

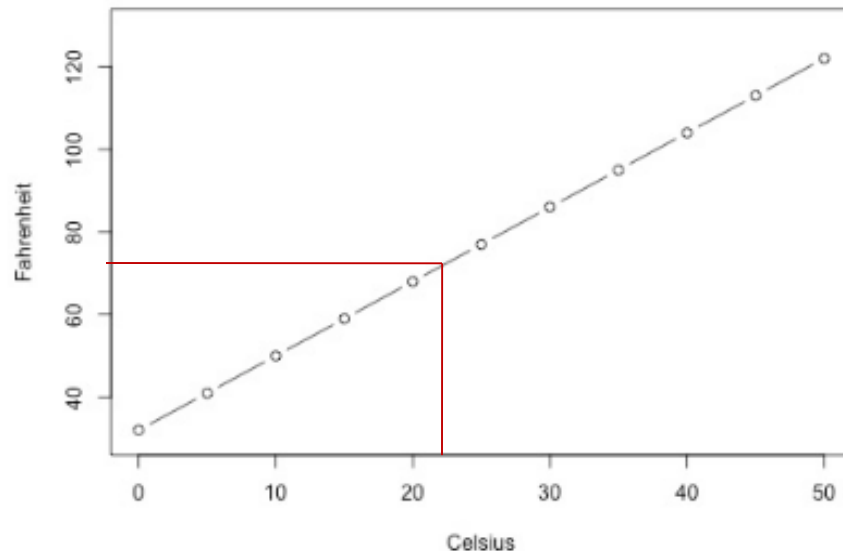
# Unit 2

---

## Simple Linear Regression

# Relationships among Variables

- Temperature in Fahrenheit and degrees Celsius are related as  
$$F = (9/5)C + 32$$
- Equation is used to get exact value of temperature in Fahrenheit for the given value in degrees Celsius
- Observed values of data points fall directly on a line



deterministic relationship

# Relationships among Variables



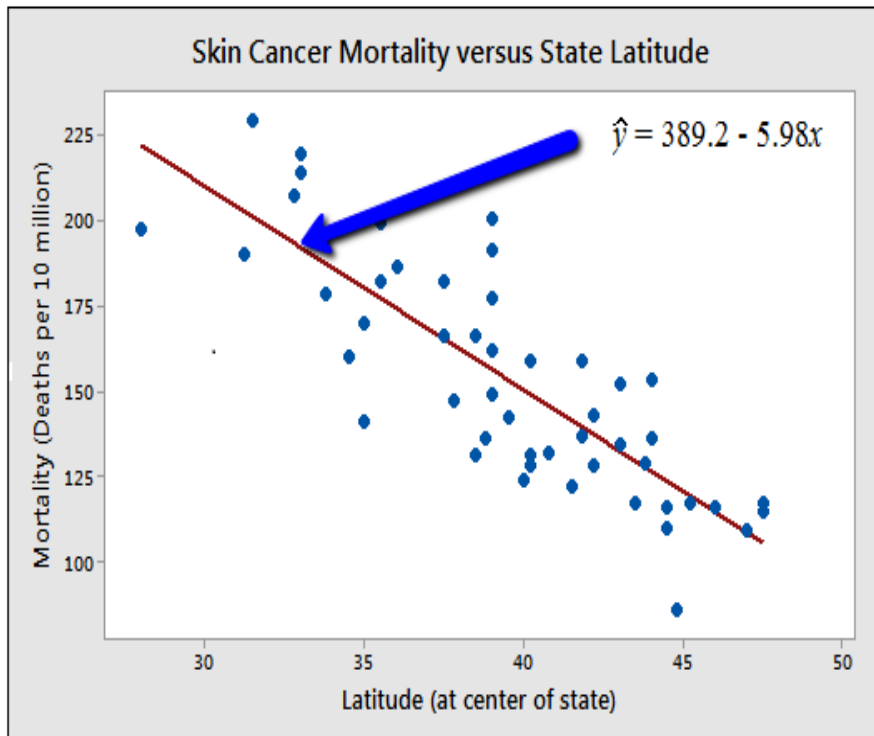
- Variables can have any of the following relationships
  - Deterministic
  - Statistical

# Examples: Deterministic Relationships

- Circumference =  $\pi \times \text{diameter}$
- Ohm's Law:  $I = V/r$   
where  $V$  = voltage applied,  $r$  = resistance, and  
 $I$  = current
- For each of these deterministic relationships, the equation *exactly* describes the relationship between the two variables
- In statistical relationships, the relationship between the variables is not perfect

