STAT011 Statistical Methods I

Lecture 5 Correlation and Regression

Lu Chen
Swarthmore College
2/5/2019

Review

- Density curve
  - Properties: area under the curve = 1; area = proportion;
  - Normal curve: symmetric, unimodal, bell-shaped
- Normal distribution: $N(\mu, \sigma)$
  - Density function
  - The 68-95-99.7 rule
  - Standard Normal distribution: $N(0, 1)$
    - `dnorm()`, `pnorm()`, `qnorm()`
  - Assessing Normality: Normal Q-Q plot
    - `qqnorm()`, `abline()`

Outline

- Relationships between variables
- Relationship between two quantitative variables
- Correlation coefficient
  - Definition and formula
  - Examples
- Least squares regression
  - How to find the best fitting line
  - Least squares regression in R

Relationships between variables

- Lecture 2~4: exploratory analysis of a single variable.
- But most statistical problems involve two or more variables.
- Lecture 5~7: exploring the relationship between two variables.

Association between variables

Two variables measured on the same observations are associated if knowing the values of one of the variables tells you something about the values of the other variable.

Relationships between variables

Examples:

- Smoking versus lung cancer
- Number of courses you planed to take versus number of courses you are taking right now
- SAT score versus first year college GPA
- Size versus price of a cup of coffee
- Height versus shoe length
- Mileage versus market value of a car

Note:

- In most relationships, one variable explains/causes the changes in the other variable.

Relationships between variables

A **response variable** measures an outcome of a study.

An **explanatory variable** explains or causes changes in the response variable.

- Usually the response variable is the variable of interest. In the following relationships, which variable is the explanatory/reponse variable?
  - Smoking versus lung cancer
  - Number of courses you planed to take versus number of courses you are taking right now
  - SAT score versus first year college GPA
  - Size versus price of a cup of coffee
  - Height versus shoe length
  - Mileage versus market value of a car

STAT 011 map

<table>
<tr><th colspan="2" rowspan="2">Exploratory Data Analysis</th><th rowspan="2"><u>No Explanatory</u></th><th colspan="2"><u>Explanatory</u></th></tr>
<tr><th>Categorical</th><th>Quantitative</th></tr>
<tr><td rowspan="2"><u>Response</u></td><td>Categorical</td><td>• Table of counts and proportions
• Bar plot
• Pie chart
(Lecture 2)</td><td>• Two-way tables
- Joint distribution
- Marginal distribution
- Conditional distribution
• Bar plot
(Lecture 6)</td><td>——</td></tr>
<tr><td>Quantitative</td><td>• Mean, SD
• Median, IQR
• Histogram, density curve
• Boxplot
(Lecture 2~4)</td><td>• Table of summary statistics
• Histogram, density curve
• Boxplot
(Lecture 7)</td><td>• Correlation
• Regression
• Scatterplot
(Lecture 5~6)</td></tr>
</table>

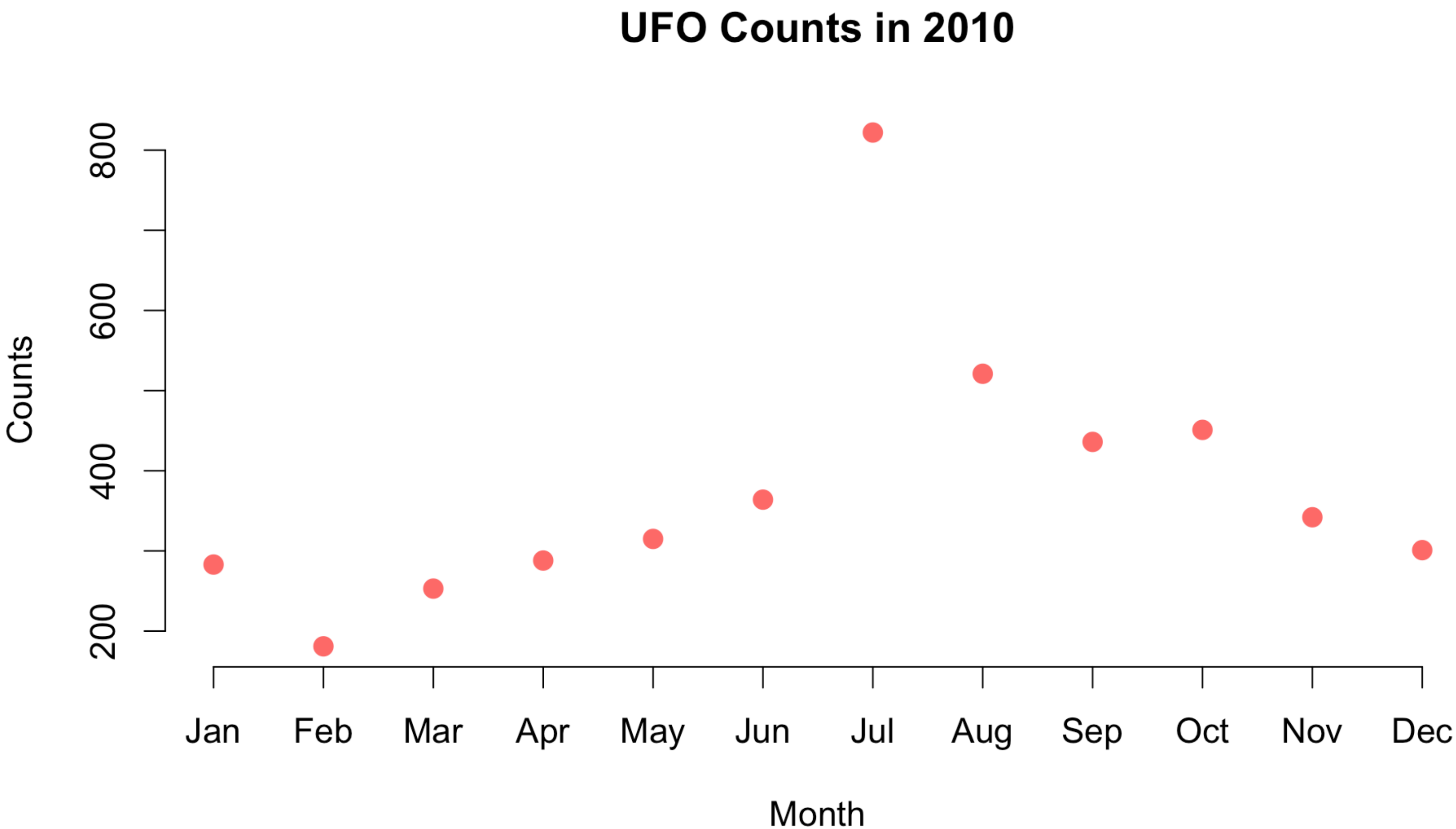
STAT 011 map

Statistical Inference		**No Explanatory**	**Explanatory**		
			Binary	**Categorical**	**Quantitative**
Response	**Binary**	Inference of a proportion *(Lecture 18)*	Inference of two proportions *(Lecture 19)*		——
	Categorical	Goodness-of-fit test *(Lecture 20)*	Chi-squared test *(Lecture 20)*		——
	Quantitative	One-sample t test *(Lecture 15)*	Two-sample t test *(Lecture 16~17)*	——	Linear regression *(Lecture 22~25)*

