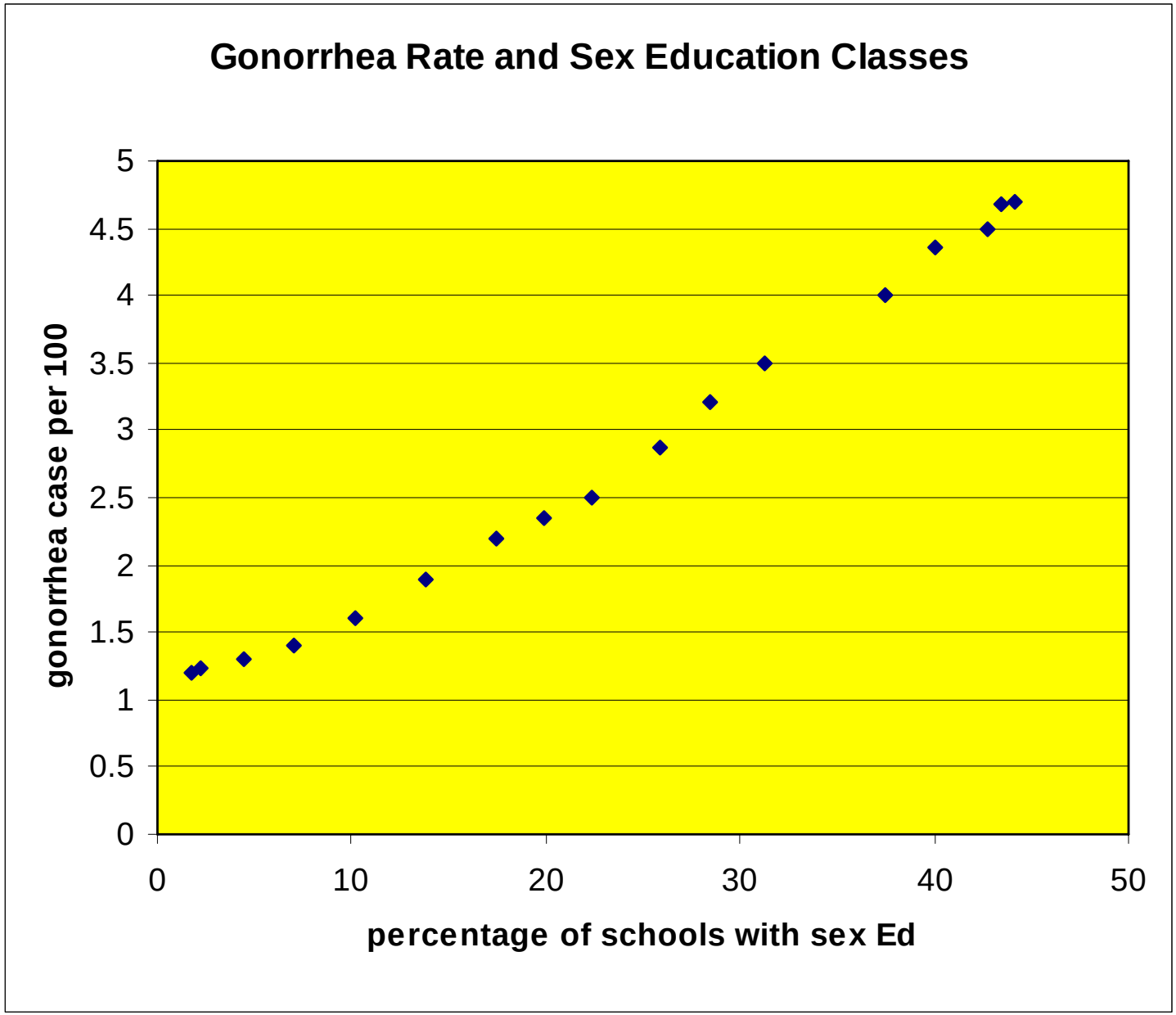


Figure 9.1 Rate of gonorrhea cases per 1,000 population (actual estimates) and percentage of students (largely fictional estimates) in high schools with sex education programs

