STATISTICS 108

Outline for today:

- Go over syllabus
- Provide requested information – I will hand out blank paper and ask questions
- Brief introduction and hands-on activity

Information Sources

- Class webpage (also linked to my.ucdavis page for the class):

  http://stat.ucdavis.edu/~utts/st108

- Includes link to syllabus

  Syllabus.pdf

Purple Paper

Please provide the following information:

1. Name
2. Major
3. Year in school
4. Something interesting about yourself
5. Why you are taking this class

Purple paper, continued

6. On a 1 to 5 scale, how familiar and comfortable are you with these?

   1=none, 5 = completely

   a. Summation notation
   b. Normal distribution
   c. Hypothesis testing
   d. Confidence intervals
   e. Sampling distributions
   f. Scatter plots and simple linear regression

Purple paper, continued

7. The following data:
   a. Your height, in *inches*
   b. Your "handspan" in *centimeters*, defined as the distance covered on the ruler by your stretched hand from the tip of the thumb to the tip of the small finger.
   c. Your "residual" (to be explained!)

Regression

- Used to describe the relationship between a "response" variable and one or more "predictor" variables.
- Used to predict a future response using known, current values of the predictors.
- Switch to power point slides from Brooks/Cole to accompany "Mind On Statistics" by Utts/Heckard

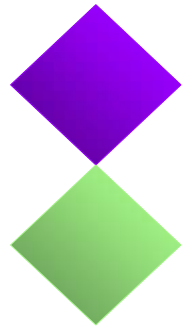
IMPORTANT NOTE

The remaining slides are from Power point presentations to accompany *Mind on Statistics,* by Utts and Heckard and are copyright Brooks/Cole. They are not to be copied or used for purposes other than this class.

Three Tools we will use …

- **Scatterplot**, a two-dimensional graph of data values
- **Correlation**, a statistic that measures the *strength* and *direction* of a linear relationship
- **Regression equation**, an equation that describes the average relationship between a response and explanatory variable

5.1 Looking for Patterns with Scatterplots



Questions to Ask about a Scatterplot

- What is the *average* pattern? Does it look like a straight line or is it curved?
- What is the direction of the pattern?
- How much do individual points vary from the average pattern?
- Are there any unusual data points?

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.

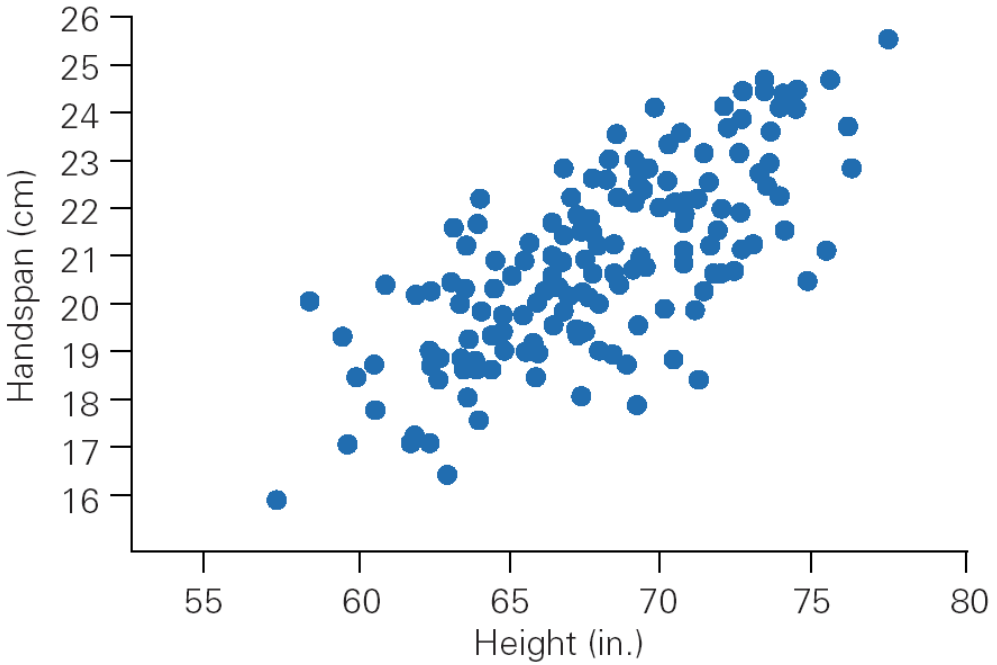
Data shown are the first 12 observations of a data set that includes the heights (in inches) and fully stretched handspans (in centimeters) of 167 college students.

