

**Today: Sections 5.3 and 5.4**  
(We will do 5.5 Wed of next week)

**Homework:**  
Chapter 5: #40  
(use R to draw plot, or do by hand)

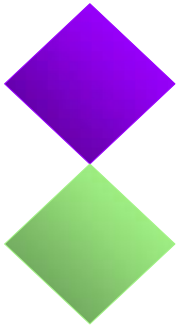
# Three tools for studying relationships between two quantitative variables:



- **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:

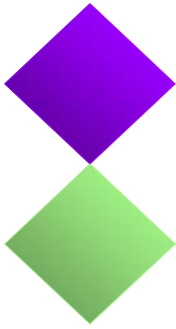
| Height (in.) | Span (cm) |
|--------------|-----------|
| 71           | 23.5      |
| 69           | 22.0      |
| 66           | 18.5      |
| 64           | 20.5      |
| 71           | 21.0      |
| 72           | 24.0      |
| 67           | 19.5      |
| 65           | 20.5      |
| 76      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

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.



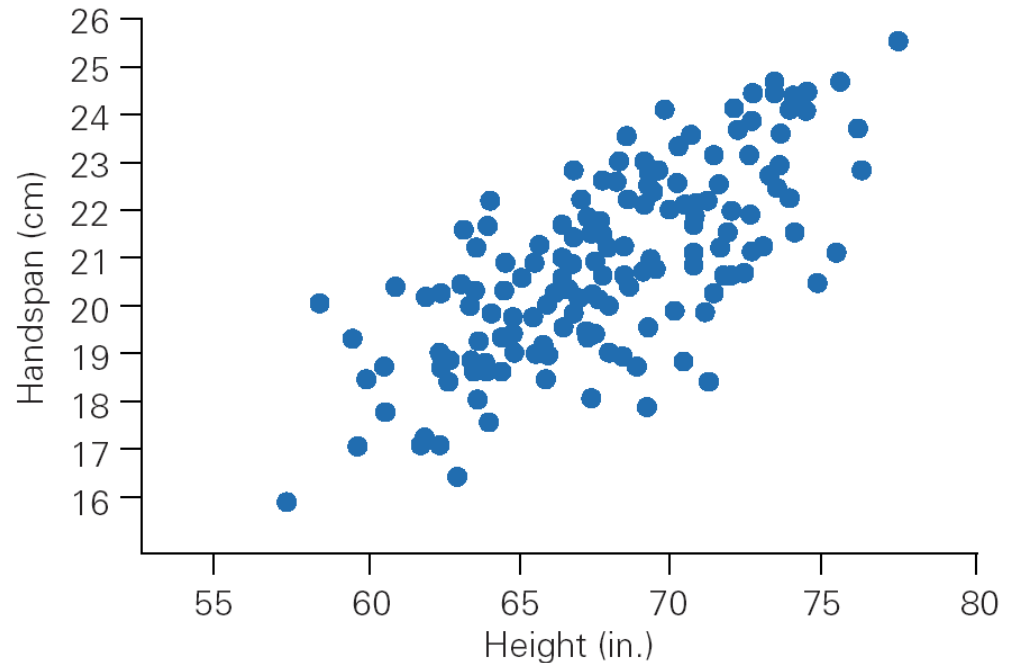
# Positive Association: *Height and Handspan*



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**.



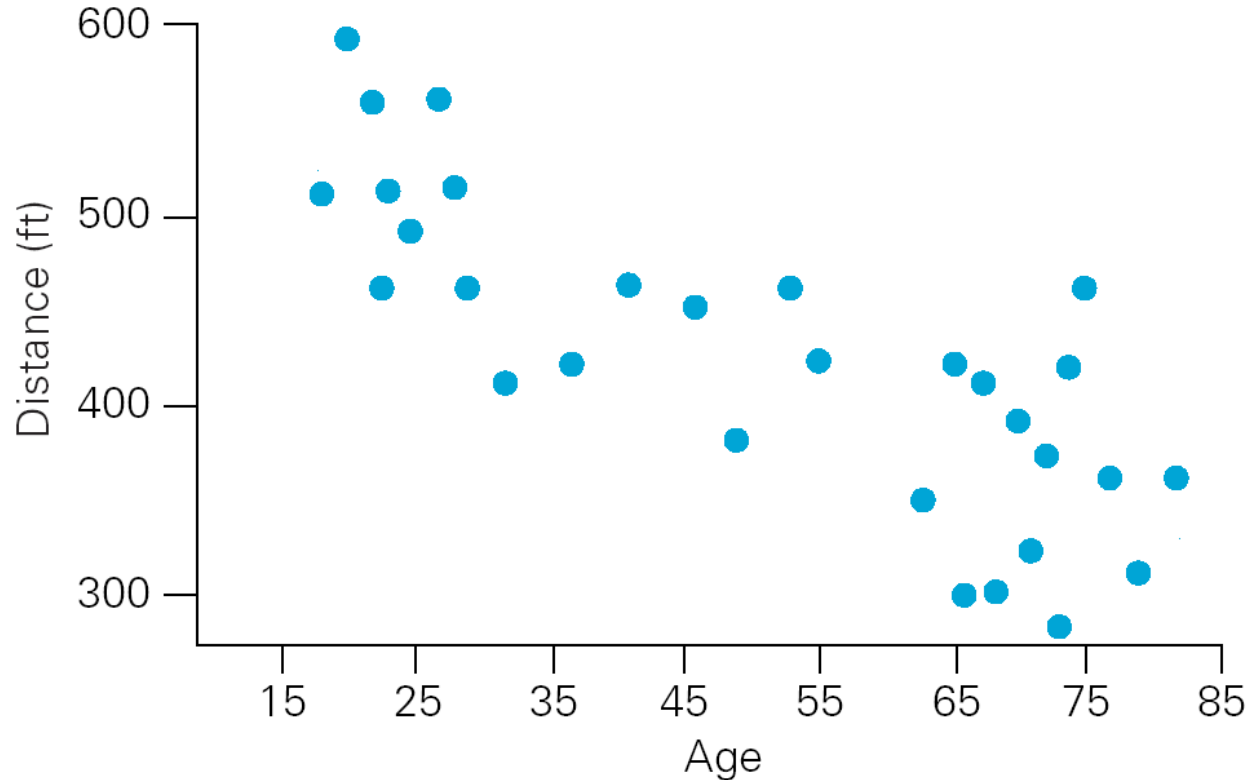
# Negative Association:

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



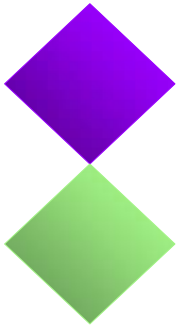
- A research firm determined the **maximum distance** at which each of 30 drivers could read a newly designed sign.
- The 30 participants in the study ranged in **age** from 18 to 82 years old.
- We want to examine the **relationship** between age and the sign legibility distance.