# Example: Statistical Relationship

- Mortality due to skin cancer (number of deaths per 10 million people) and the latitude (degrees North) at the center of each of states in the U.S.
- The scatter plot supports such a hypothesis



- A person living in the higher latitudes is less exposed the harmful rays of the sun
- Therefore, person has less risk of death due to skin cancer
- Relationship is not perfect

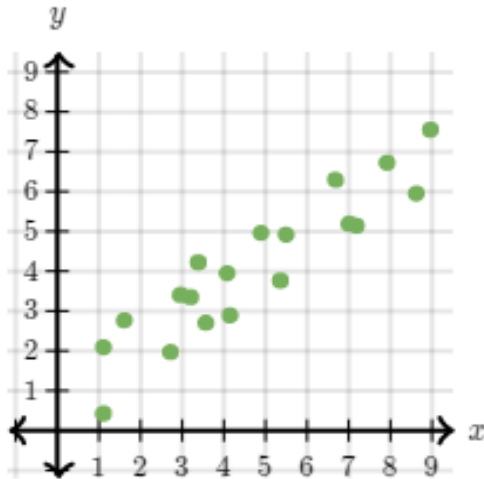
# Examples: Statistical Relationship

- Height and weight
  - As height increases, we expect weight to increase
  - It does not increase perfectly
- Driving speed and gas mileage
  - As driving speed increases, we expect fuel mileage to decrease
  - Does not decrease perfectly