Example 5.1 *Height and Handspan*

Taller people tend to have greater handspan measurements than shorter people do.

When two variables tend to increase together, we say that they have a **positive association**.

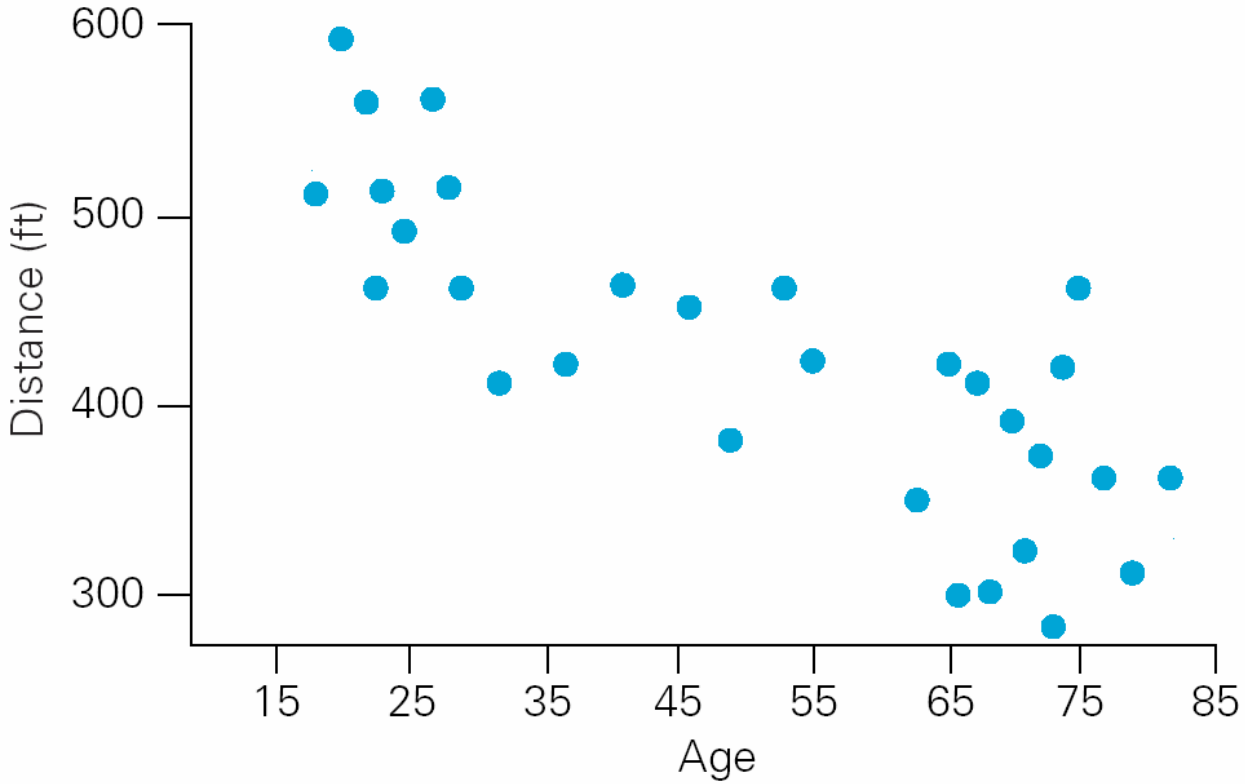
The handspan and height measurements may have a **linear relationship**.