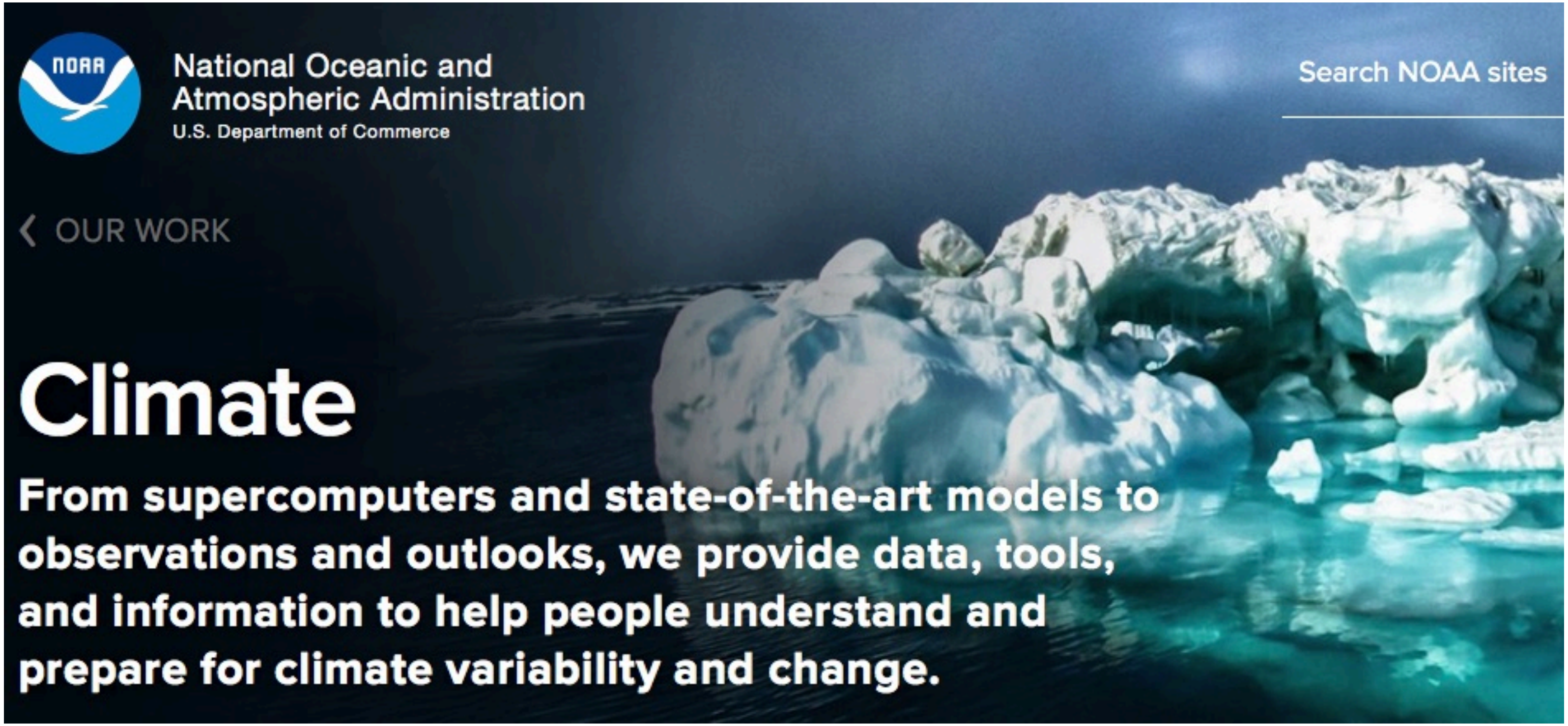
Data source

http://www.nuforc.org

UFO counts - response variable

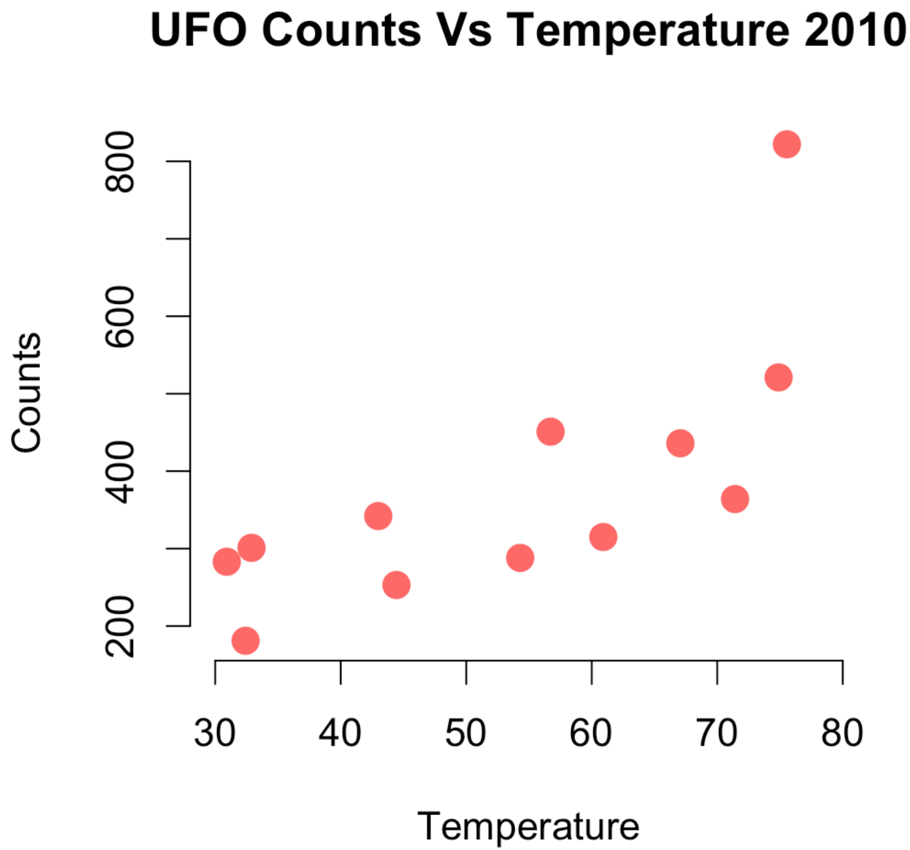


Data source



http://www.noaa.gov/climate

UFO counts versus temperature



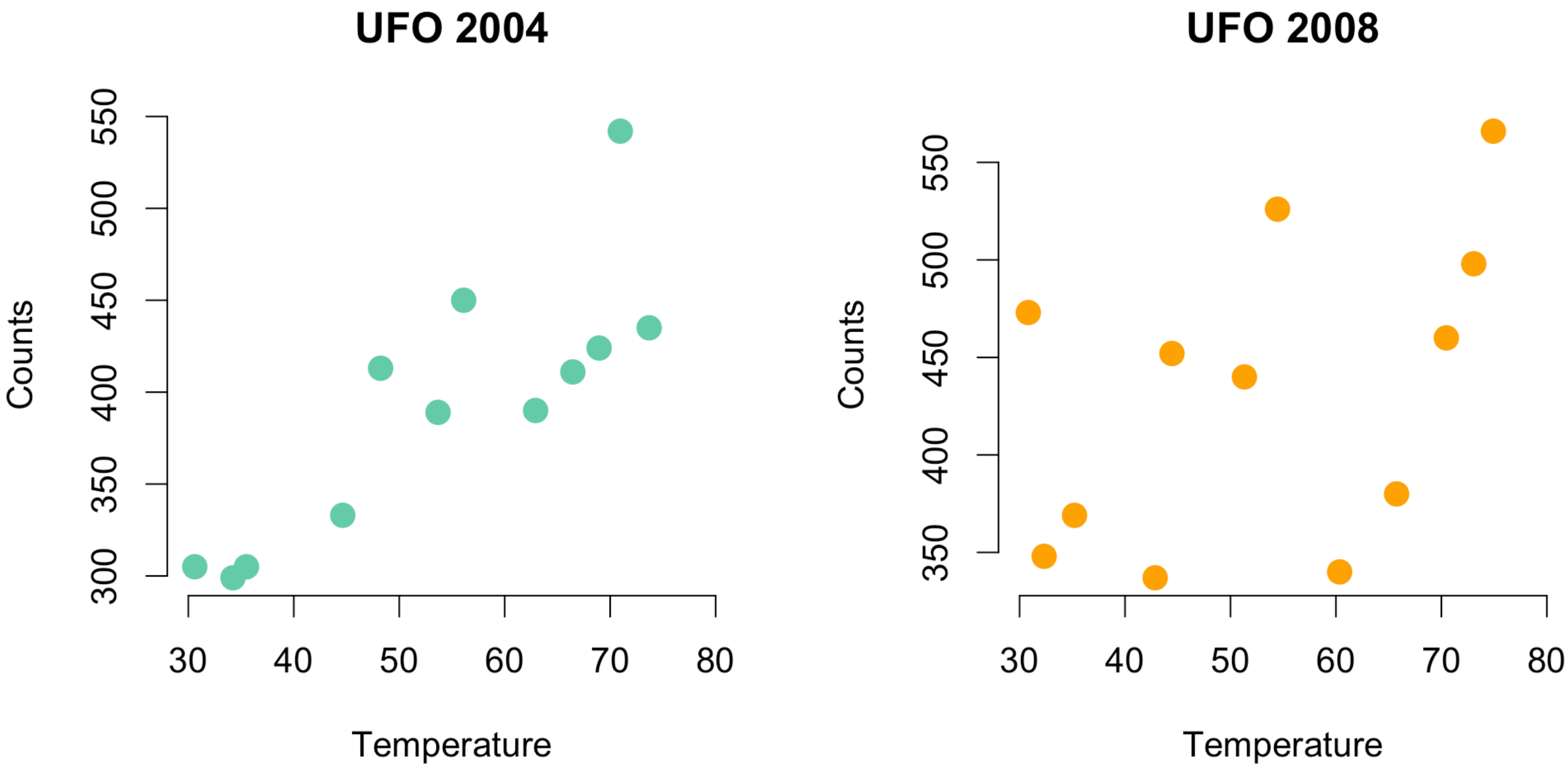
Scatterplot

- x-axis: explanatory variable - Temperature.
- y-axis: repsonse variable - UFO Counts.
- Each point represents an observation with a certain x and y value.

Describe the relationship

- Form: linear or curved or none?
- Direction: positive or negative?
- Strength: strong or weak?
- Any outlier?