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



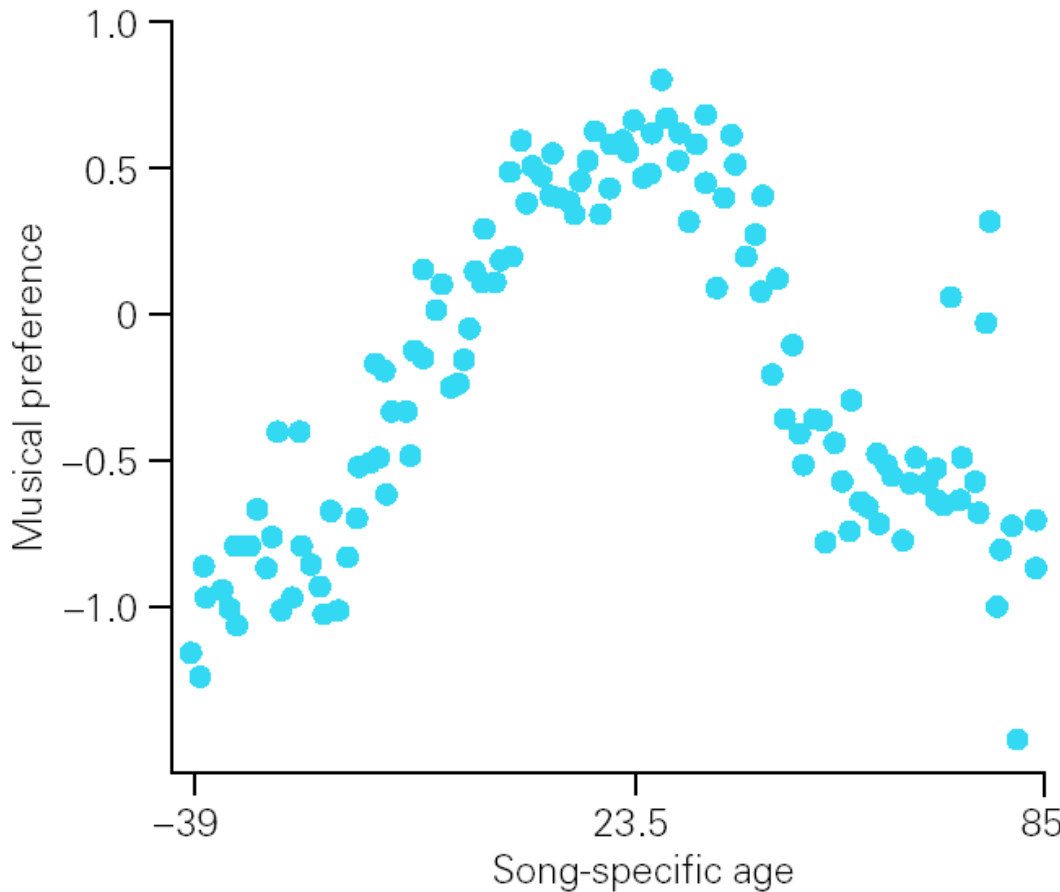
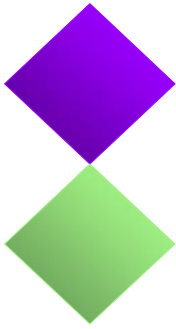
- We see a **negative** association with a **linear** pattern.
- We will use a **straight-line equation** to model this relationship.

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



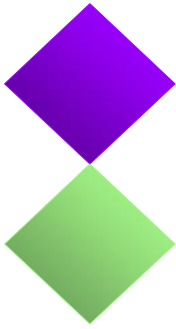
- The 108 participants in the study ranged in age from 16 to 86 years old.
- We want to examine the **relationship** between **song-specific age** (age in the year the song was popular) and **musical preference** (positive score => above average, negative score => below average).

# Example 5.3 *The Development of Musical Preferences*



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

# Review of what we do with a regression line

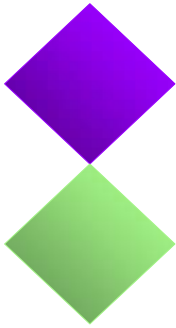


When the best equation for describing the relationship between  $x$  and  $y$  is a straight line, the equation is called the **regression line**.

## **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

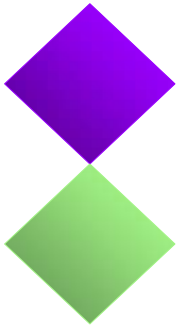
## 5.3 Measuring Strength and Direction with Correlation



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

- The **strength** of the 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 $r$



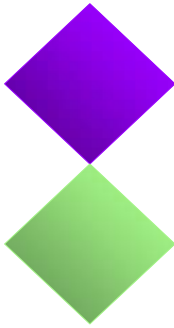
- $r$  is always between  $-1$  and  $+1$
- $r = -1$  or  $+1$  indicates a perfect linear relationship
  - $r = +1$  means *all* points are on a line with *positive* slope
  - $r = -1$  means *all* points are on a line with *negative* slope
- **Magnitude** of  $r$  indicates the strength of the *linear* 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 $r$

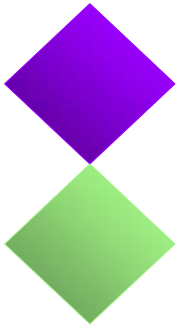
$$r = \frac{1}{n-1} \sum \left( \frac{x_i - \bar{x}}{s_x} \right) \left( \frac{y_i - \bar{y}}{s_y} \right)$$

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.



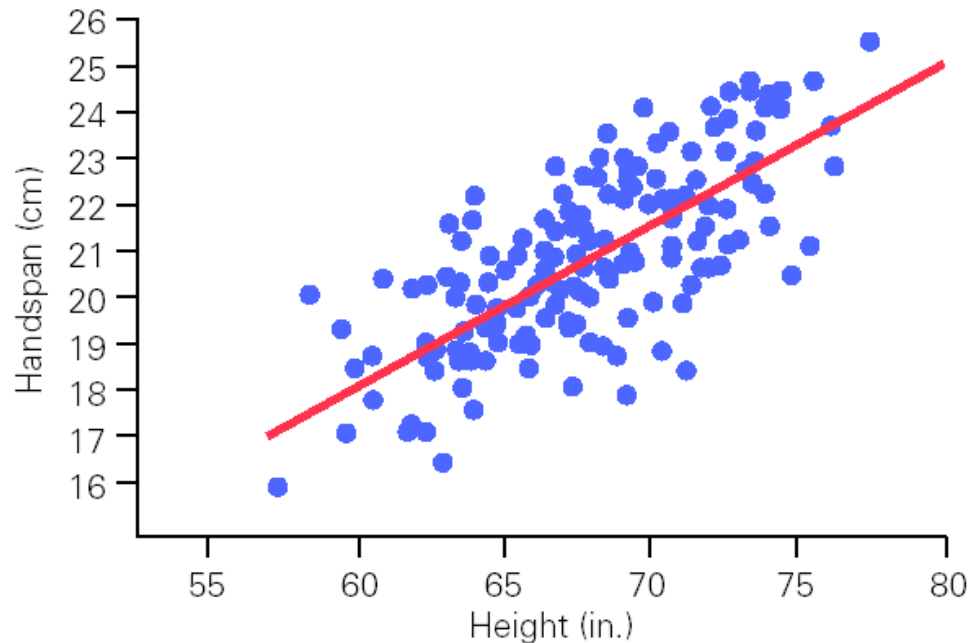
# Example 5.1 *Height and Handspan*



**Regression equation:**  $\text{Handspan} = -3 + 0.35 \text{ Height}$

**Correlation  $r = +0.74 \Rightarrow$**

a somewhat **strong positive linear** relationship.