Time Series suggests possible relateness

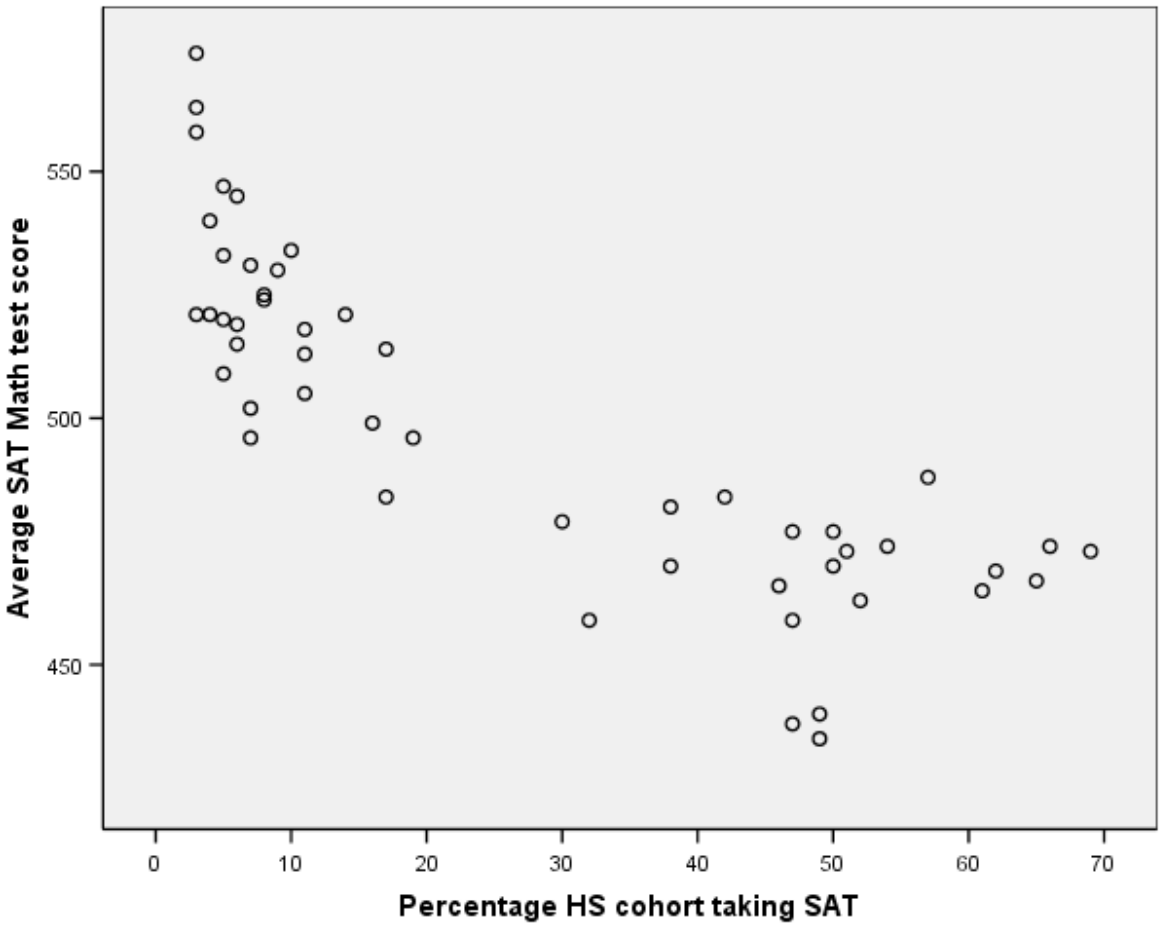
Scatter diagram/ scatter plot



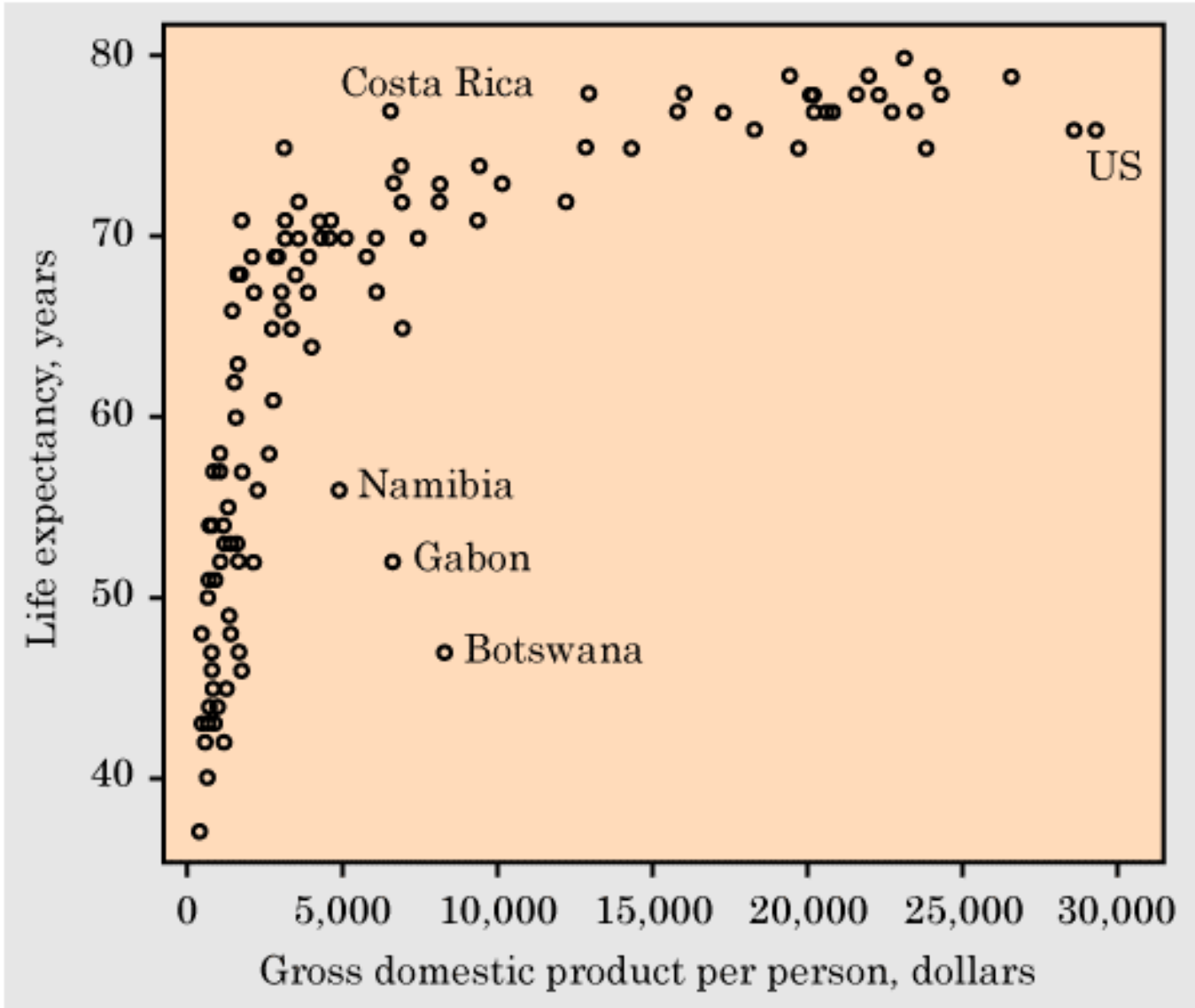
Scatterplot

- A Scatterplot is a way of showing the relationship between two variables

Plot

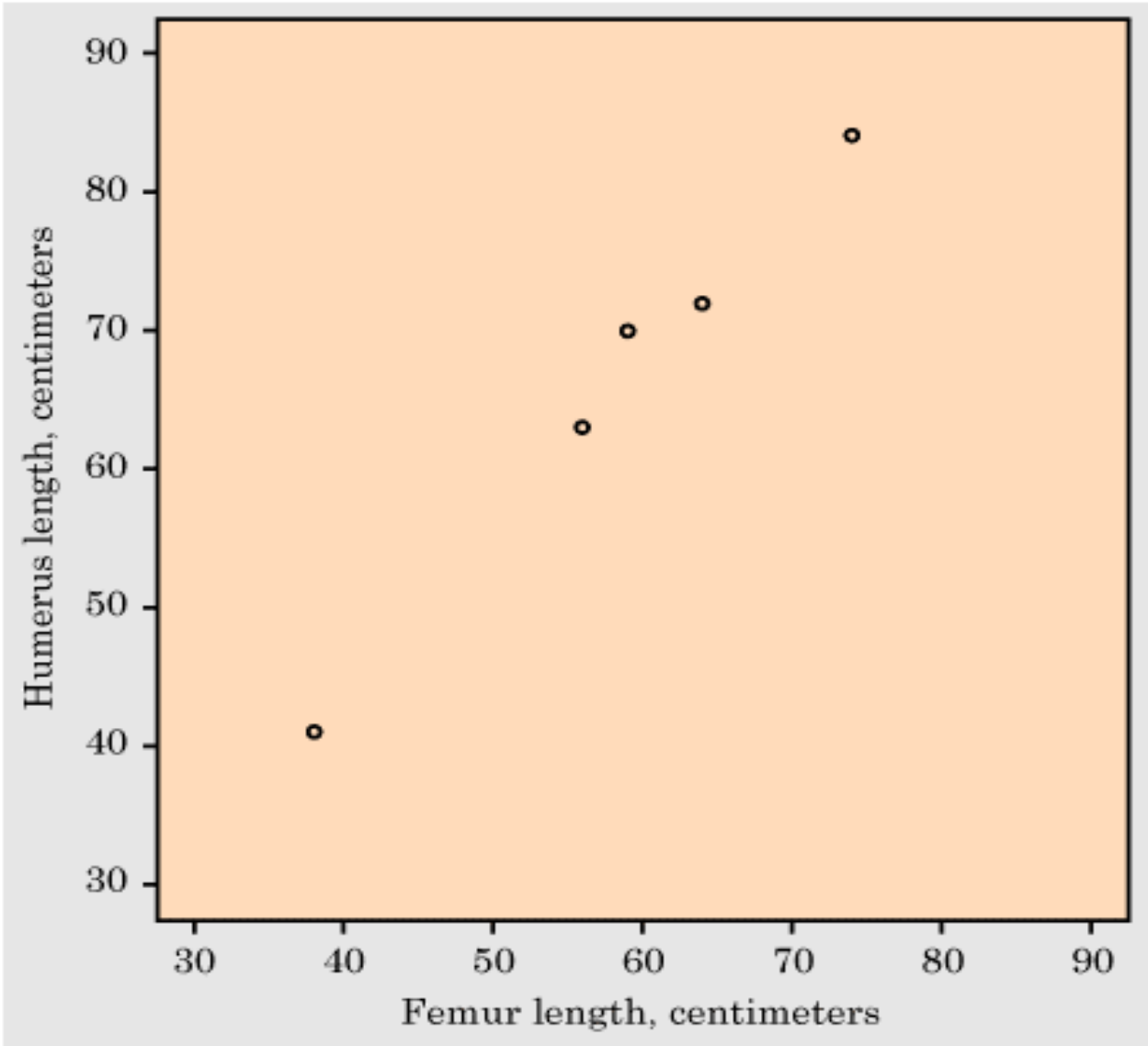


state	percent	math
ALABAMA	6	515
ALASKA	30	479
ARIZONA	11	505
ARKANSAS	4	521
CALIFORNIA	38	482
COLORADO	17	514
CONNECTICUT	69	473
DELAWARE	50	470
FLORIDA	38	470
GEORGIA	49	440
HAWAII	47	477
IDAHO	7	502
ILLINOIS	14	521
INDIANA	47	459
IOWA	3	574


80
70
60
50
40
Life expectancy, years
Costa Rica
US
Namibia
Gabon
Botswana
0
5,000
10,000
15,000
20,000
25,000
30,000
Gross domestic product per person, dollars

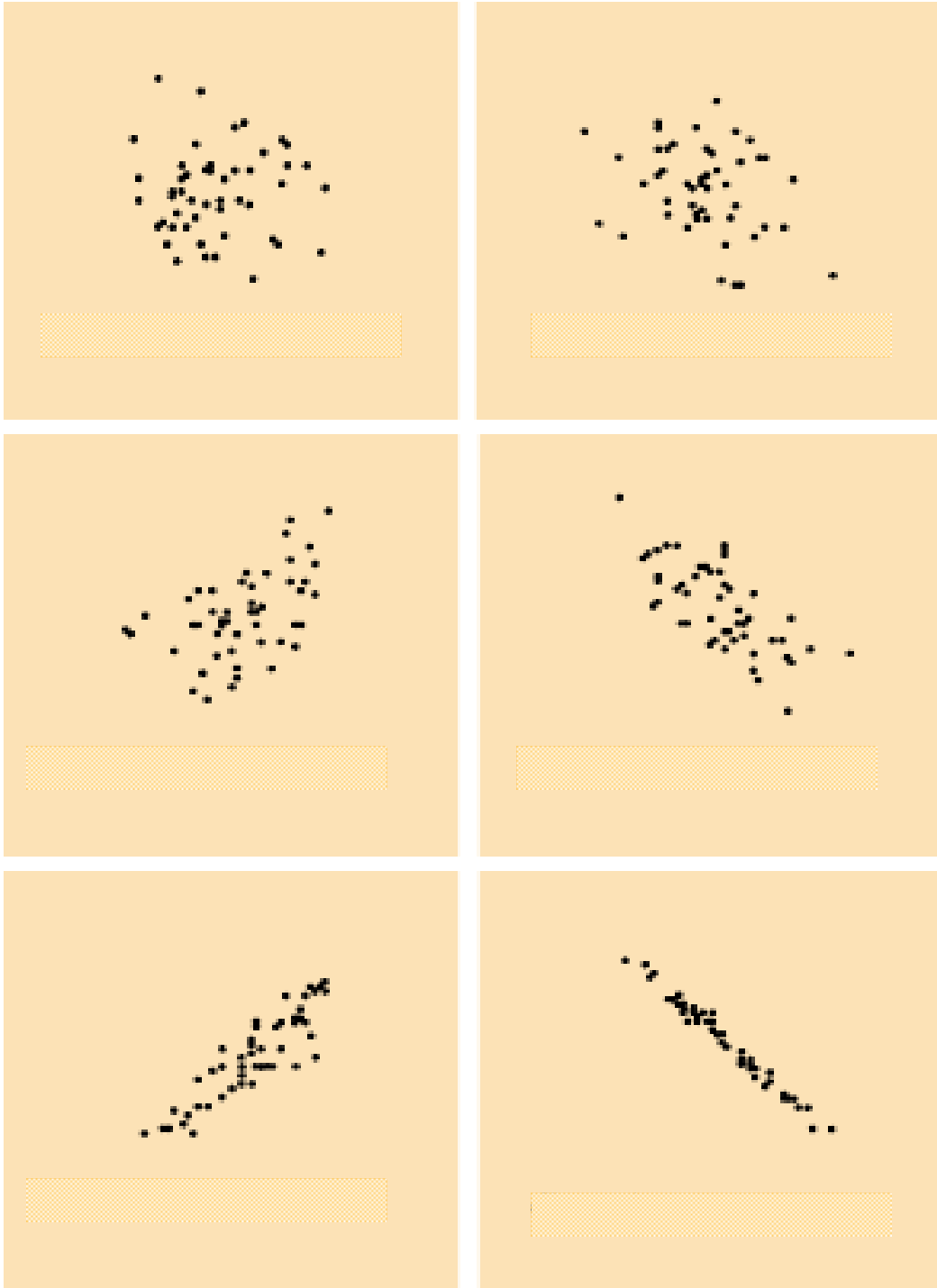
One Special Kind of Relationship

Some relationships are such that the points of a scatterplot tend to fall along a straight line -- **linear** relationship