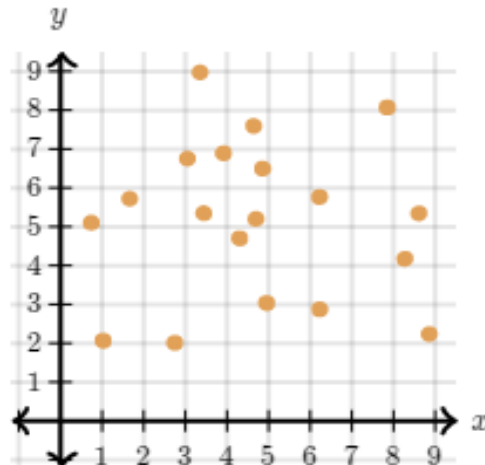
# Types of Association

Scatter plots are used to see relationships between variables

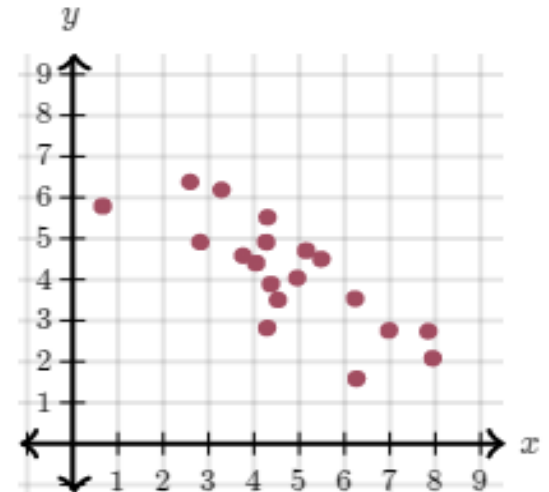
Positive association



No association

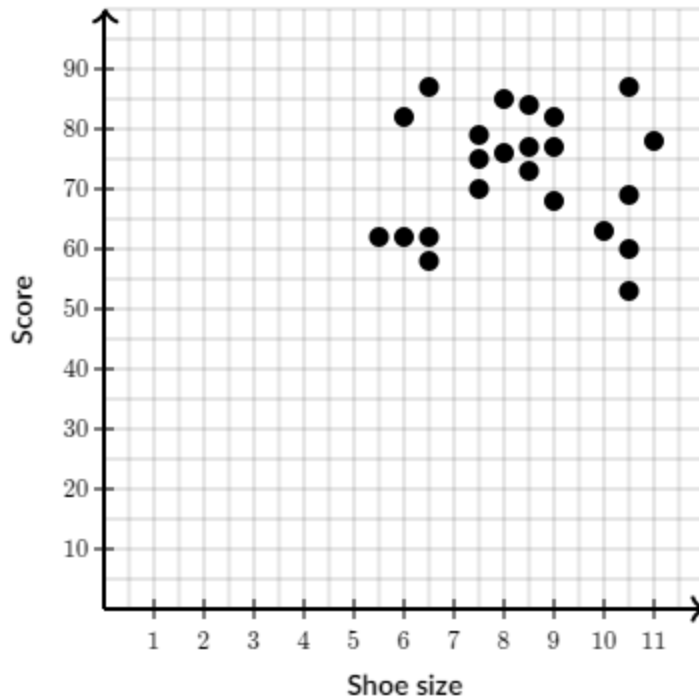


Negative association



# Ex 1: Shoe sizes and test scores

- Data set shows test grades and shoe sizes of students in a class
- The data is shown in the scatter plot



What is the best description of the relationship between shoe size and test scores?

Positive association

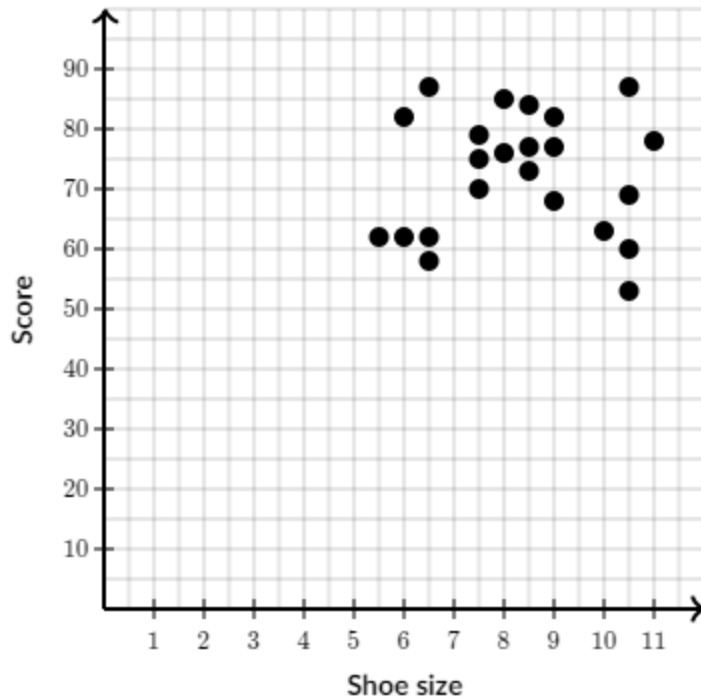
Negative association

No association



# Ex 1: Shoe sizes and test scores

- Data set shows test grades and shoe sizes of students in a class
- The data is shown in the scatter plot



What is the best description of the relationship between shoe size and test scores?

Positive association

Negative association

**No association**

## Ex 2: Flower height and petal length

- Measured the height and petal length (in centimeters) of all the flowers in a garden

Height (cm)	30	20	15	35	10	40
Petal length (cm)	6	4	2	8	1.5	8.5

- What is the best description of the relationship between height and petal length for the flowers?  
Positive association  
Negative association  
No association

## Ex 2: Flower height and petal length

- Measured the height and petal length (in centimeters) of all the flowers in a garden

Height (cm)	30	20	15	35	10	40
Petal length (cm)	6	4	2	8	1.5	8.5

- What is the best description of the relationship between height and petal length for the flowers?

Positive association

Negative association

No association

# What Is Regression?

- Regression searches for relationships among variables
- Ex 1: Observe several employees of a company and try to understand how their salaries depend on the features, such as experience, level of education, role, city they work in, and so on
- Data related to each employee represent one observation
- Experience, education, role, and city are **independent** features
- Salary **depends** on independent features
- Ex 2: Establish a mathematical dependence of the prices of houses on their areas, numbers of bedrooms, distances to the city center, and so on.

# Regression

- Dependent variables are called outputs or responses
- Independent variables are called inputs or predictors
- Regression problems usually have one continuous and unbounded dependent variable
- Inputs, can be continuous, discrete, or even categorical data such as gender, nationality, brand, and so on
- It is a common practice to denote the outputs with  $y$  and inputs with  $x$
- If there are two or more independent variables, they can be represented as  
 $\mathbf{x} = (x_1, \dots, x_r)$ , where  $r$  is the number of inputs