Example 5.2 *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*

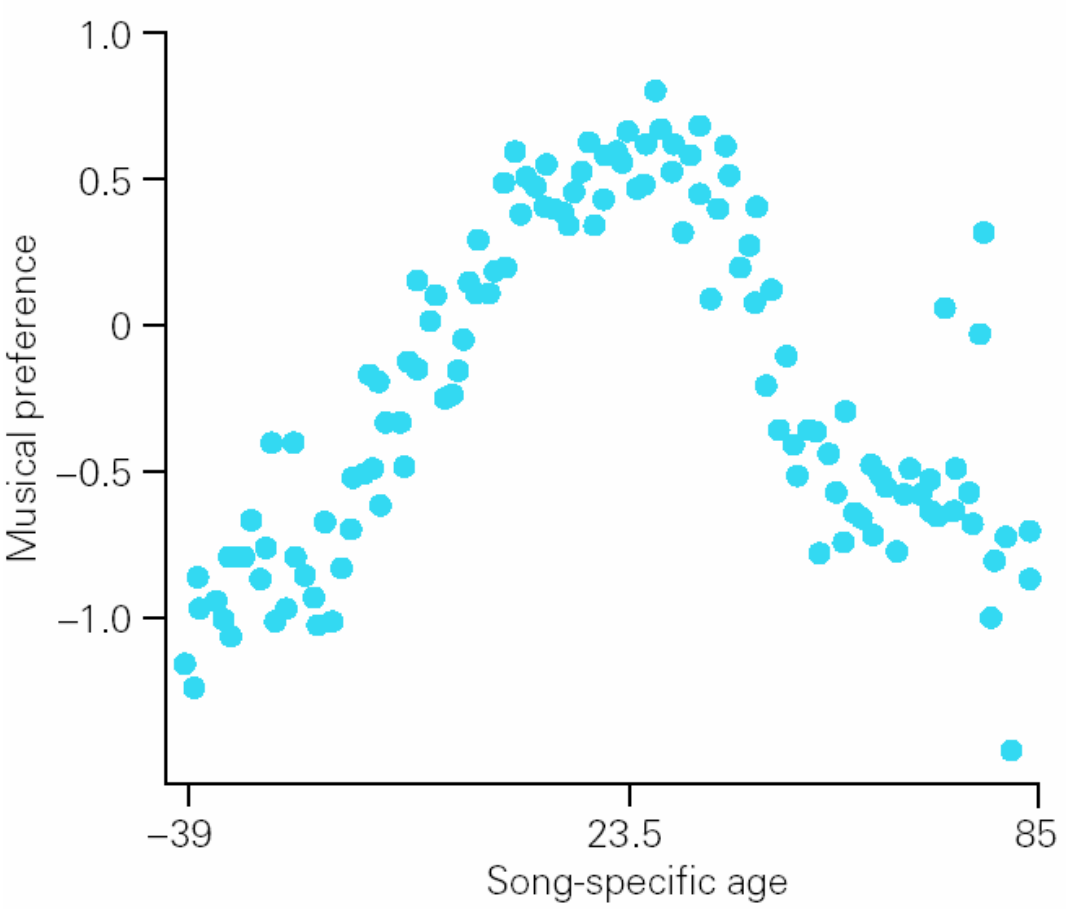


- We see a **negative** association with a **linear** pattern.
- We will use a **straight-line equation** to model this relationship.

Example 5.3 *The Development of Musical Preferences*

- The 108 participants in the study ranged in age from 16 to 86 years old.
- We want to examine the **relationship** between **song-specific age** (age in the year the song was popular) and **musical preference** (positive score => above average, negative score => below average).