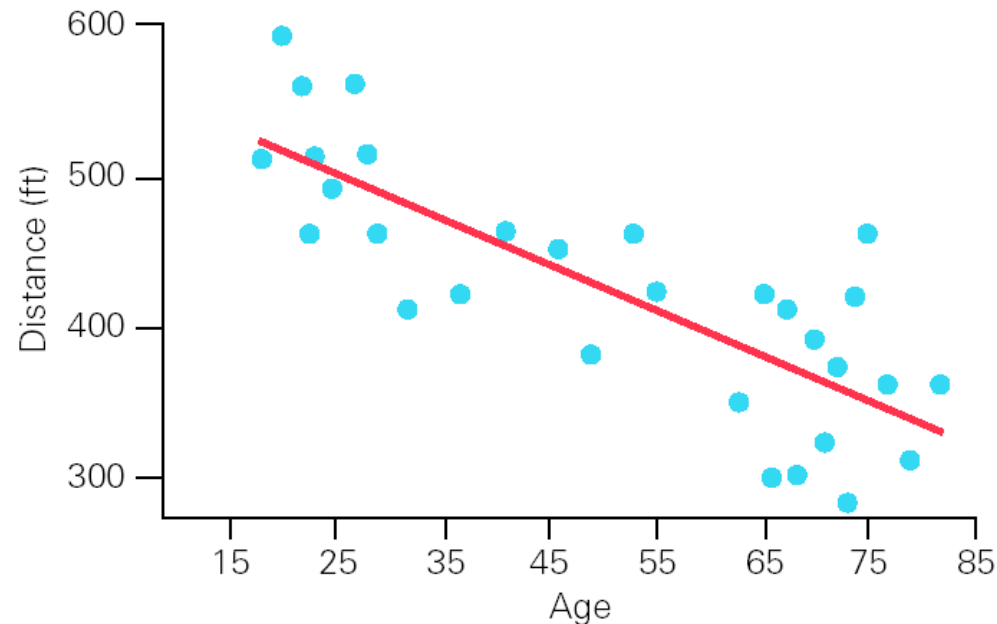


## Example 5.2 *Driver Age and Maximum Legibility Distance of Highway Signs (cont)*

Regression equation:  $\text{Distance} = 577 - 3 \text{ Age}$

Correlation  $r = -0.8 \Rightarrow$

a somewhat **strong negative linear** association.