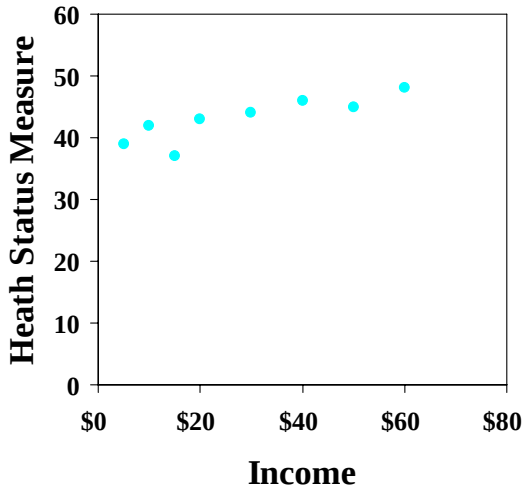

Humerus length, centimeters
90
80
70
60
50
40
30
30
40
50
60
70
80
90
Femur length, centimeters

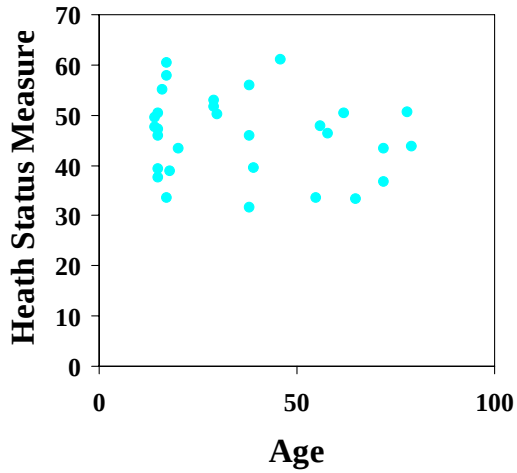
Properties of Linear relationships

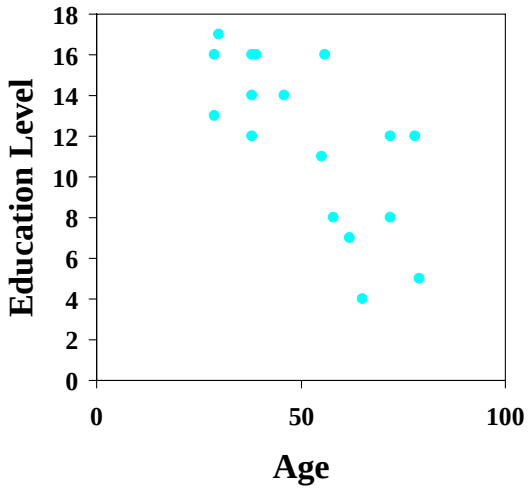
1. Direction
2. Strength

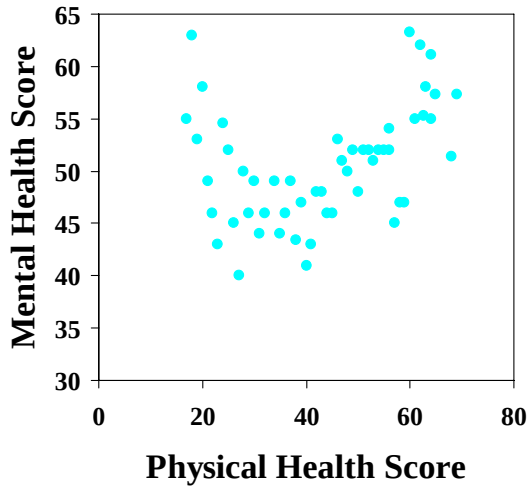


Examples of Relationships









Measuring Strength & Direction of a Linear Relationship

- How closely does a non-horizontal straight line fit the points of a scatterplot?
- The correlation coefficient (often referred to as just ***correlation***): **r**
  - measure of the *strength* of the relationship: the stronger the relationship, the larger the magnitude of r.
  - measure of the *direction* of the relationship: positive r indicates a positive relationship, negative r indicates a negative relationship.

Correlation Coefficient

- special values for r :
  - a perfect positive linear relationship would have r = +1
  - a perfect negative linear relationship would have r = -1
  - if there is no *linear* relationship, or if the scatterplot points are best fit by a horizontal line, then r = 0
  - *Note: r must be between -1 and +1, inclusive*

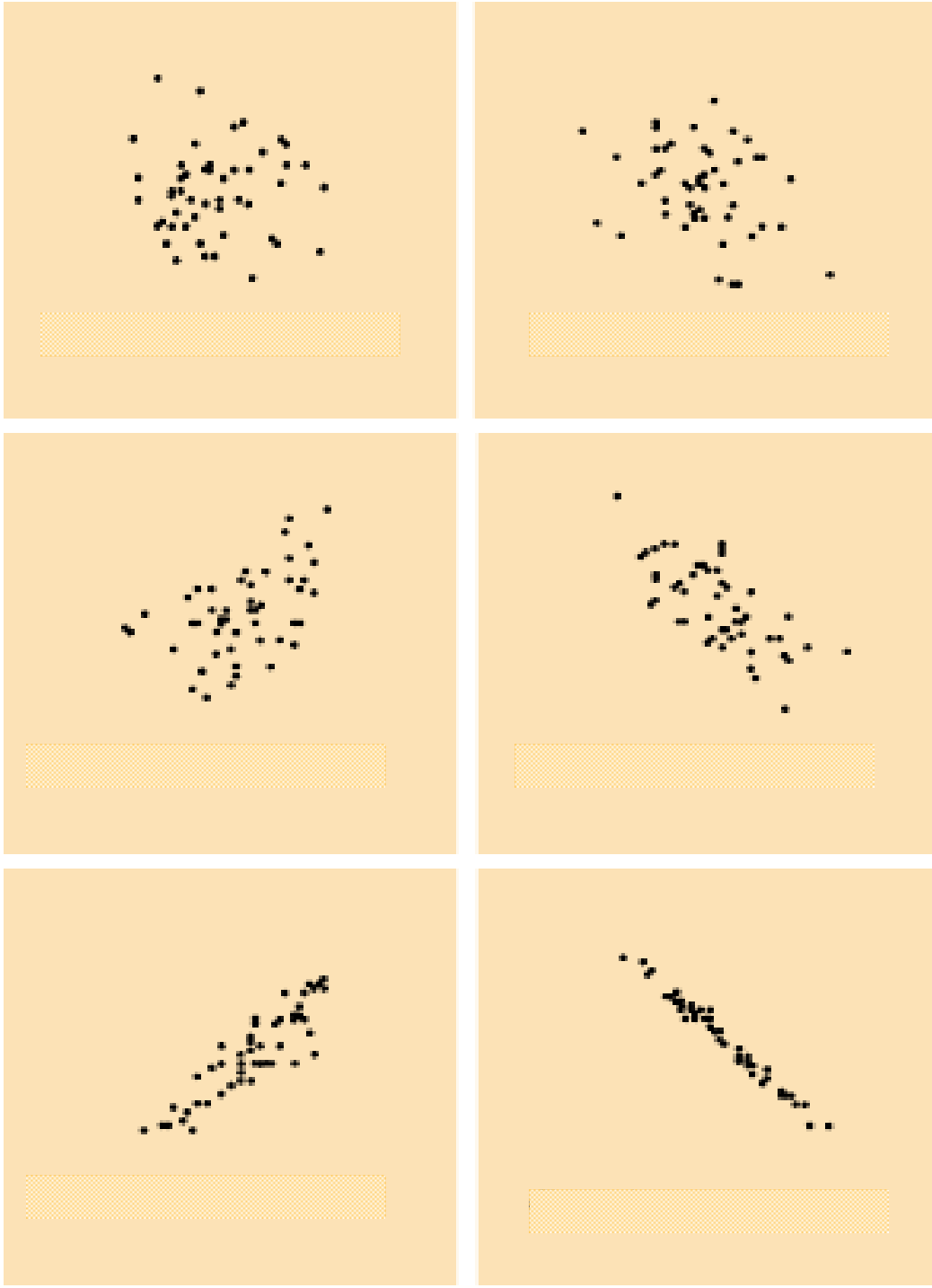
Plot

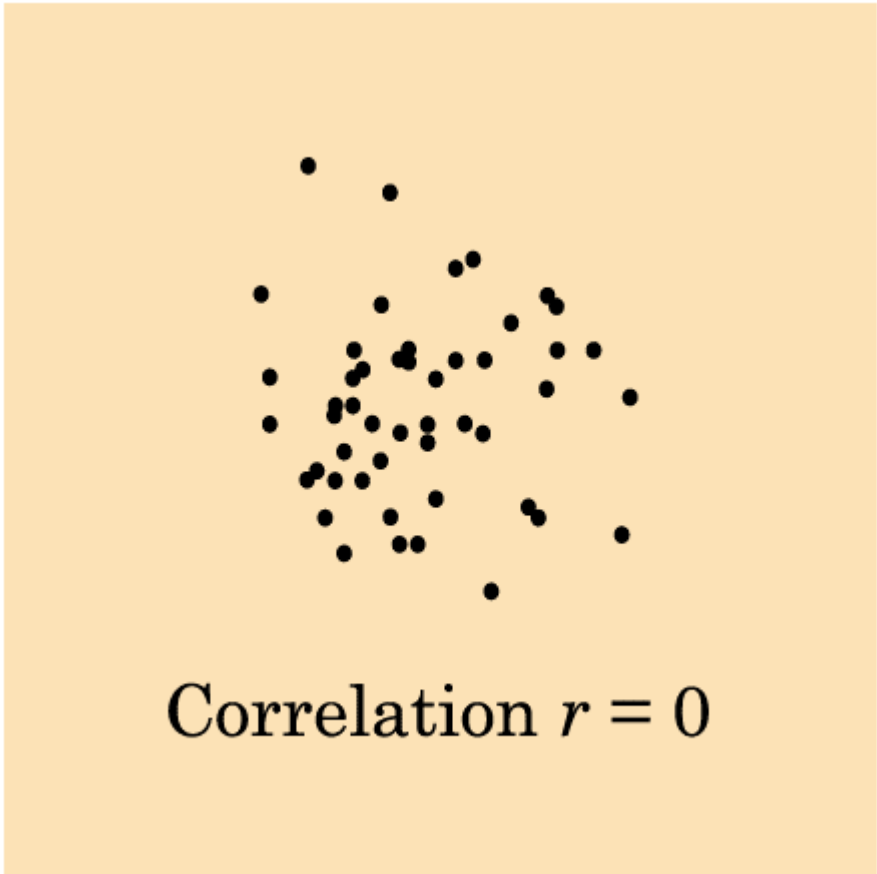
- r > 0: as one variable changes, the other variable tends to change in the *same* direction
- r < 0: as one variable changes, the other variable tends to change in the *opposite* direction