Example 5.3 *The Development of Musical Preferences*



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

5.2 Describing Linear Patterns 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:

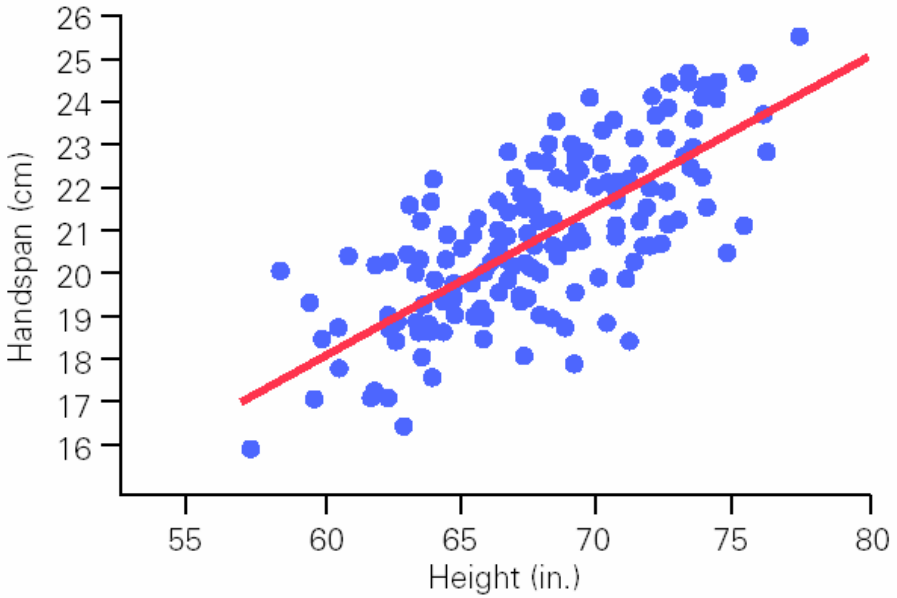
- to **estimate the average** value of y at any specified value of x
- to **predict the value** of y for an **individual**, given that individual's x value

Example 5.1 *Height and Handspan (cont)*

Regression equation: Handspan = -3 + 0.35 Height

Estimate the average handspan for people 60 inches tall:
Average handspan = -3 + 0.35(60) = 18 cm.

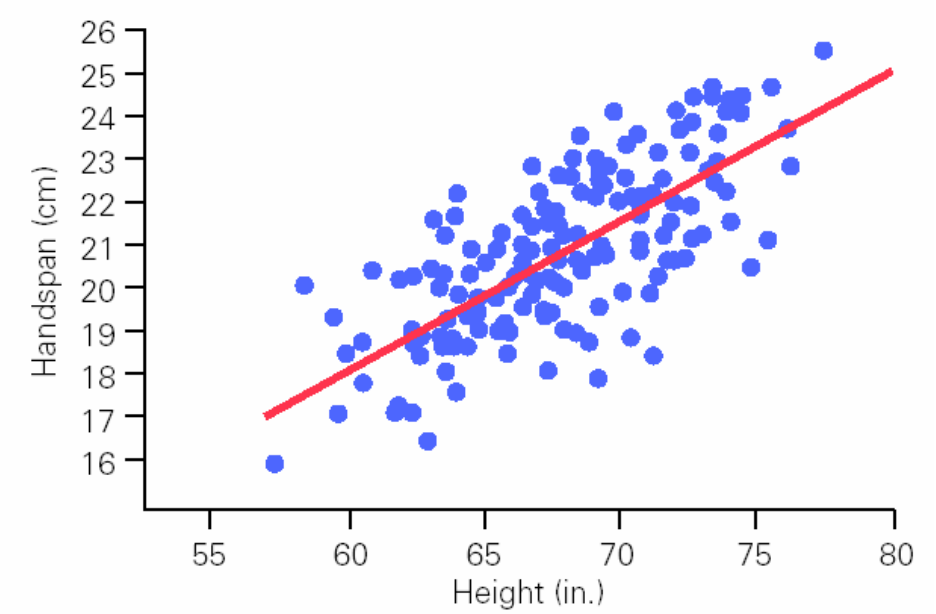
Predict the handspan for someone who is 60 inches tall:
Predicted handspan = -3 + 0.35(60) = 18 cm.