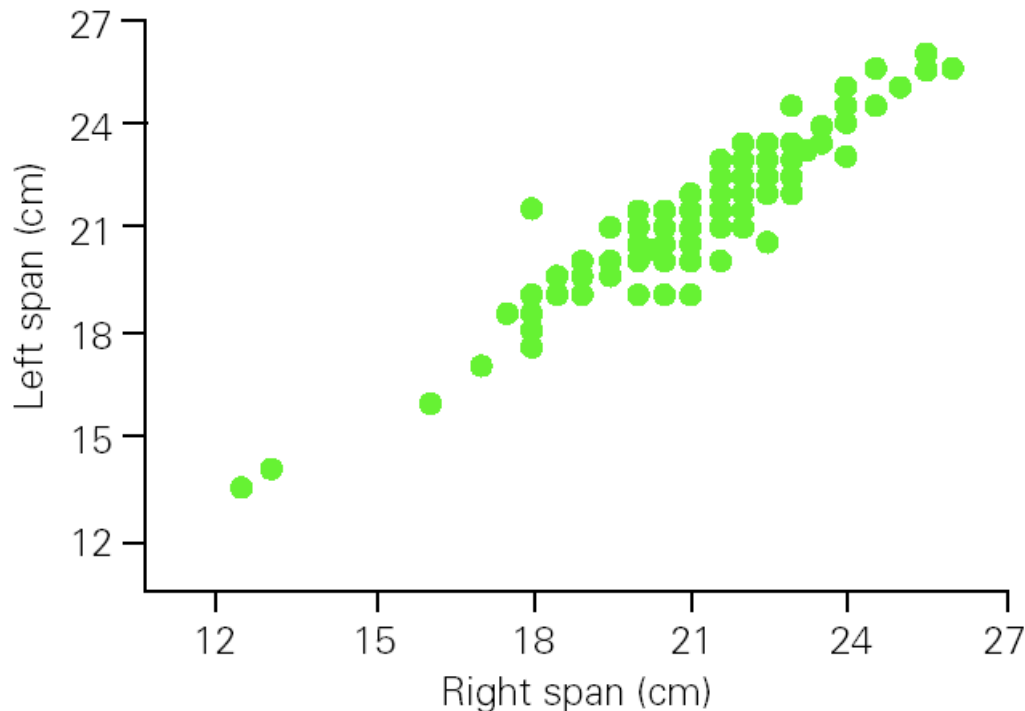


# Example 5.11 *Left and Right Handspans*

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

**Correlation  $r = +0.95 \Rightarrow$**

**a very strong positive linear relationship.**

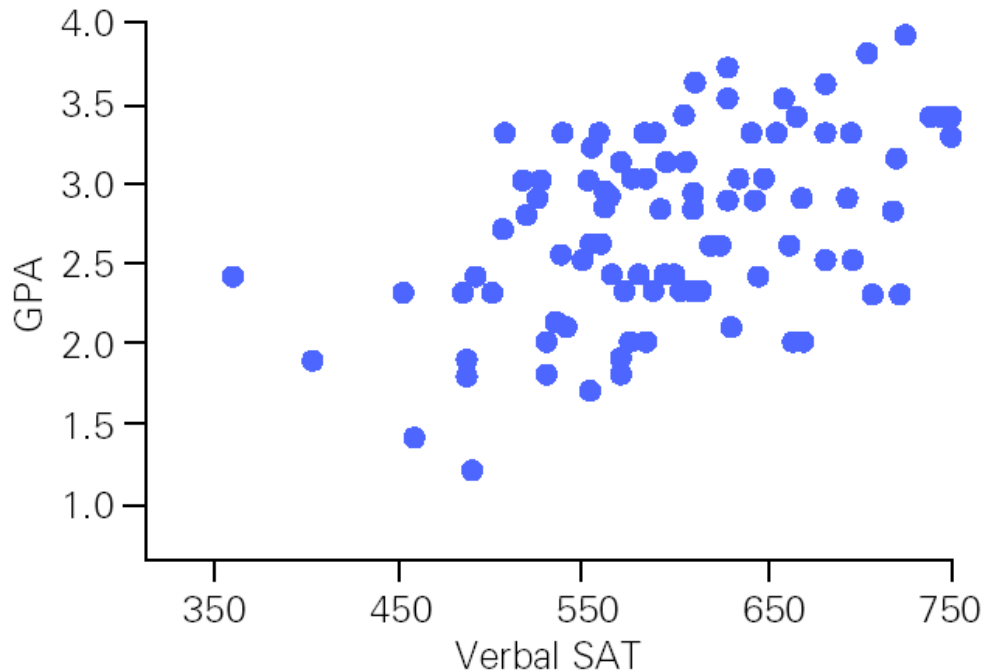


# Example 5.12 *Verbal SAT and GPA*

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

**Correlation  $r = 0.485 \Rightarrow$**

**a moderately strong positive linear relationship.**

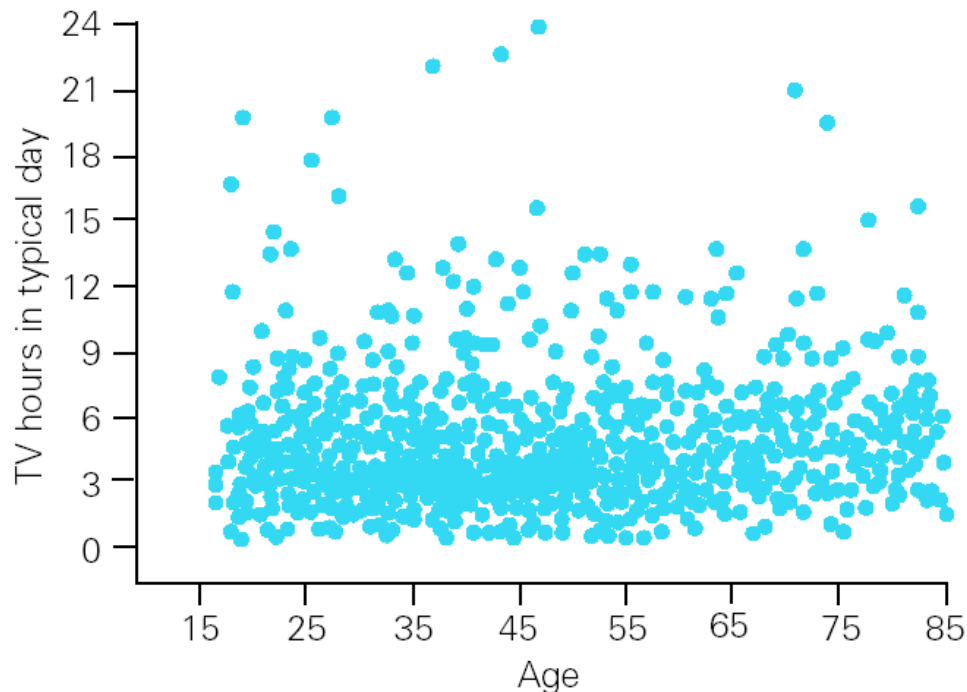


## Example 5.13 *Age and Hours of TV Viewing*

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

**Correlation  $r = 0.12 \Rightarrow$  a weak connection.**

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



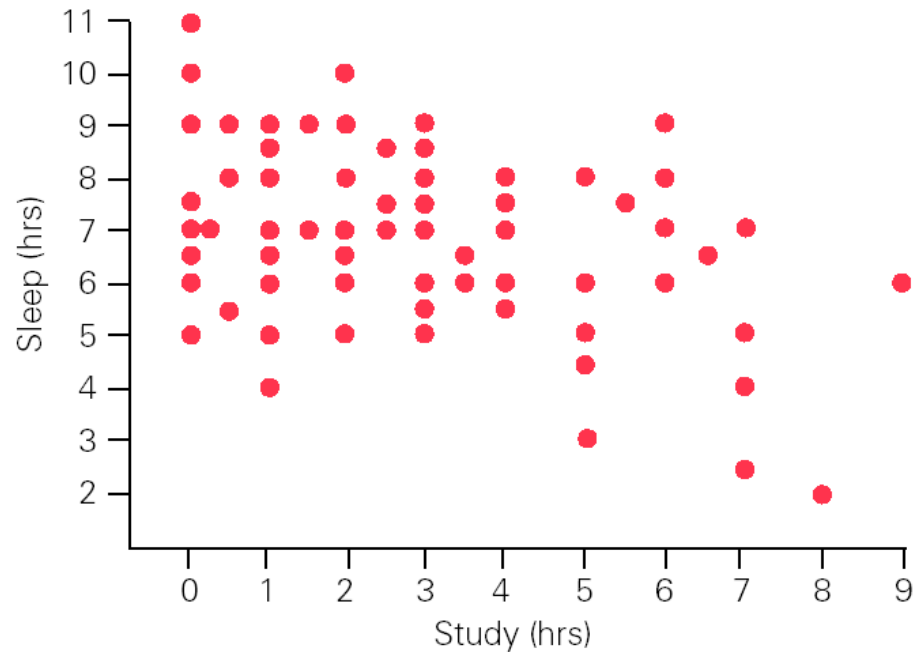
# Example 5.14 *Hours of Sleep and Hours of Study*



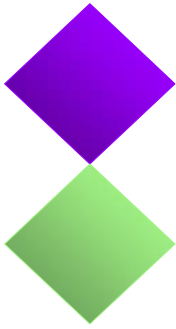
Relationship between reported hours of sleep the previous 24 hours and the reported hours of study during the same period for a sample of 116 college students.

**Correlation  $r = -0.36$**

**$\Rightarrow$  a not too strong  
negative association.**



# Recall: The Equation for the Regression Line:

$$\hat{y} = b_0 + b_1x$$


**Prediction Error** = difference between the observed value of  $y$  and the predicted value

- **Residual** =  $y - \hat{y}$
- **Least Squares Regression Line:**  
minimizes **SSE** = the sum of the squared residuals.

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



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

| <b><math>x = \text{Age}</math></b> | <b><math>y = \text{Distance}</math></b> | <b><math>\hat{y} = 577 - 3x</math></b> | <b>Residual</b>   |
|------------------------------------|---|--|-------------------|
| 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  $\Rightarrow$  observed value higher than predicted.**