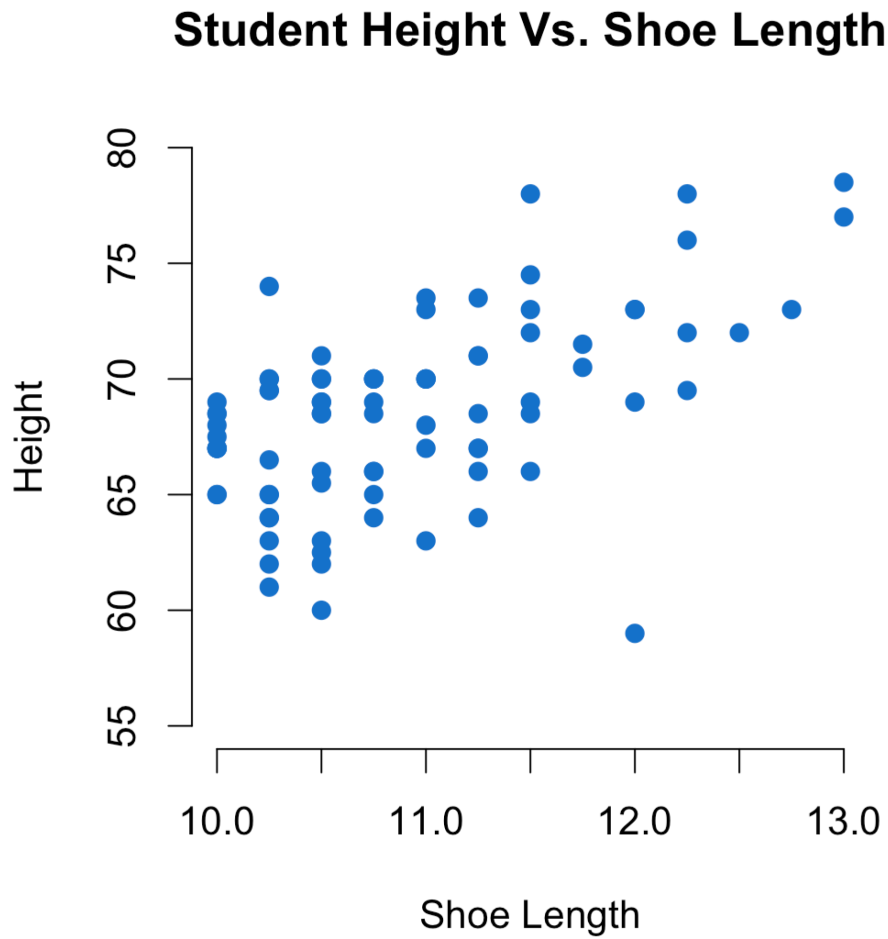
UFO counts versus temperature



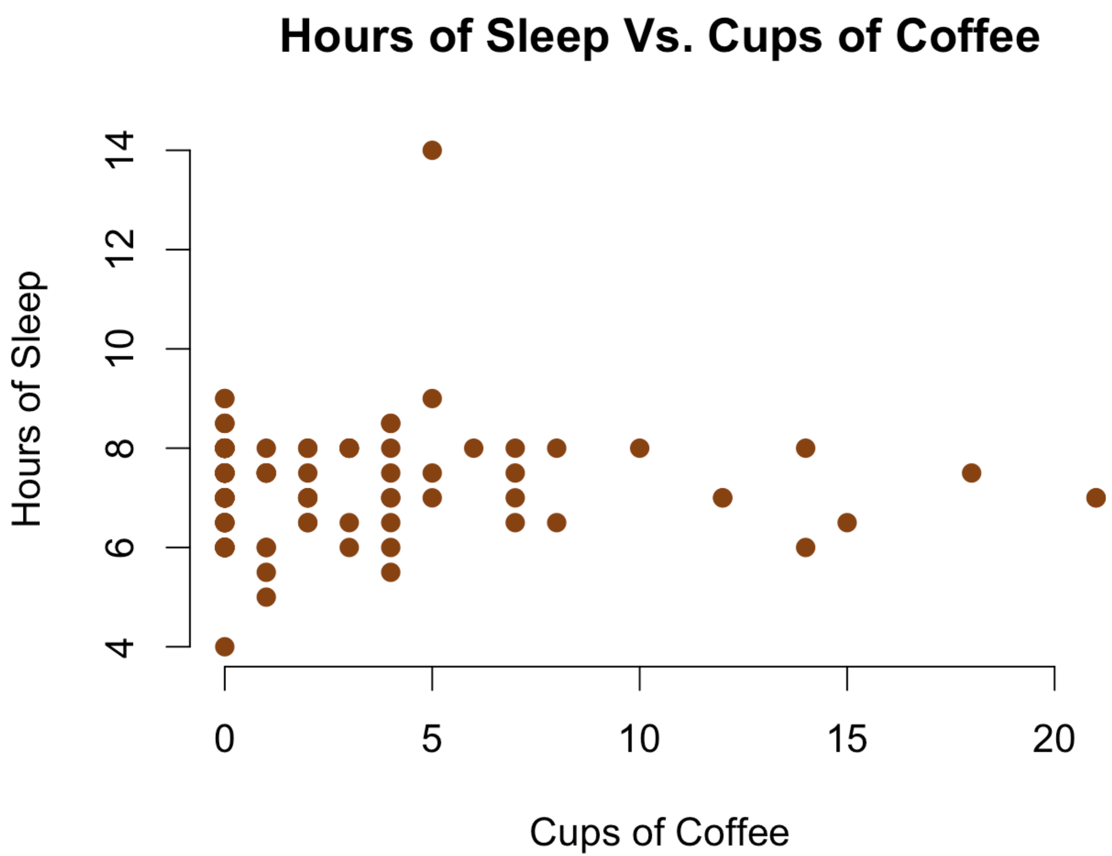
Another example

R codes for scatterplot:

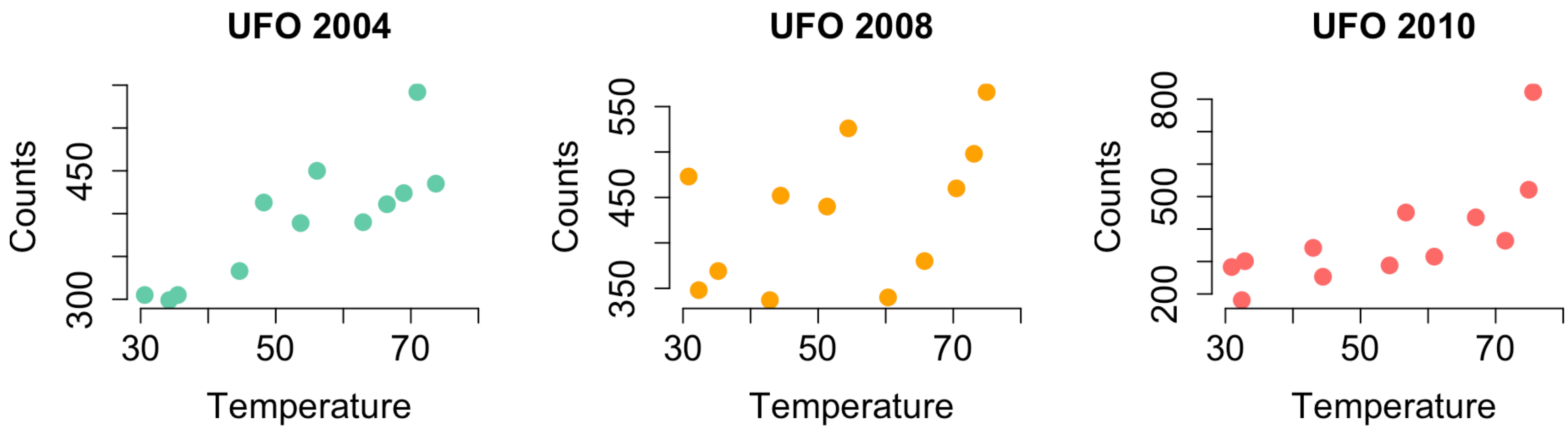
```
plot(x=Survey$ShoeLength,
     y=Survey$Height,
     main="Student Height
           Vs. Shoe Length",
     xlab="Shoe Length",
     ylab="Height",
     col="dodgerblue3", pch=19)
```