Example 5.1 *Height and Handspan (cont)*

Regression equation: Handspan = -3 + 0.35 Height

Slope = 0.35 => Handspan increases by 0.35 cm,
on average, for each increase of 1 inch in height.



In a **statistical relationship**, there is variation from the average pattern.

The Equation for the Regression Line

$$\hat{y} = b_0 + b_1 x$$

$\hat{y}$ is spoken as "***y*-hat**," and it is also referred to either as predicted y or estimated y.

b_0 is the **intercept** of the straight line. The intercept is the value of y when $x = 0$.

b_1 is the **slope** of the straight line. The slope tells us how much of an increase (or decrease) there is for the y variable when the x variable increases by one unit. The sign of the slope tells us whether y increases or decreases when x increases.

Prediction Errors and Residuals

- **Prediction Error** = difference between the observed value of y and the predicted value $\hat{y}$.

- **Residual** $= (y - \hat{y})$

Let's predict your handspan
Record these on your purple sheet

Regression equation: $\hat{y} = b_0 + b_1 x$

Handspan (cm) = -3 + 0.35 Height (inches)

Calculate your predicted handspan:

Examples: -3 + (.35)(60 inches) = 18 cm

-3 + (.35)(65 inches) = 19.75 cm

-3 + (.35)(70 inches) = 21.5 cm

Find your residual:

(actual handspan – predicted handspan)

5.3 Measuring Strength and Direction with Correlation

Correlation *r* *indicates the* ***strength*** *and the* ***direction*** *of a straight-line relationship.*

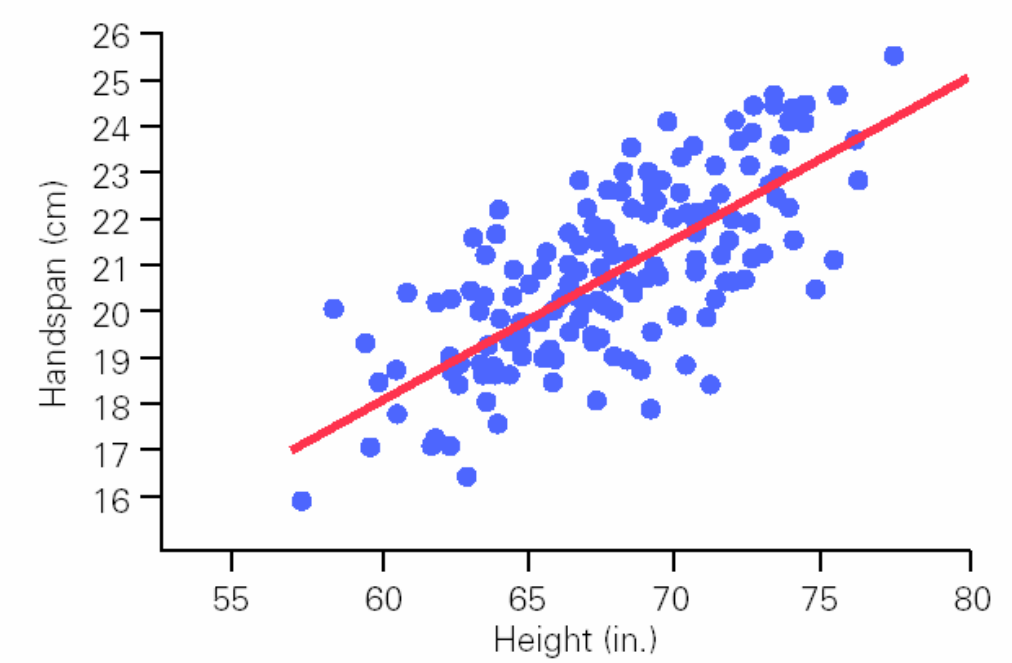
- The ***strength*** of the relationship is determined by the *closeness of the points to a straight line.*
- The ***direction*** is determined by whether one variable generally increases or generally decreases when the other variable increases.