**Negative residual  $\Rightarrow$  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.

# Interpretation of $r^2$ and a formula



**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.

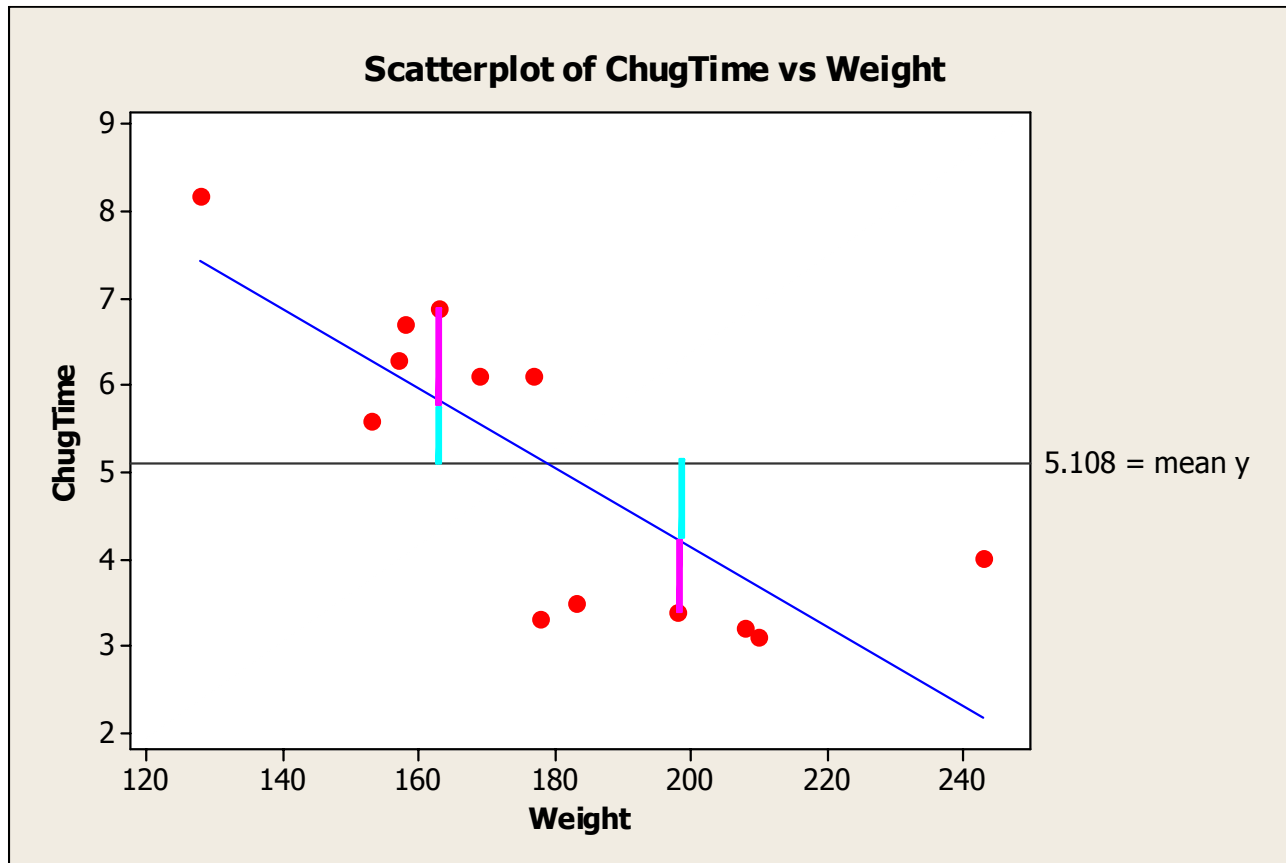
$$r^2 = \frac{SSTO - SSE}{SSTO}$$

**Total variation** for each point =  $(y - \text{mean } y)$

**Unexplained part** = residual =  $(\text{actual } y - \text{predicted } y)$

**Explained by knowing  $x$**  =  $(\text{predicted } y - \text{mean } y)$

Data from Exercise 5.73

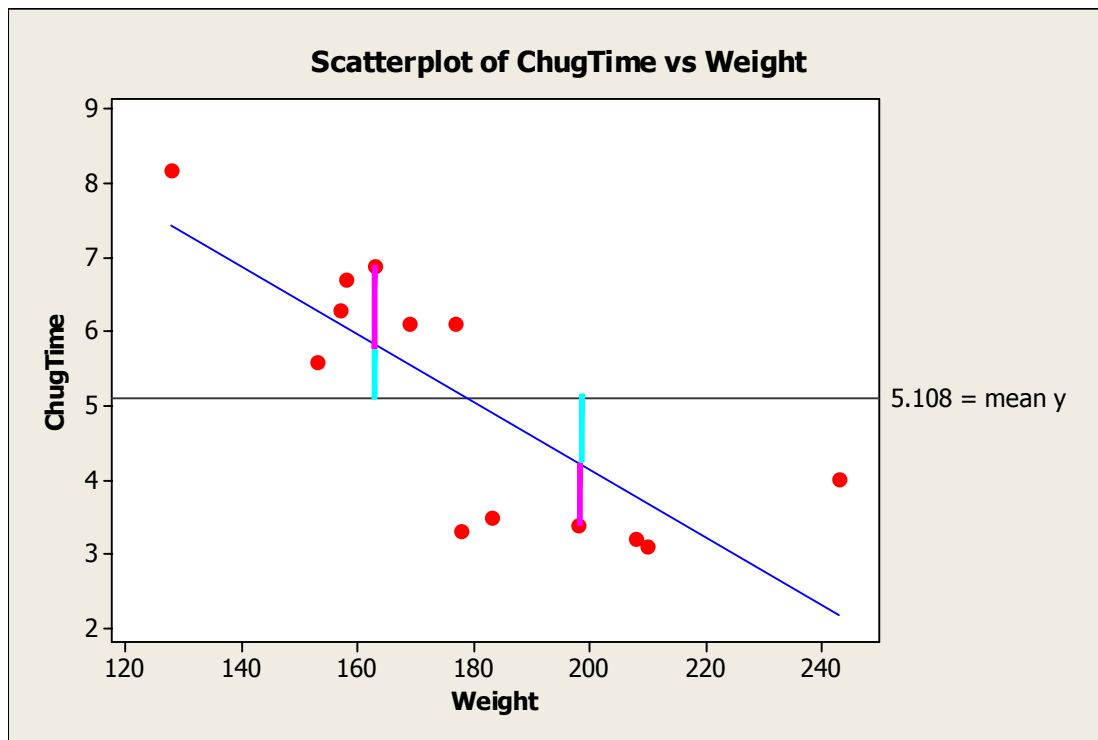


**Total variation** summed over all points =  $SSTO = 36.6$

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

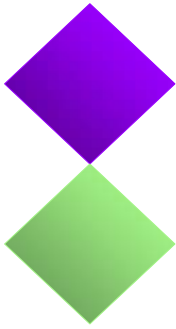
**Explained by knowing  $x$**  summed over all points =  $22.7$

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



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

# Interpretation of $r^2$ for other examples



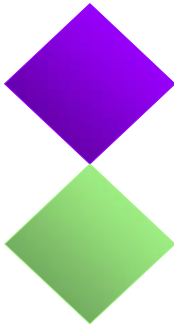
## **Example 5.11: *Left and Right Handspans***

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

## **Example 5.13: *TV viewing and Age***

$r^2 = 0.014 \Rightarrow$  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*