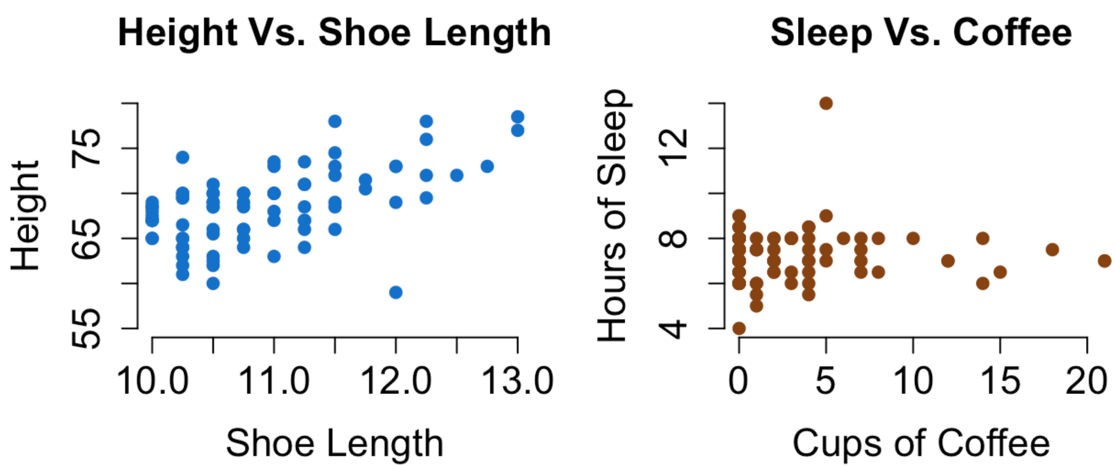


Another example



Linear relationships





- How do we measure the direction and strength of a linear relationship?

Correlation coefficient

The **correlation** measures the *direction* and *strength* of the **linear relationship** between two quantitative variables. Correlation is usually written as ***r***.

Suppose that we have data on variables X and Y for n individuals. The means and standard deviations of the two variables are $\bar{x}$ and s_x for the x-values, and $\bar{y}$ and s_y for the y-values. The correlation r between X and Y is

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

- $\frac{x_i - \bar{x}}{s_x}$ and $\frac{y_i - \bar{y}}{s_y}$: standardized x and y values - no units.
- $\left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)$ can be positive or negative.

Correlation coefficient

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

- $-1 \leq r \leq 1$
- $r > 0$: positive relationship
- $r < 0$: negative relationship
- $r = 0$: no relationship
- $r = \pm 1$: perfect relationship
  - For example: $y = 2x$

Correlation coefficient

```r
cor(Survey$Height, Survey$ShoeLength)
```

```
## [1] NA
```

```r
cor(Survey$Height, Survey$ShoeLength, na.rm=T)
```

```
## Error in cor(Survey$Height, Survey$ShoeLength, na.rm = T): unused argument (na.rm
```

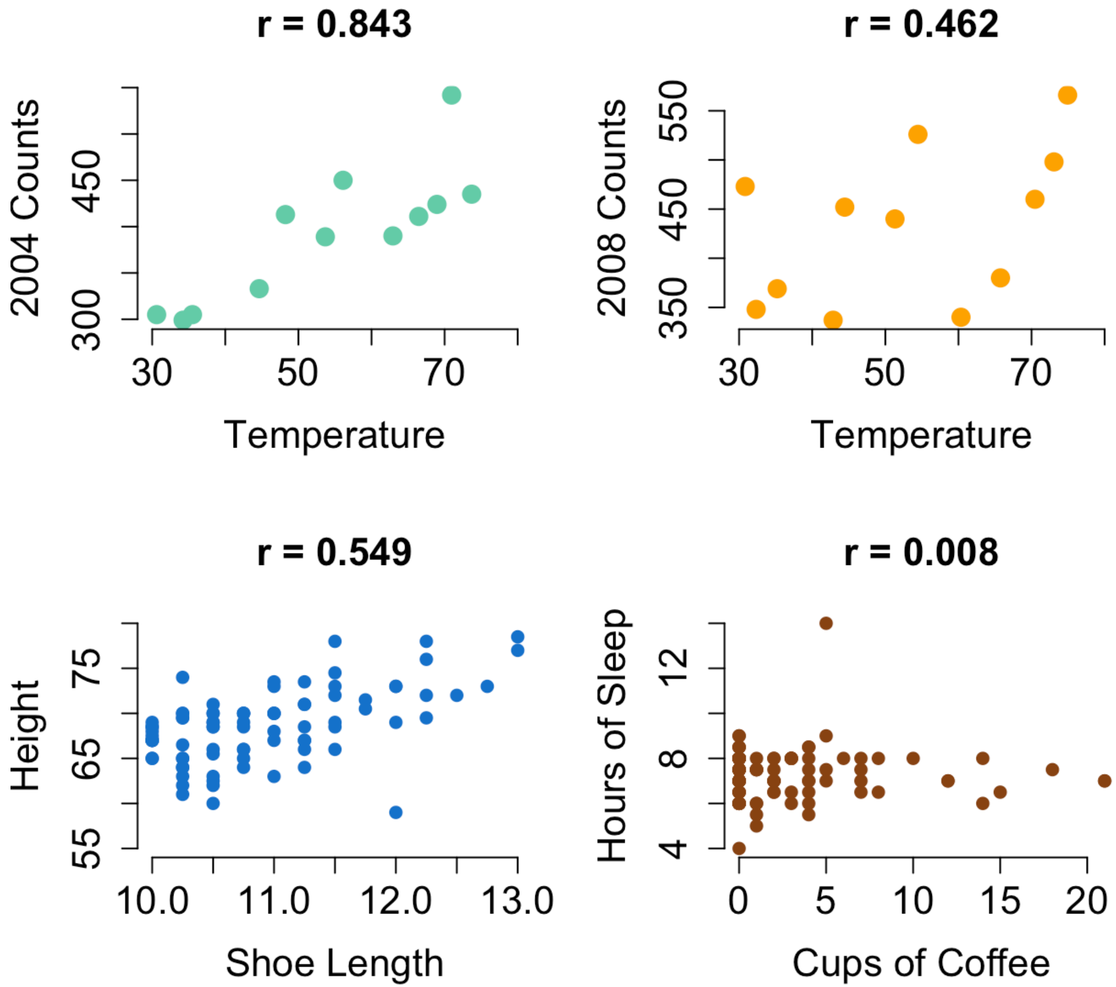
```r
# Remove NAs using use = "complete.obs"
cor(Survey$Height, Survey$ShoeLength, use = "complete.obs")
```

```
## [1] 0.54941
```

```r
# The order of the two variables does not matter
cor(Survey$ShoeLength, Survey$Height, use = "complete.obs")
```