Example 5.1 *Height and Handspan (cont)*

Regression equation: Handspan = -3 + 0.35 Height

Correlation r = +0.74 =>

a somewhat **strong positive linear** relationship.

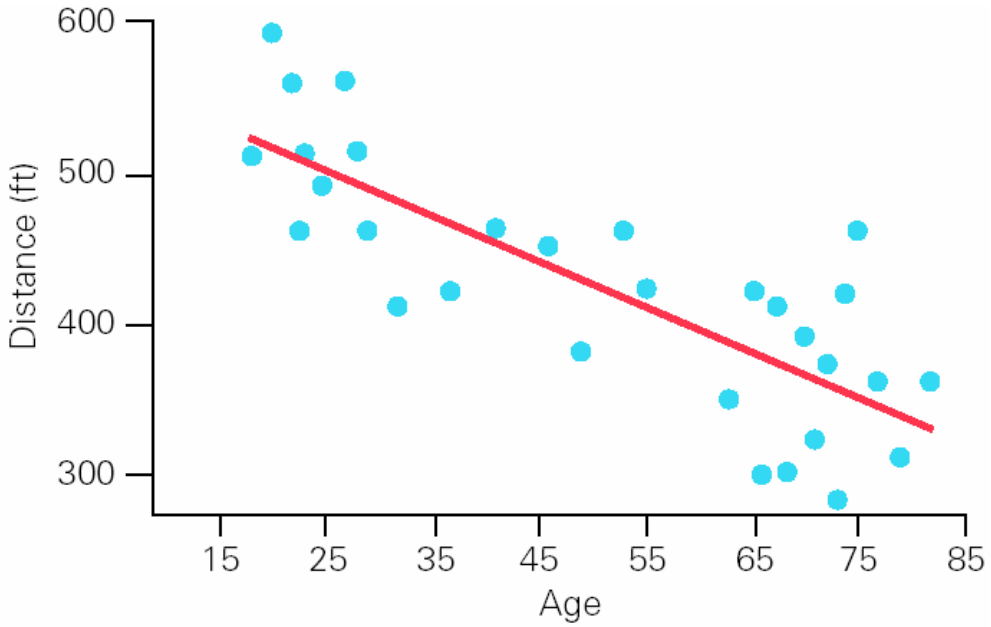


Example 5.2 *Driver Age and Maximum Legibility Distance of Highway Signs (cont)*

Regression equation: Distance = 577 - 3 Age

Correlation *r* = -0.8 =>

a **somewhat strong negative linear** association.