Examples of Correlations

- Husband's versus Wife's ages
  - r = .94
- Husband's versus Wife's heights
  - r = .36
- Professional Golfer's Putting Success: Distance of putt in feet versus percent success
  - r = -.94

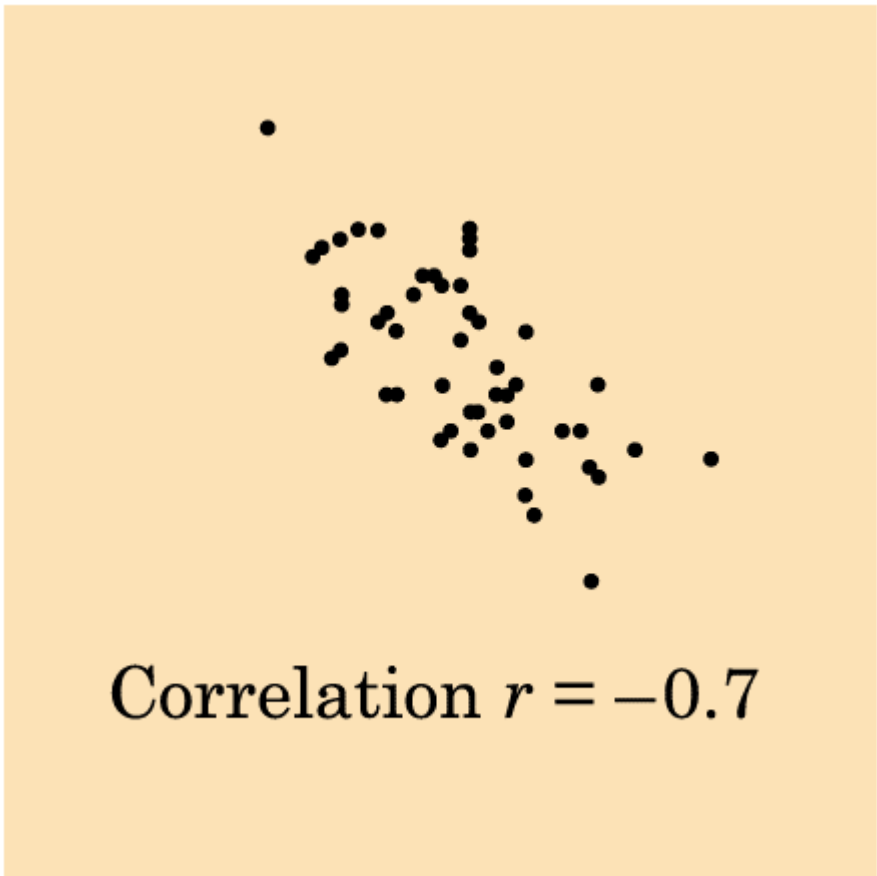
Plot



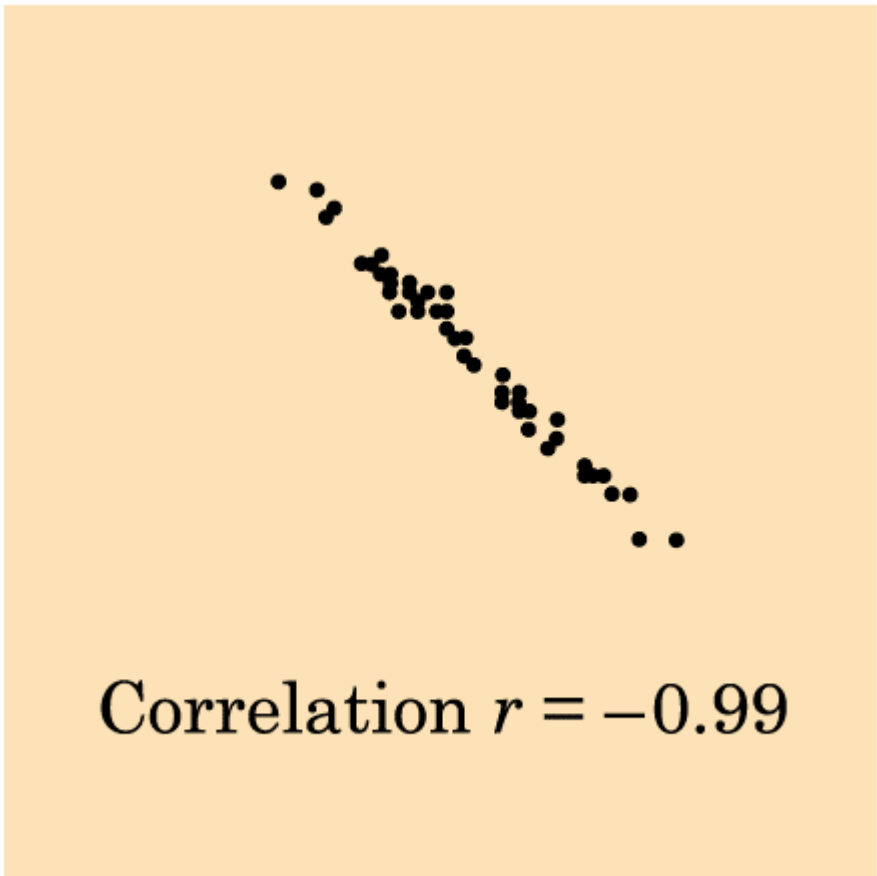

Correlation $r = 0$

Correlation $r = -0.3$

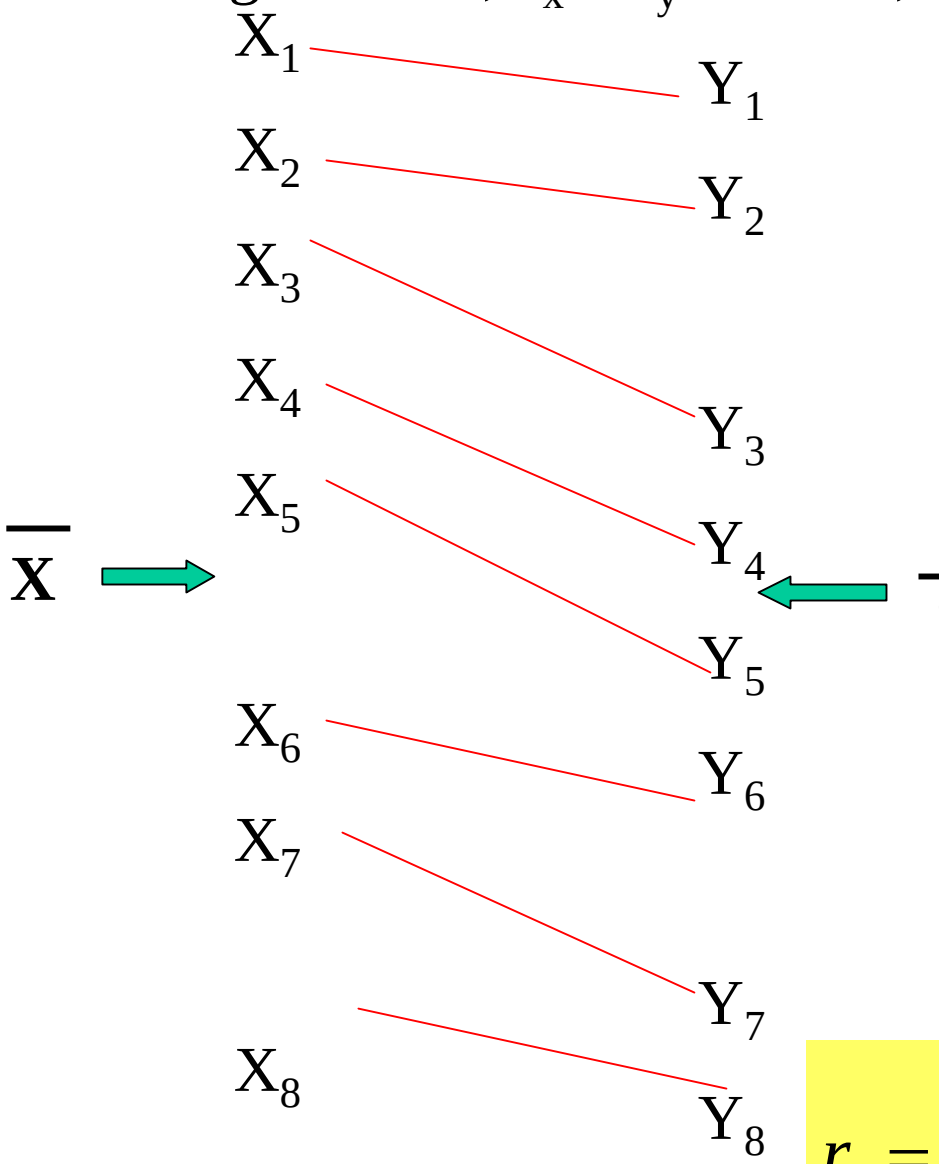
Correlation $r = 0.5$


Correlation $r = -0.7$

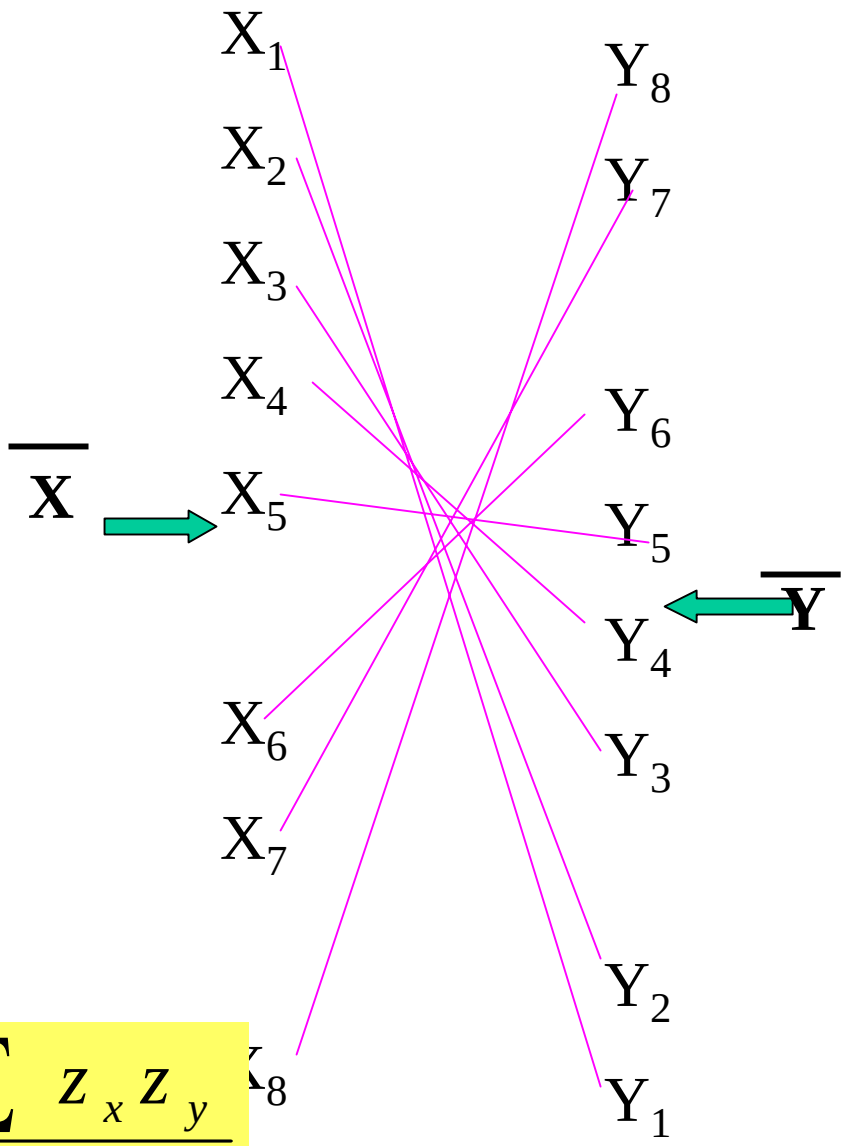