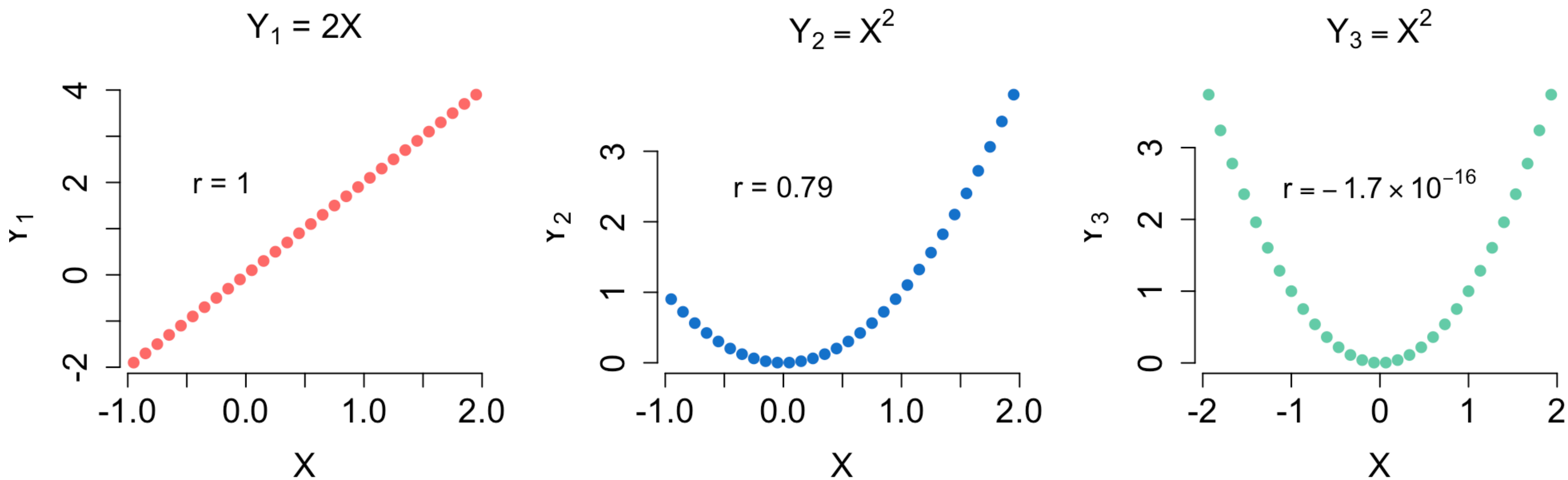
```
## [1] 0.54941
```

Correlation coefficient

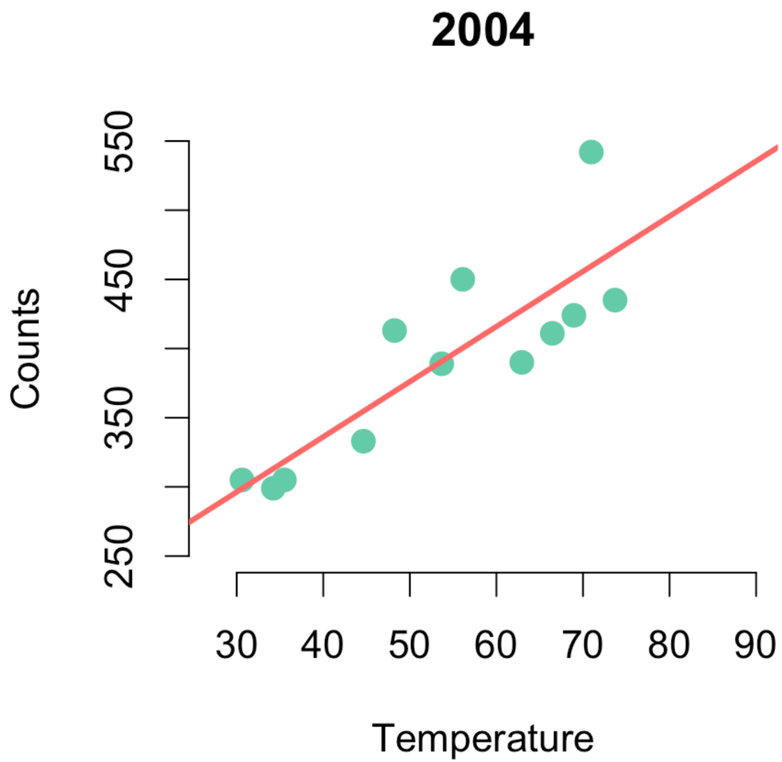


Correlation coefficient

- When calculating correlation, there is no distinction in explanatory or response variable.
- Both varaibles must be quantitative.
- Linear transformation does not alter the value of correlation.
- Correlation only captures the linear relationship betweem two variables.



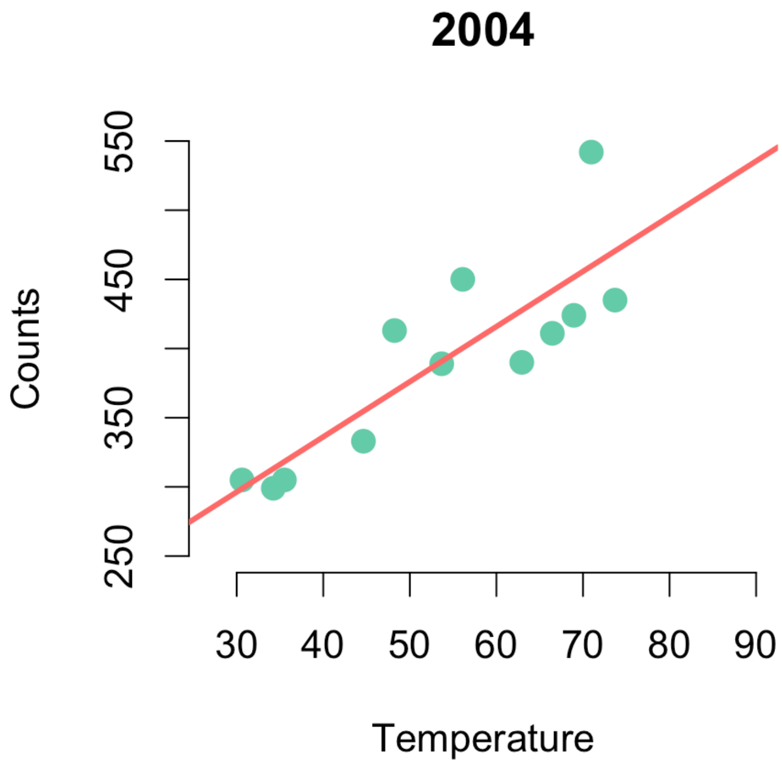
Regression line



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

- Y: response variable (Counts)
  - y: observed values of variable Y
  - $\hat{y}$: predicted values of variable Y
- X: explanatory variable (Temperature)
  - x: observed values of variable X
- e: difference between the observed and the predicted values of Y
- b_0: **intercept**. The value of $\hat{y}$ when $x = 0$
- b_1: **slope**. The amount by which $\hat{y}$ changes when x increases by one unit.