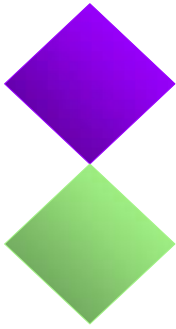
- Coefficients:

- ```
                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

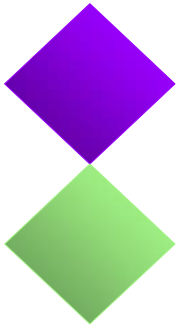
- Multiple R-squared: 0.9023, Adjusted R-squared: 0.9018

## 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.

# Problem 5.6: 20 cities in US

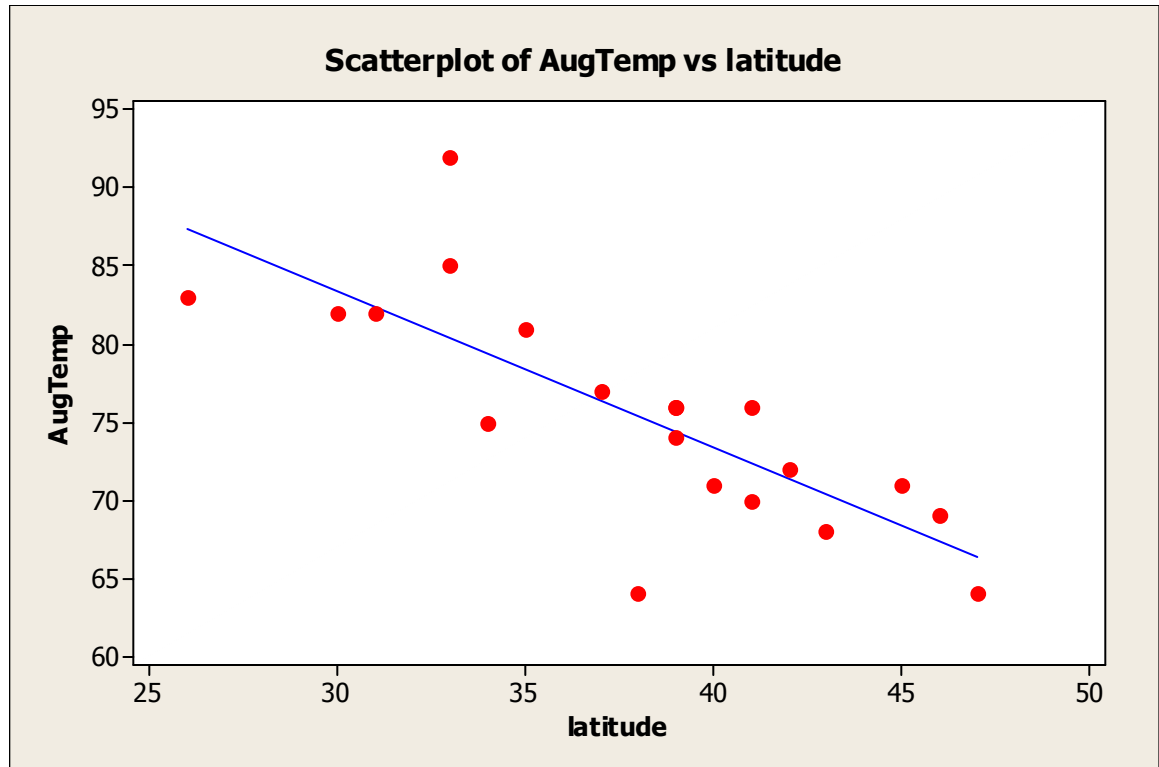
$x$ =latitude,  $y$ =average Aug temp

Intercept = 114

Slope =  $-1.00$

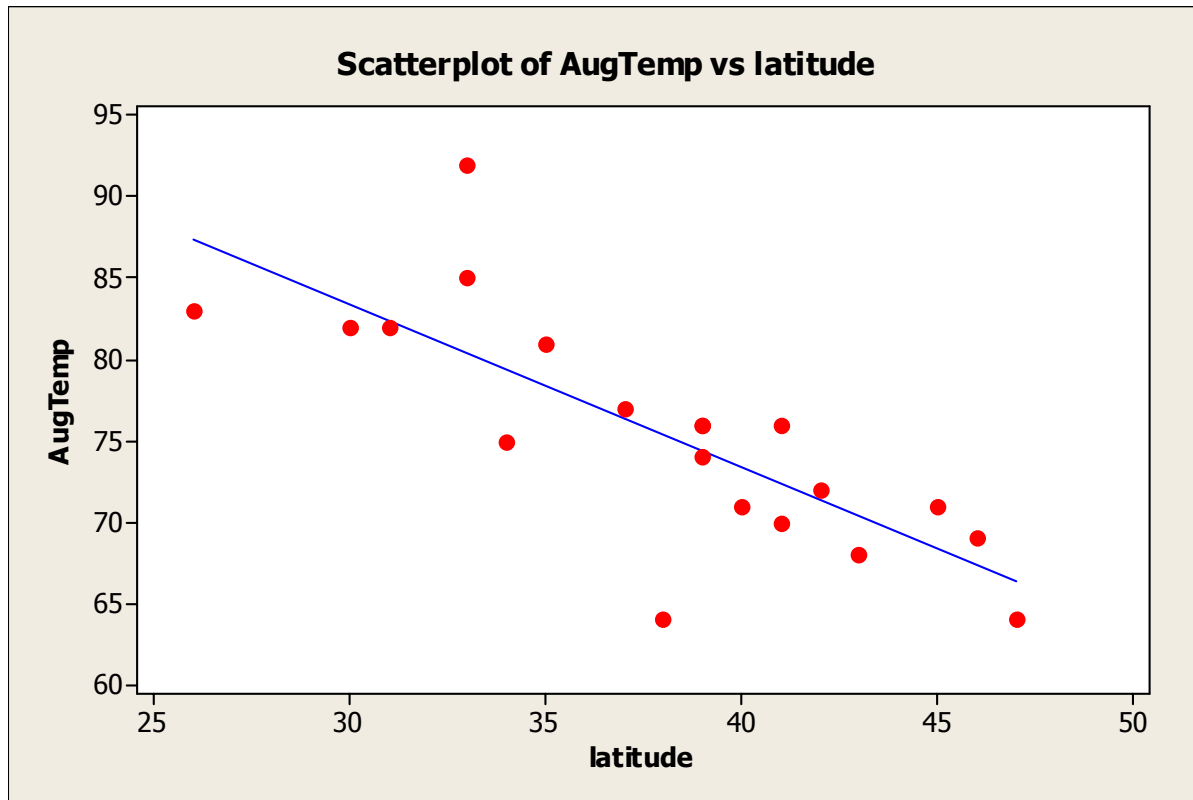
For instance,

Irvine latitude  
= 33.4, so  
predicted avg  
Aug temp is  
 $114 - 33.4 =$   
80.6 degrees



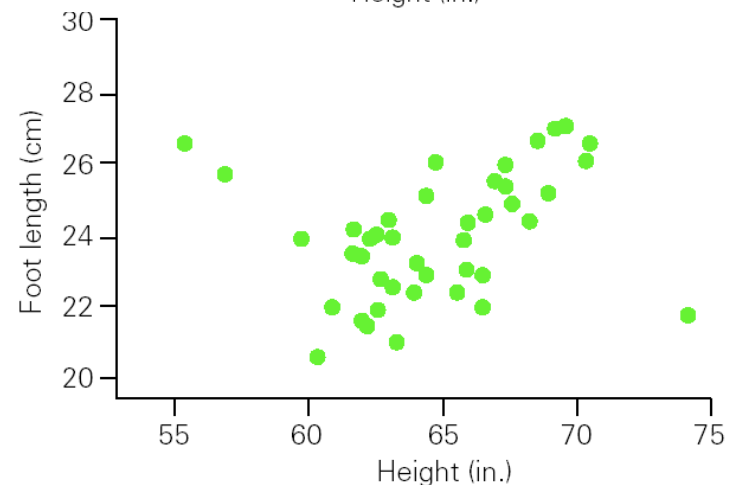
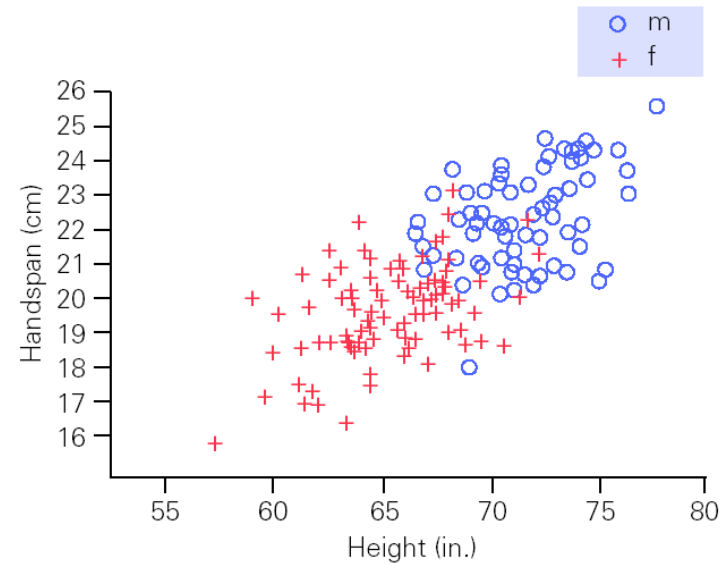
# Extrapolation

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 temps; intercept = 126.