Correlation $r = 0.9$


Correlation $r = -0.99$

Strong Positive, $z_x * z_y$ Positive,

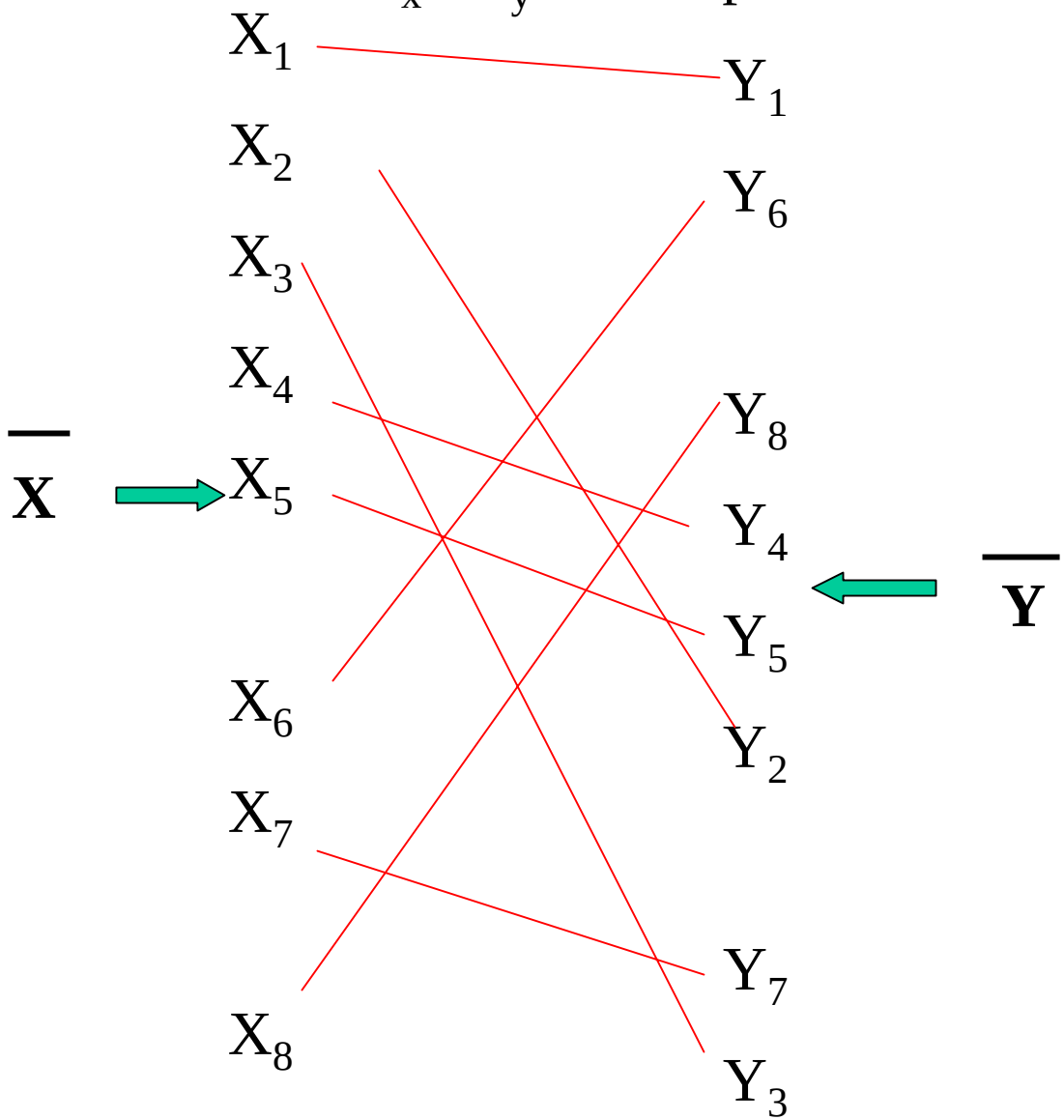


Strong Negative $z_x * z_y$ Positive



$$r = \frac{\sum z_x z_y}{n - 1}$$

Mixed $z_x * z_y$ mixed positive and negative

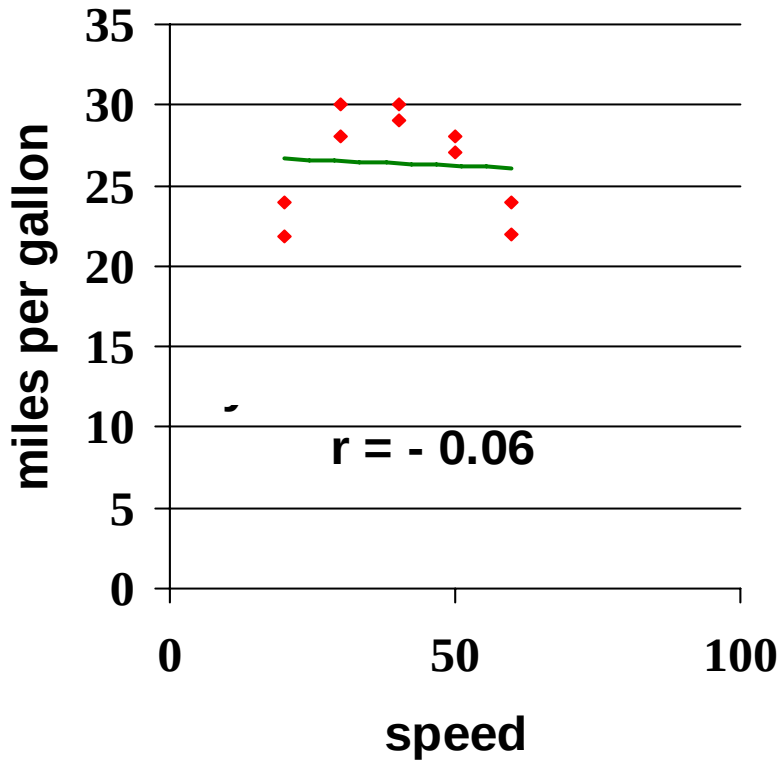


$$r = \frac{\sum z_x z_y}{n-1}$$

Not all Relationships are Linear
Miles per Gallon versus Speed

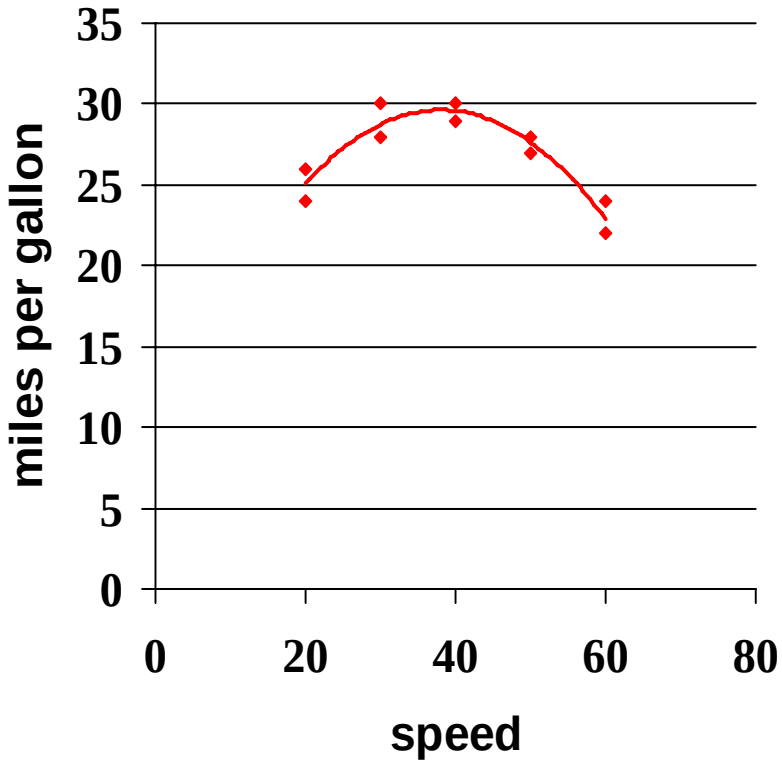
- Linear relationship?