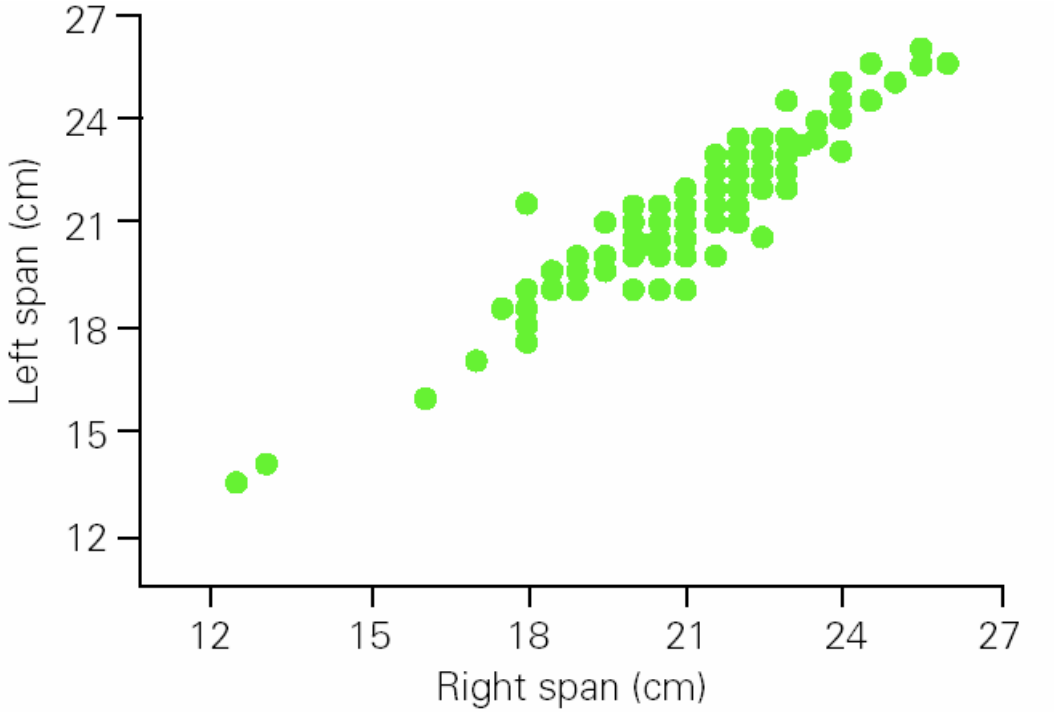


Example 5.6 *Left and Right Handspans*

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.

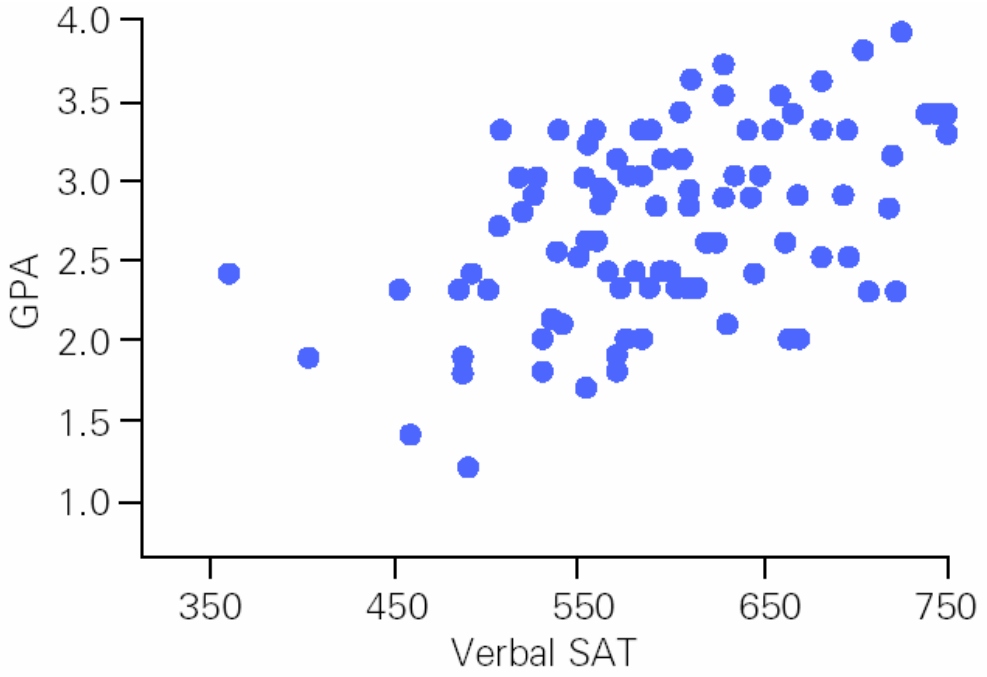


Example 5.7 *Verbal SAT and GPA*

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.