Regression line

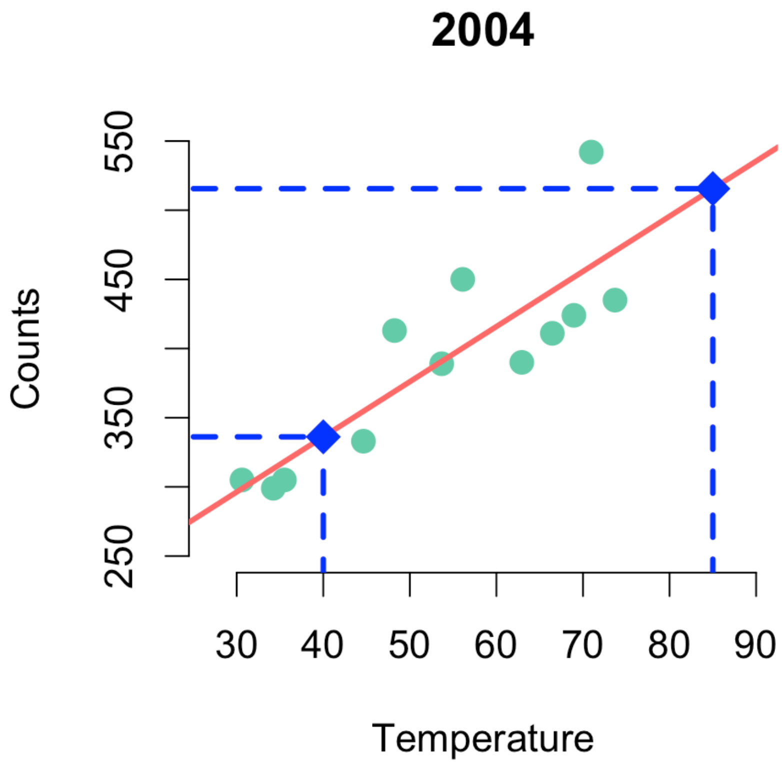


$$\hat{y} = b_0 + b_1 x = 176.7 + 4.0x$$

Interpretation

- $b_0 = 176.7$: the predicted UFO *Count* is 176.7 when *Temperature* is 0.
  - b_0 is the **baseline** value.
  - Sometimes, value of b_0 does not have practical meaning.
- $b_1 = 4.0$: the predicted UFO *Count* increases 4 when *Temperature* increases 1 $^{\circ}$F.
  - b_1 measures the **rate of change**.

Regression line



$$\hat{y} = b_0 + b_1 x = 176.7 + 4.0x$$

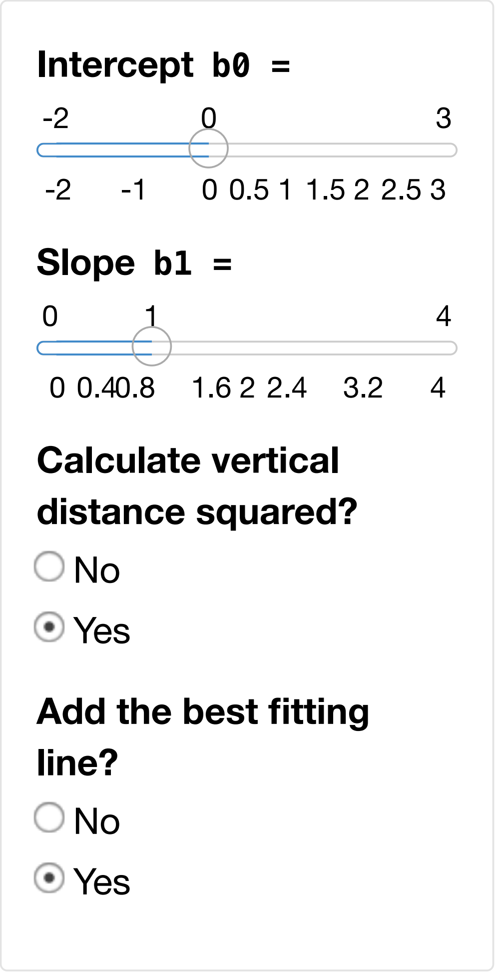
Prediction

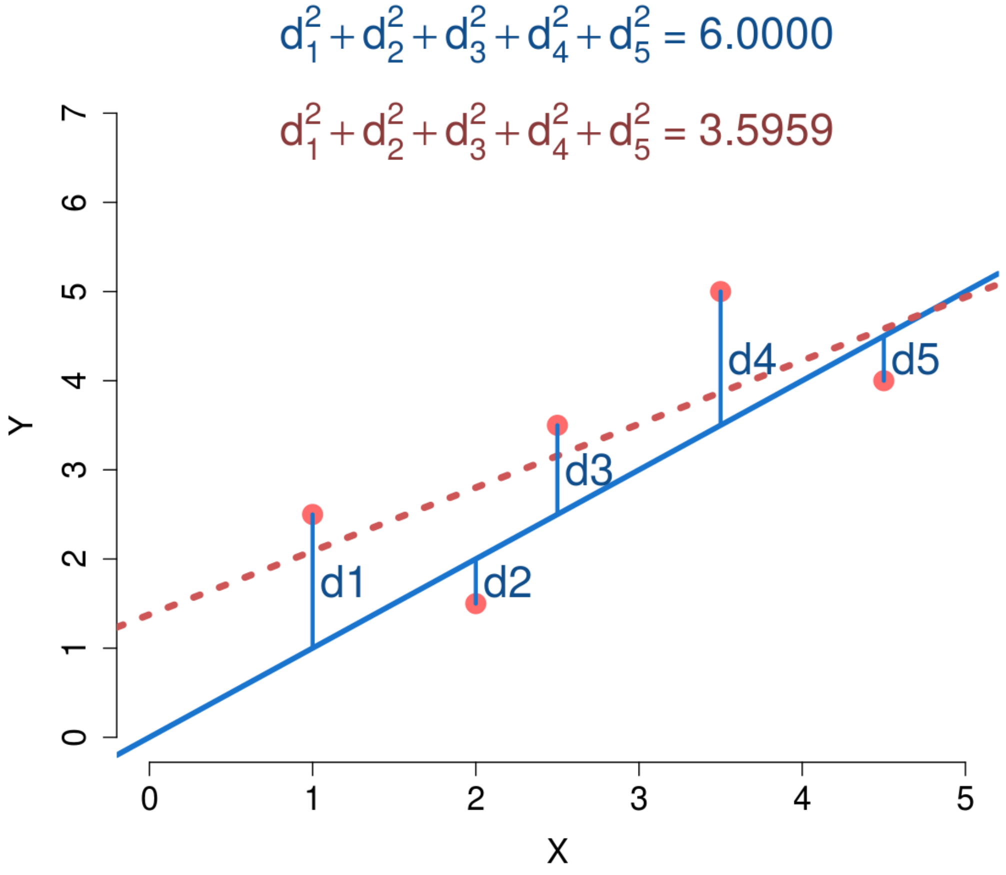
- When Temperature is 40, the predicted UFO count is
  - $\hat{y} = 176.7 + 4.0 \times 40 = 336.7$
- When Temperature is 85, the predicted UFO count is
  - $\hat{y} = 176.7 + 4.0 \times 85 = 516.7$
  - Problem! predicting y with x values outside the range of x is called extrapolation and we should AVOID it.