- Speed chosen for each subject varies from 20 mph to 60 mph
- MPG varies from trial to trial, even at the same speed
- Statistical relationship



Not all Relationships are Linear
Miles per Gallon versus Speed

- Curved relationship (*r* is misleading)
- Speed chosen for each subject varies from 20 mph to 60 mph
- MPG varies from trial to trial, even at the same speed
- Statistical relationship



Problems with Correlations

Plot

- Outliers can inflate or deflate correlations
- Groups combined inappropriately may mask relationships (a third variable)
  - groups may have different relationships when separated

Linear Regression

- Objective: To *quantify* the linear relationship between an explanatory variable and response variable.

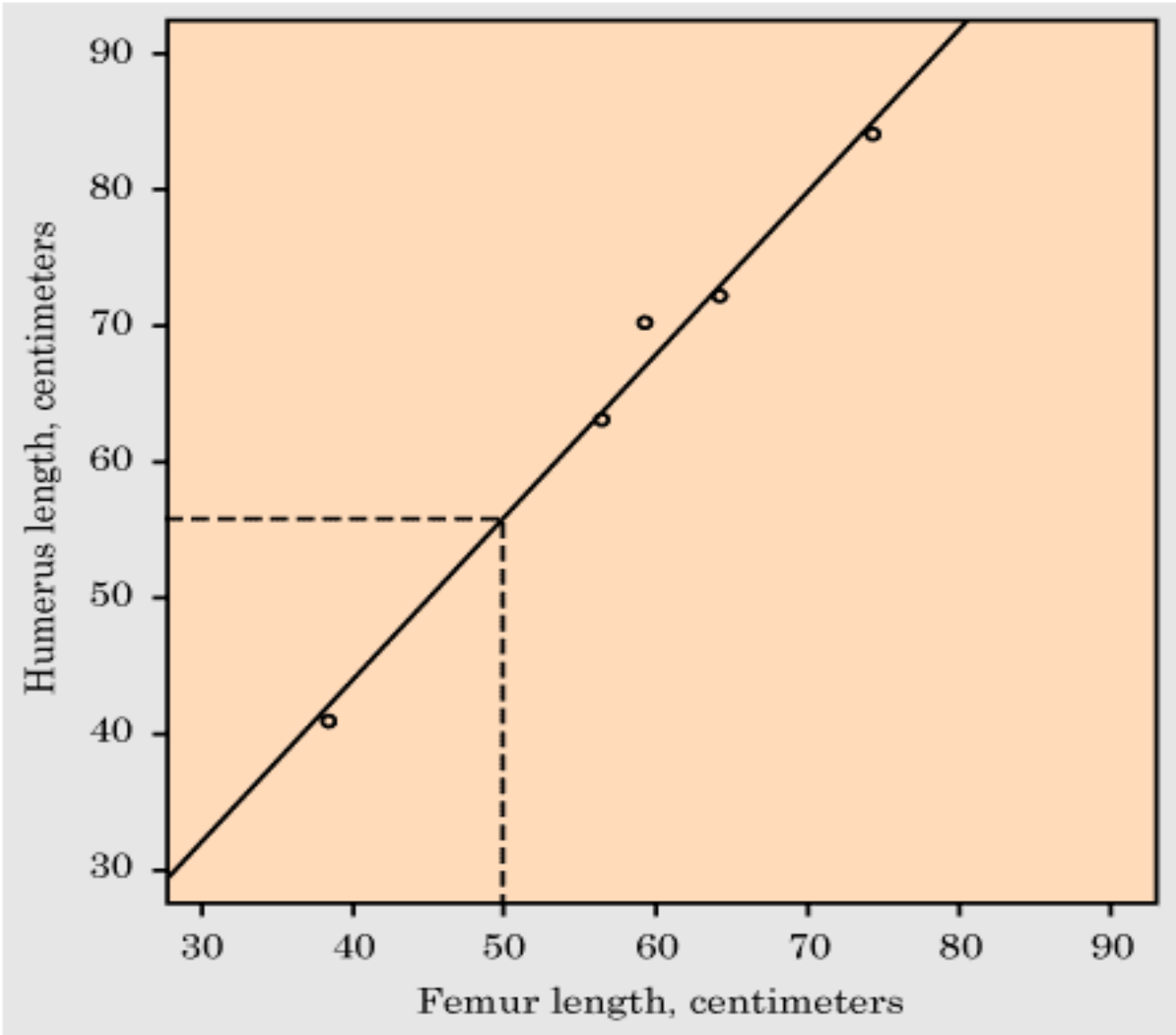
  We can then ***predict*** the average response for all subjects with a given value of the explanatory variable.

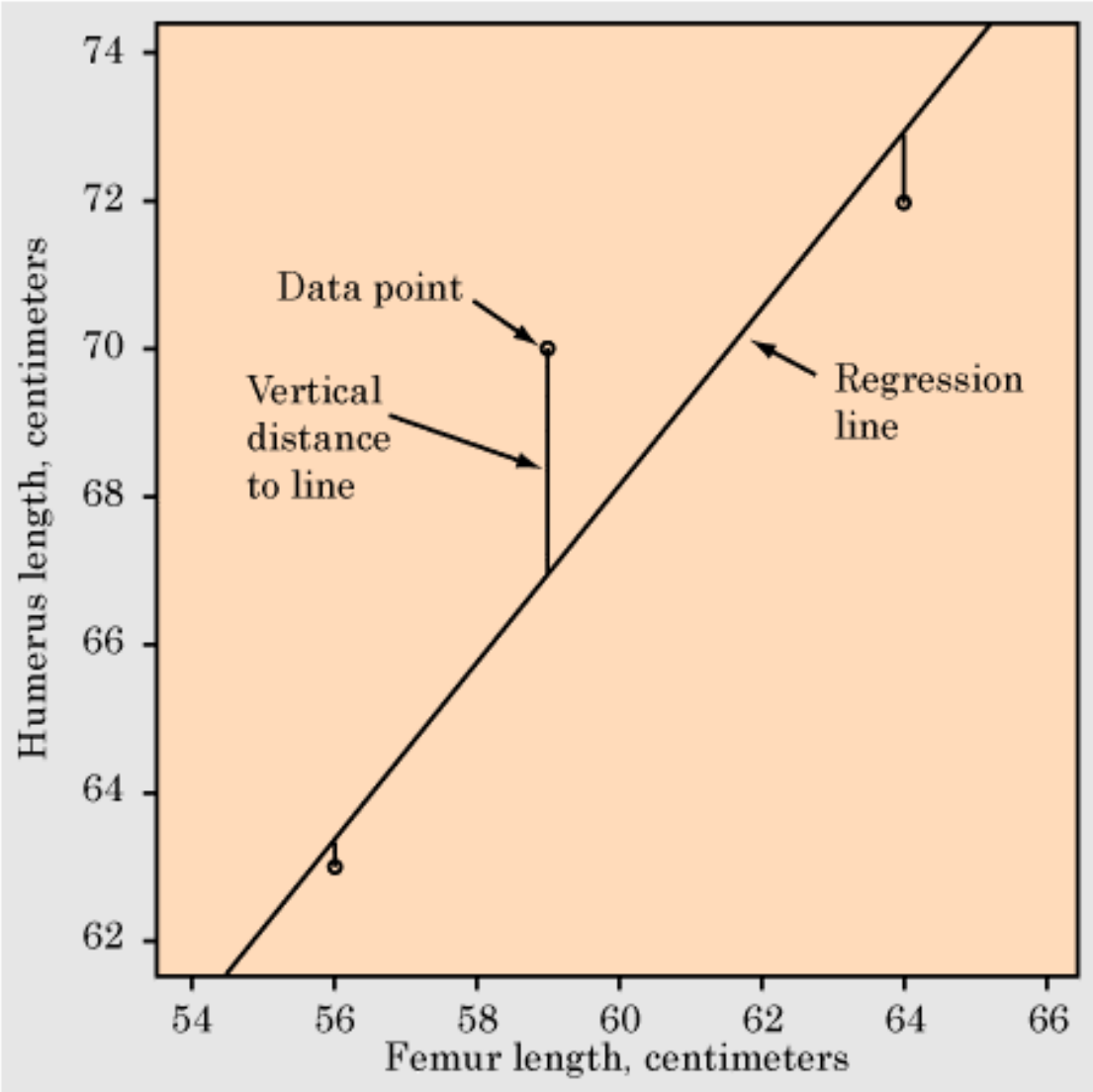
- Regression equation: $y = a + bx$ Plot
  - ***x*** is the value of the explanatory variable
  - ***y*** is the average value of the response variable

  - note that ***a*** and ***b*** are just the intercept and slope of a straight line
  - note that ***r*** and ***b*** are not the same thing, but their signs will agree

Least Squares

- Used to determine the “best” line
- We want the line to be as close as possible to the data points in the vertical (y) direction (since that is what we are trying to predict)
- **Least Squares**: use the line that minimizes the sum of the squares of the vertical distances of the data points from the line


Humerus length, centimeters
90
80
70
60
50
40
30
30
40
50
60
70
80
90
Femur length, centimeters