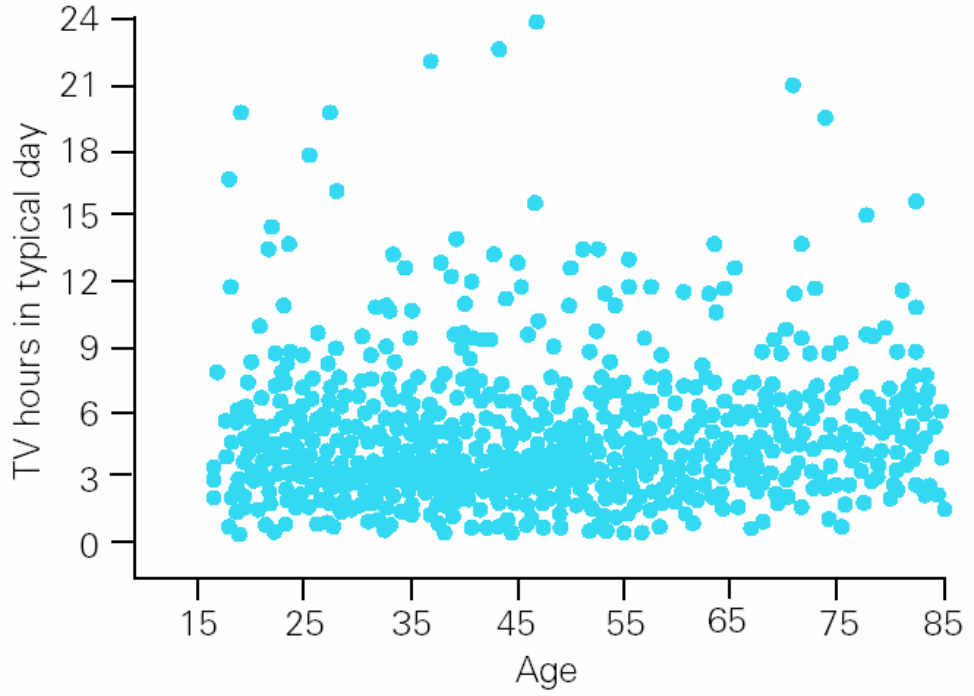


Example 5.8 *Age and Hours of TV Viewing*

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

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

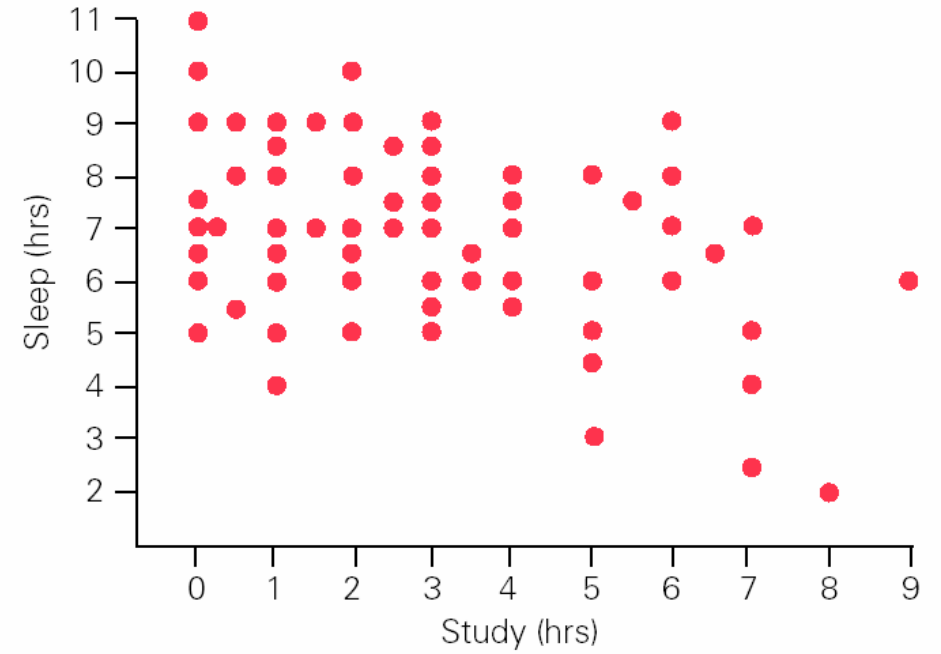
Note: a few claimed to watch more than 20 hours/day!



Example 5.9 *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.



Summary

Regression is used to do two things:

- Predict future values using information available now
- Estimate the average relationship between a "response" and one or more other variables
- Regression only works for *linear* relationships