How to find the best fitting line?

- Connect the point with smallest x value to the one with largest x value
- Find the line that has same number of points above it and below it
- Find the line that is closest to the points in distance
- Find the line that is cloest to the points in the horizontal direction
- Find the line that is closest to the points in the vertical direction

How to find the best fitting line?





Least Square Regression (LSR)

> The **least-squares regression** line of y on x is the line that minimizes the sum of the squares of the vertical distances from the data points to the line.

- Observed data (x_i, y_i) for the i^{th} data point; the total number of data points is n.
- Predicted values $\hat{y_i} = b_0 + b_1 x_i$ for the i^{th} data point.
- Vertical distance from the data points to the line:

$$\text{Residual} = \text{Observed } y - \text{Predicted } y$$
$$e = y - \hat{y}$$
$$e_i = y_i - \hat{y_i}$$

- In least squares regression, we minimize

$$\sum (\text{residual})^2 = \sum_{i=1}^{n} \left(y_i - \hat{y_i}\right)^2 = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2$$

Least Square Regression (LSR)

To minimize

$$\sum(\text{residual})^2 = \sum_{i=1}^{n}\left(y_i - \hat{y_i}\right)^2 = \sum_{i=1}^{n}(y_i - b_0 - b_1 x_i)^2$$

- We search for the values of b_0 and b_1 that result in the smallest $\sum(\text{residual})^2$.

$$\text{Slope } b_1 = r\frac{s_y}{s_x},$$
$$\text{Intercept } b_0 = \bar{y} - b_1\bar{x},$$

where $\bar{x}$ $(\bar{y})$ and s_x (s_y) are the mean and standard deviatioin of x (y); r is the correlation between x and y.

Least Square Regression (LSR) in R

```r
# We use the lm() function in R to obtain the least squares regression line
m <- lm(Height ~ ShoeLength, data=Survey) # y ~ x
m
```

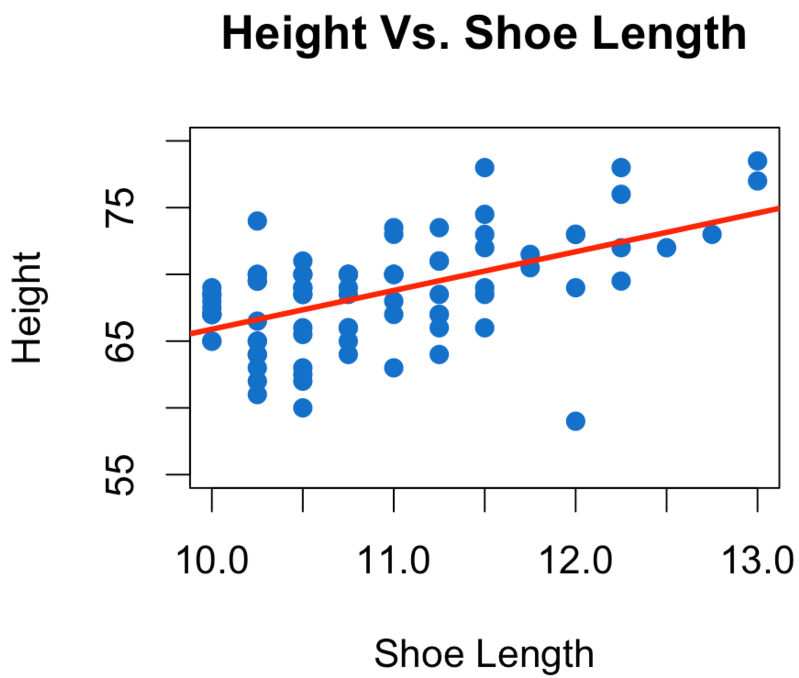
```
##
## Call:
## lm(formula = Height ~ ShoeLength, data = Survey)
##
## Coefficients:
## (Intercept)   ShoeLength
##      36.880        2.902
```

$$\hat{y} = 36.88 + 2.90x$$

$$\widehat{Height} = 36.88 + 2.90 \times ShoeLength$$

Least Square Regression (LSR) in R

```
plot(Survey$ShoeLength, Survey$Height,
     xlab="Shoe Length", ylab="Height", main="Height Vs. Shoe Length",
     col="dodgerblue3", pch=19, ylim=c(55,80))
# Add the regression line to the scatterplot
abline(reg=m, lwd=3, col="red") # or abline(a = 36.88, b = 2.90)
```



Least Square Regression (LSR) in R

```r
# Prediction
predict(m, data.frame(ShoeLength = 10))
```

```
##       1
## 65.8953
```

```r
predict(m, data.frame(ShoeLength = c(10.7, 12, 13.9)))
```

```
##        1        2        3
## 67.92635 71.69830 77.21116
```

```r
predict(m) # Predicting y from ALL the x values in the data set
```

```
##        2        3        5        8       10       11       12       13
## 68.07142 68.79680 65.89530 72.42368 69.52218 74.59981 66.62067 66.62067
##       15       16       17       18       19       21       22       23
## 70.97293 66.62067 67.34605 71.69830 68.79680 69.52218 66.62067 66.62067
##       24       25       27       28       29       30       31       33
## 68.07142 67.34605 68.79680 65.89530 68.07142 66.62067 68.07142 65.89530
##       35       36       38       39       40       41       42       43
## 65.89530 71.69830 67.34605 70.24755 66.62067 65.89530 68.07142 67.34605
##       44       45       47       50       53       54       56       57
## 73.14905 68.07142 69.52218 69.52218 66.62067 66.62067 70.24755 67.34605
```

Summary

- Relationships between variables
- Relationship between two quantitative variables
- Correlation coefficient r, `cor()`
  - Definition and formula
  - Examples
- Least squares regression
  - How to find the best fitting line
    - *Minimize the sum of squares of vertical distances from the points to the line*
  - Least squares regression in R
    - `lm()`, `predict()`

Next lecture: assessing least squares regression line and relationship between two categorial variables