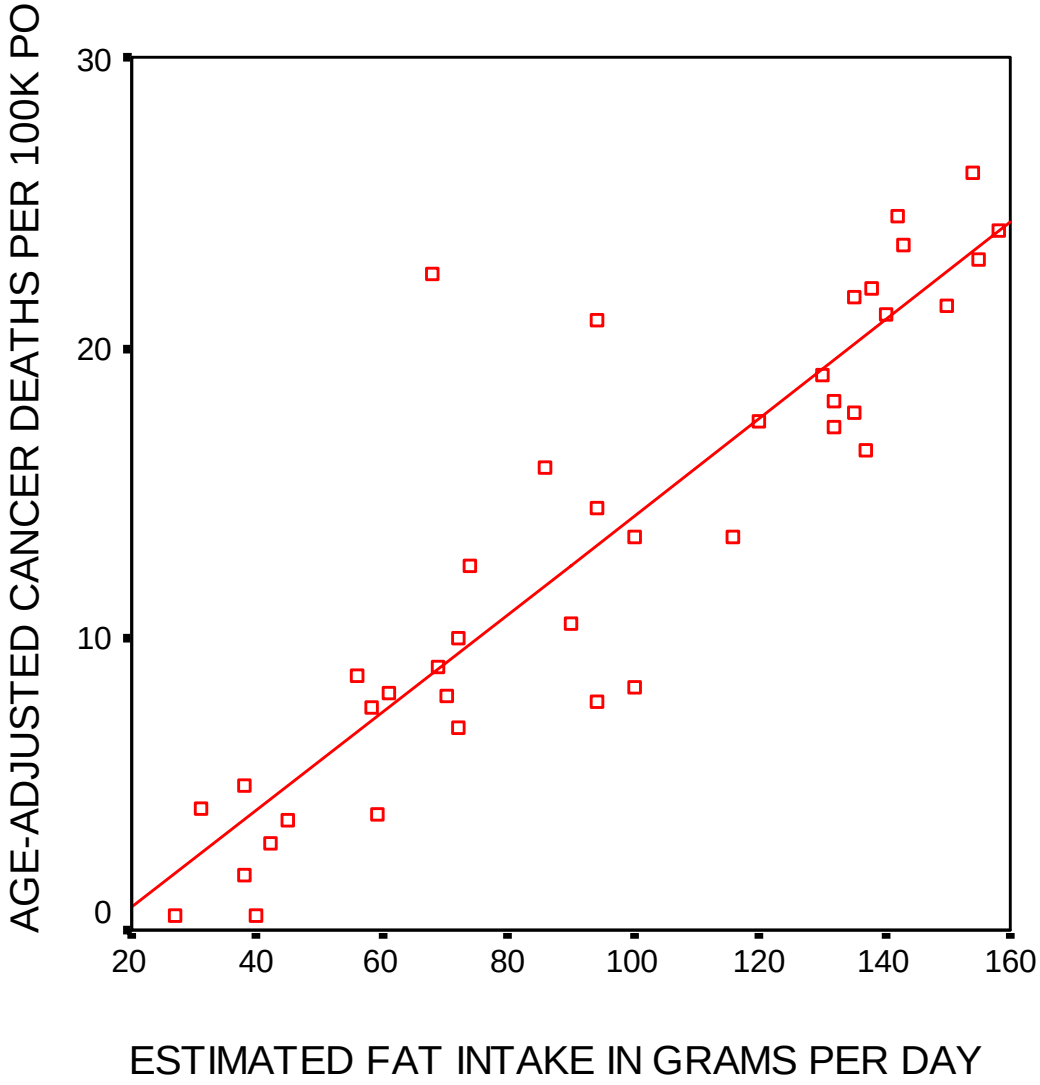

74
72
70
68
66
64
62
Humerus length, centimeters
Data point
Vertical
distance
to line
Regression
line
54
56
58
60
62
64
66
Femur length, centimeters

Gonorrhea Rate= .83 + .085 Percent, $r = .995$, $r^2 = .990$

Coefficient of Determination (R^2)

- Measures usefulness of regression prediction
- **R^2** (or **r^2**, the square of the correlation): measures how much variation in the values of the response variable (*y*) is explained by the regression line
  - r=1: R^2=1: regression line explains all (100%) of the variation in *y*
  - r=.7: R^2=.49: regression line explains almost half (50%) of the variation in *y*

Orientation of Data Least-Squares Regression Line



Equation for a line

$y = a + bx$

y-intercept

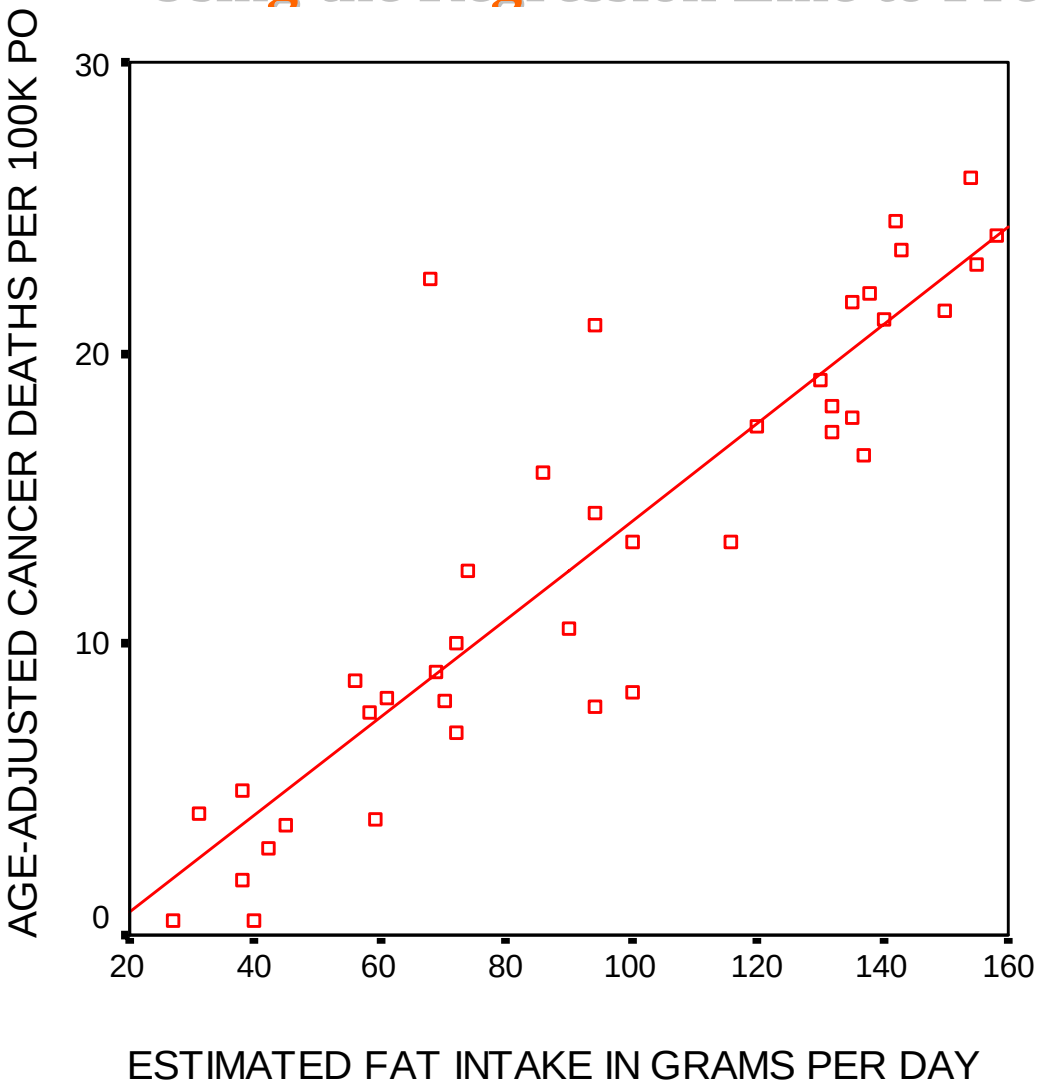
slope

Predicted y= -2.5 + .17x

$r = .894$, $r^2 = .799$

About 80% (79.9%) of the variation in cancer rate is (statistically) explained by the variation in fat intake

Using the Regression Line to Predict



Equation for a line

$y = a + bx$

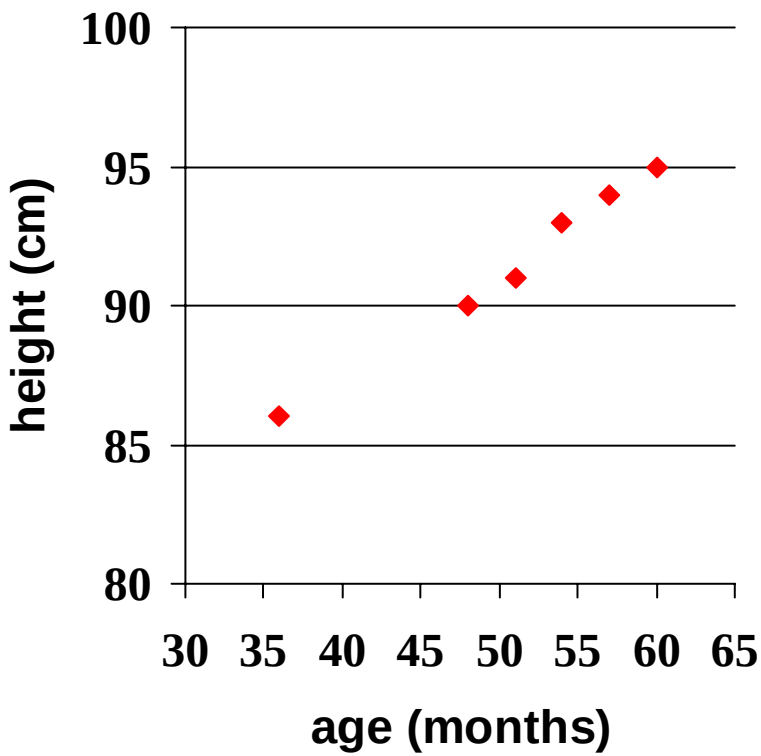
y-intercept **slope**

Predicted y= -2.5 + .17x

To predict a y for a given x, just plug into the equation. For example if the fat intake is 100 grams per day what cancer rate can we expect? Y= - 2.5 + .17*100 = -2.5 + 17 = 14.5 deaths/100k