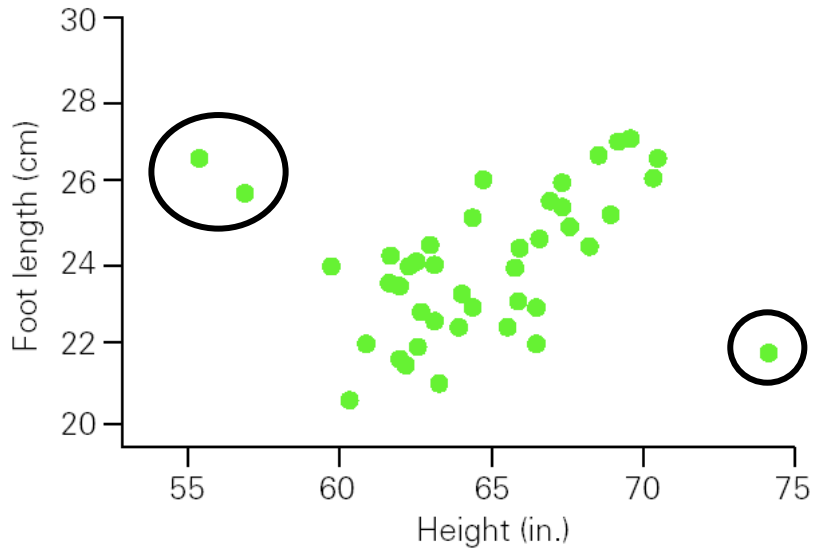


# Groups and Outliers

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



# Example 5.4 *Height and Foot Length Outliers*



Three outliers were data entry errors.

## Regression equation

uncorrected data:  $15.4 + 0.13 \text{ height}$

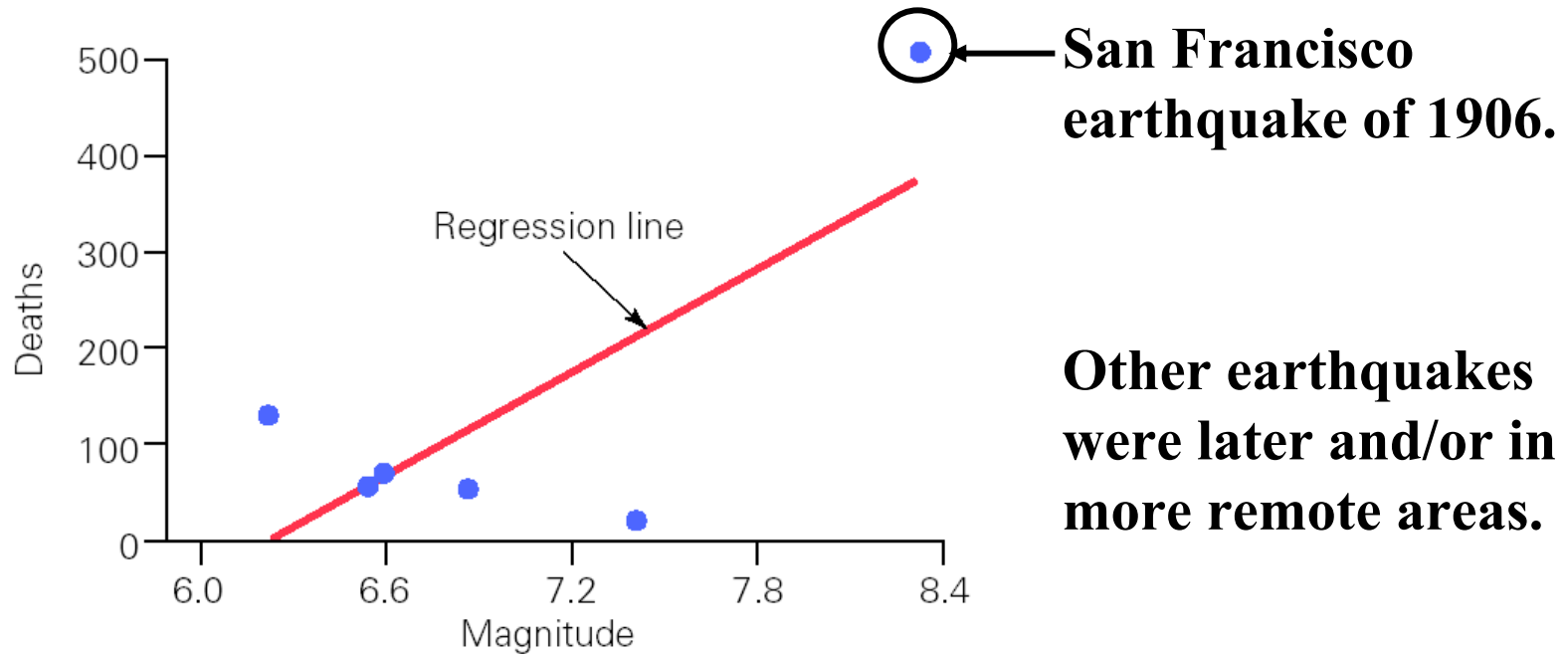
corrected data:  $-3.2 + 0.42 \text{ height}$

## Correlation

uncorrected data:  $r = 0.28$

corrected data:  $r = 0.69$

# Example 5.16 Earthquakes in US – an outlier

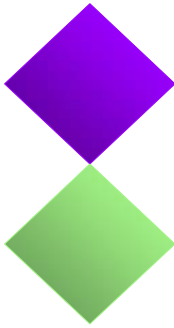


## Correlation

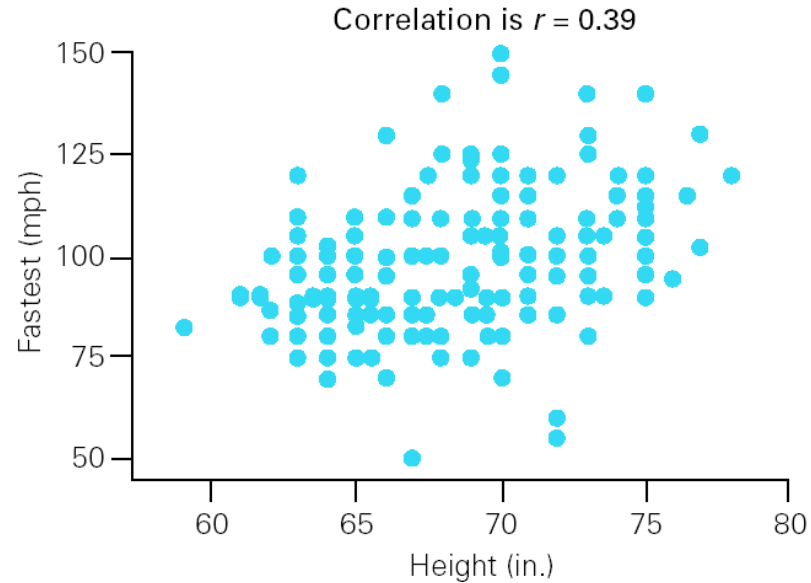
all data:  $r = 0.73$

w/o SF:  $r = -0.96$

# Example 5.17 *Height and Lead Feet*

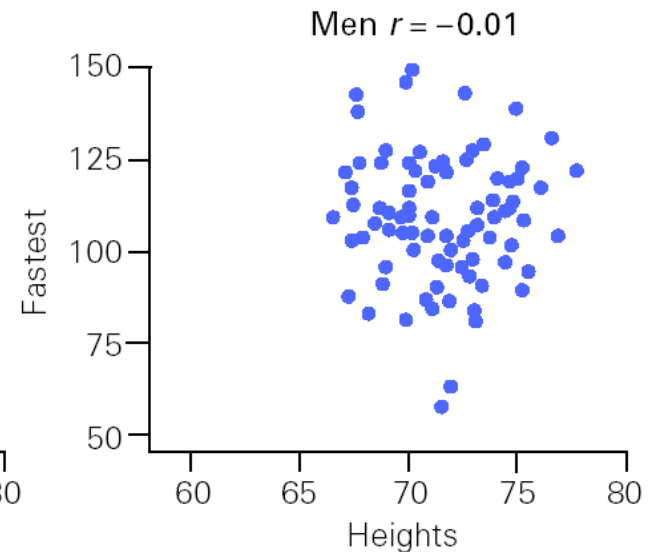
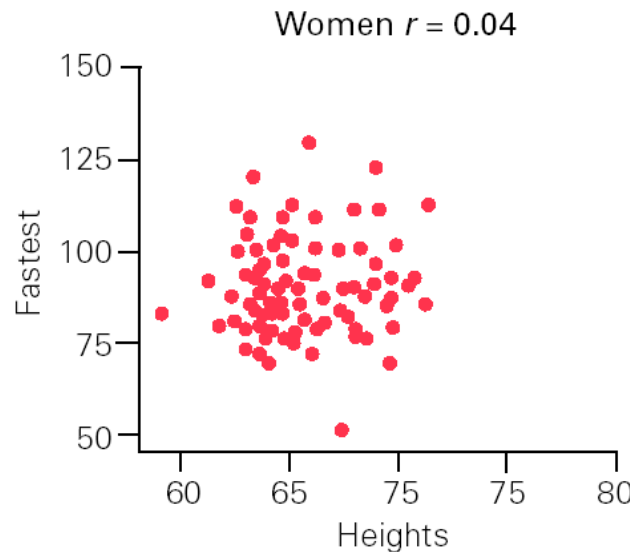