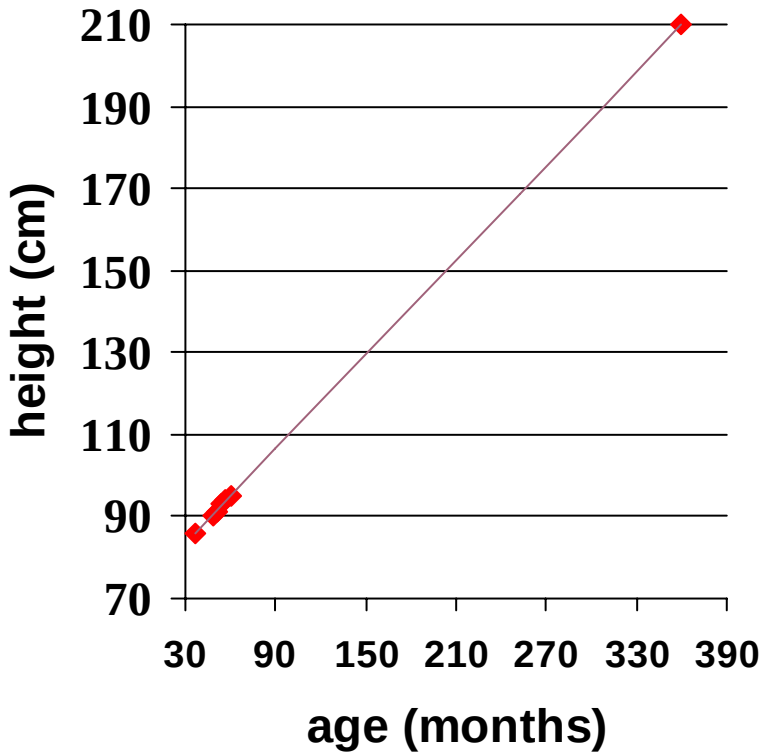
A Caution
Beware of Extrapolation

- Sarah's height was plotted against her age
- Can you predict her height at age 42 months?
- Can you predict her height at age 30 years (360 months)?



A Caution
Beware of Extrapolation

- Regression line:
  $y = 71.95 + .383\ x$
- height at age 42 months?
  $y = 88$
- height at age 30 years?
  $y = 209.8$
  - She is predicted to be 6’ 10.5” at age 30.



Correlation Does *Not* Imply Causation

Even very strong correlations may not correspond to a real causal relationship.

What makes makes for a good causal argument –Next Tuesday

What makes for a bad one. Today

Form of Argument	*Example*
associated	**associated**
A is correlated with B	**Smoking is correlated with Heart Disease**
(likely) A causes B	**(likely) Smoking causes heart disease**

Note: Correlation in the technical sense of linear correlation that can be measured by r (the Pearson correlation) is only one of a number of ways we might measure association

Five common criticism of Causal

Increase in Sex Ed classes is (positively) correlated (associated) with increased in gonorrhe

---

(likely) Increase in Sex Edu classes caused increase in gonorrhea

- **Coincidental**
  (A new strain of gonorrhea happened to emerge)
- **Both effects of the same underlying cause**
  (Increased sexual activity caused both)
- **Causal effect is genuine but insignificant**
  (Sex Ed classes encouraged risky sex for only a few)
- **Causal relation in the wrong direction**
  (Increase in gonorrhea caused introduction of more Sex Ed)
- **Causal relation may be complex**
  (Sex Ed caused changes in attitude that lead to increased sexual activity that lead to increased gonorrhea, but increased SDTs might have simultaneously caused more sex Ed courses to be introduced)

The Relationship May Be Just a Coincidence

We will see some strong correlations (or apparent associations) just by chance, even when the variables are not related in the population

1a. Coincidence (?)

Vaccines and Brain Damage

- A required whooping cough vaccine was blamed for seizures that caused brain damage
  - led to reduced production of vaccine (due to lawsuits)
- Study of 38,000 children found no evidence for the accusations (reported in *New York Times*)
  - "people confused association with cause-and-effect"
  - "virtually every kid received the vaccine…it was inevitable that, by chance, brain damage caused by other factors would occasionally occur in a recently vaccinated child"

Example 1b: In 1940 a psychologist conducted a study of the effect of propaganda on attitude toward a foreign government. He devised a test of attitude toward the German government and administered it to a group of American students . After reading German propaganda material for several months, the students were tested again to see if their attitudes had changed. Unfortunately, Germany attacked and conquered France while the experiment was in progress. There was a profound change of attitude toward the German government between test and retest. **Was the change in attitude caused by exposure to propaganda?**

Example 1c: A high school Latin teacher wished to demonstrate the favorable effect of studying Latin on mastery of English. She therefore obtained from school records the scores of all seniors on a standard English-proficiency examination. The average score for seniors who had studied Latin was much higher than the average score for those who had not. The Latin teacher concluded that "the study of Latin greatly improves one's command of English."

Taking Latin was associated with higher exam scores that showed a better command of English

---

(likely) Taking Latin causes (greatly improves) command of English

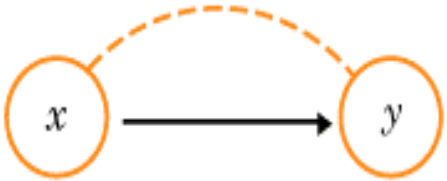
Criticize this Argument

2. Joint effect of a Common Cause

- Apparent Cause Divorce among men
- Apparent Effect : Percent abusing alcohol

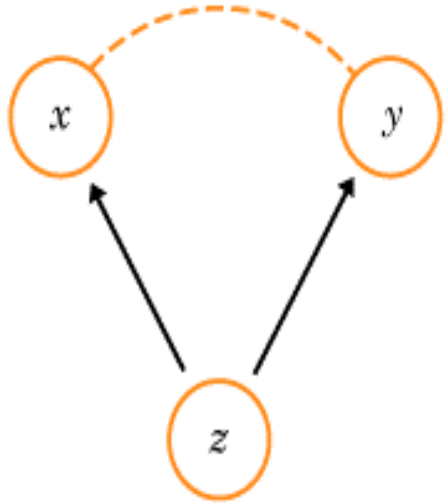
  - Both may result from an unhappy marriage.

Full Fledged Cause



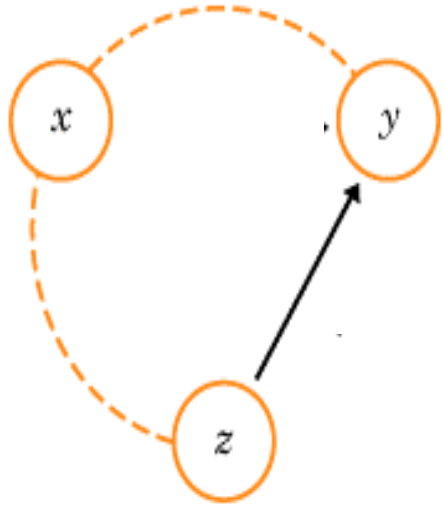
(a) Causation

Joint Effect of Common Cause



(b) Common response

Coincidental Correlation



(c) Confounding

Correlation, association

Causation

3. Apparent Cause is not the most important Contributor

- Apparent Cause: Possession of gun in home
- Apparent Effect Response: Occurrence of a homicide

  - *tendency toward violence* may be another contributor

Apparent Effect is actually the cause

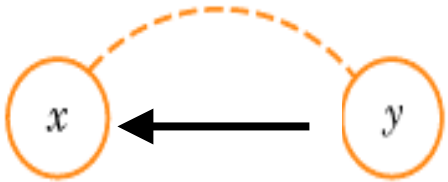
- Apparent Cause: Divorce among men
- Apparent Effect: Percent abusing alcohol

- Conclusion was that getting divorced caused alcohol abuse in men.
  - Could it be that alcohol abuse caused divorce?

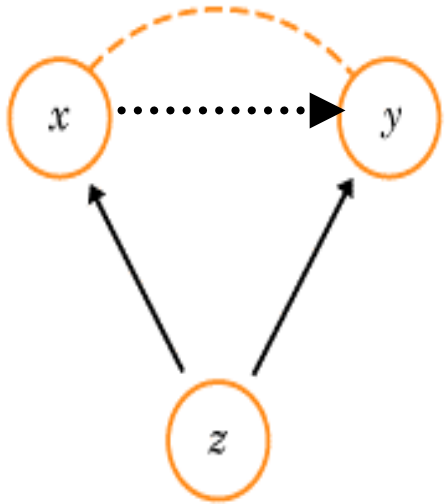
Another Complicating Factor

- Both divorces and suicides have increased dramatically since 1900.
- Are divorces causing suicides?
- Are suicides causing divorces???
- The population has increased dramatically since 1900 (causing both to increase).

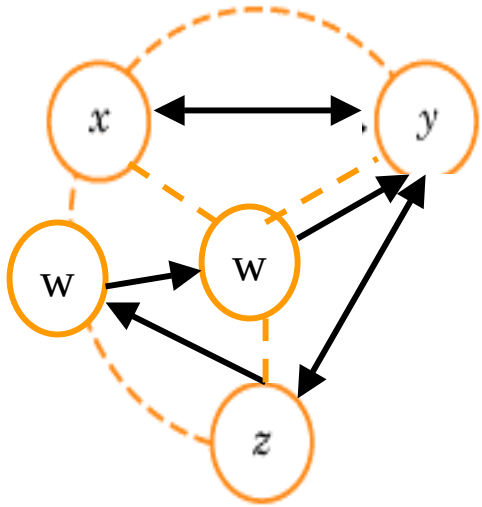
  - Better to investigate: Has the rate of divorce or the rate of suicide changed over time?



Apparent Cause in Wrong Direction



Genuine but Insignificant Cause



Complex causal Relation

But some Correlations are Causes

- Apparent Cause: pollen count from grasses
- Apparent Effect: percentage of people suffering from allergy symptoms

- Apparent Cause: amount of food eaten
- Apparent Effect: hunger level

Evidence of Causation

- A properly conducted experiment establishes the connection

Topic for Friday Morning Session


That's All Folks