**Scatterplot of all data:**  
College student heights and responses to the question “What is the fastest you have ever driven a car?”  $r = .39$



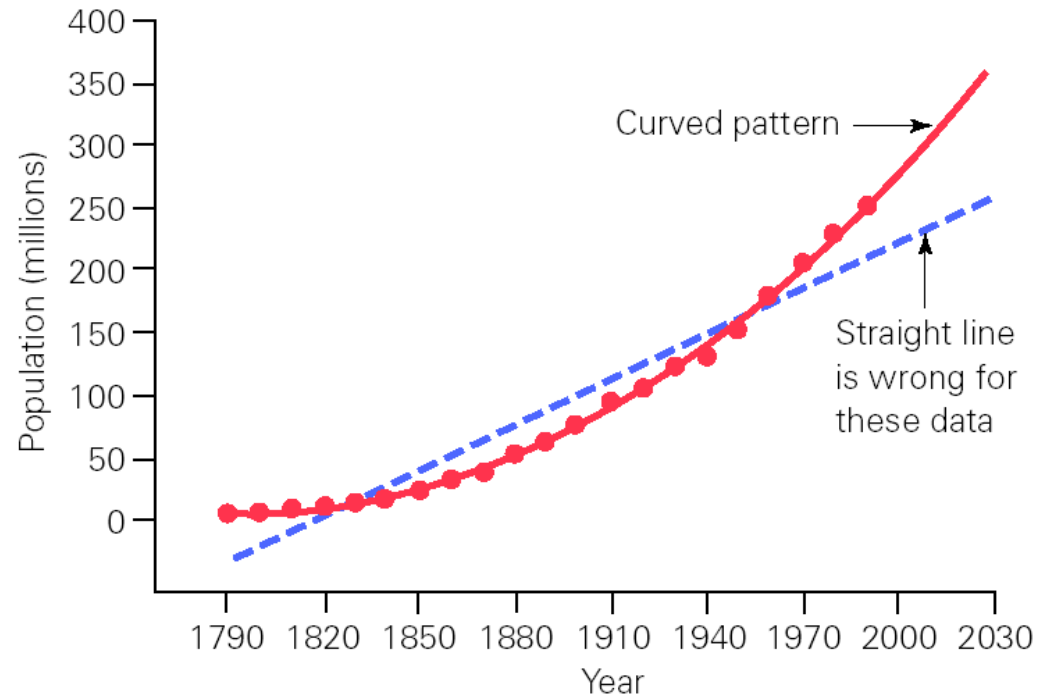
**Scatterplot by gender:**  
Combining two groups led to misleading correlation

$r = .04; -.01$



# 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:**  $\text{population} = -2218 + 1.218(\text{Year})$

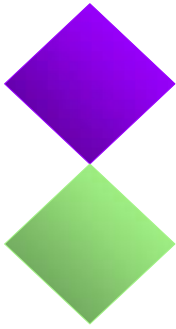
**Poor Prediction for Year 2030**  $= -2218 + 1.218(2030)$   
or about 255 million, only 6 million more than 1990.

# Applets to illustrate concepts

[http://onlinestatbook.com/stat\\_sim/reg\\_by\\_eye/index.html](http://onlinestatbook.com/stat_sim/reg_by_eye/index.html)

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

# 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