# When Do You Need Regression?

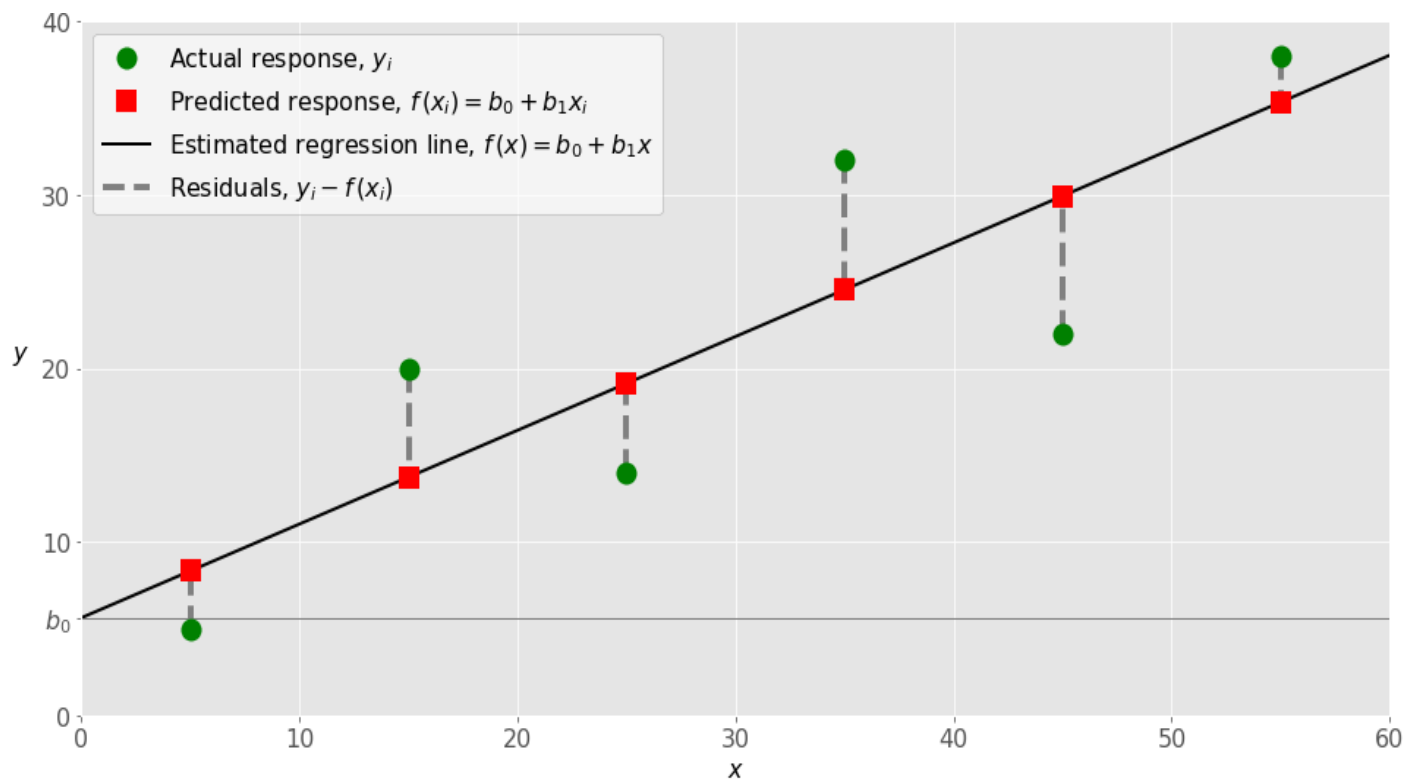
- How one or more variables influence the other
- Ex 1: Use regression to determine *if* and *to what extent* the experience or gender impact salaries
- Also useful when you want to forecast a response using a new set of predictors
- Ex 2: Try to predict electricity consumption of a household for the next hour given the outdoor temperature, time of day, and number of residents in that household
- Regression is used in fields like economy, computer science, social sciences
- Its importance rises every day with the availability of large amounts of data and increased awareness of the practical value of data

# Simple linear regression

- A statistical method that allows us to summarize and study relationships between two continuous (quantitative) variables:
- One variable, denoted  $x$ , is regarded as the predictor, explanatory, or independent variable
- The other variable, denoted  $y$ , is regarded as the response, outcome, or dependent variable.
- Simple linear regression gets its adjective "simple," because it concerns the study of only one predictor variable
- In contrast, multiple linear regression gets its adjective "multiple," because it concerns the study of two or more predictor variables

# Simple Linear Regression

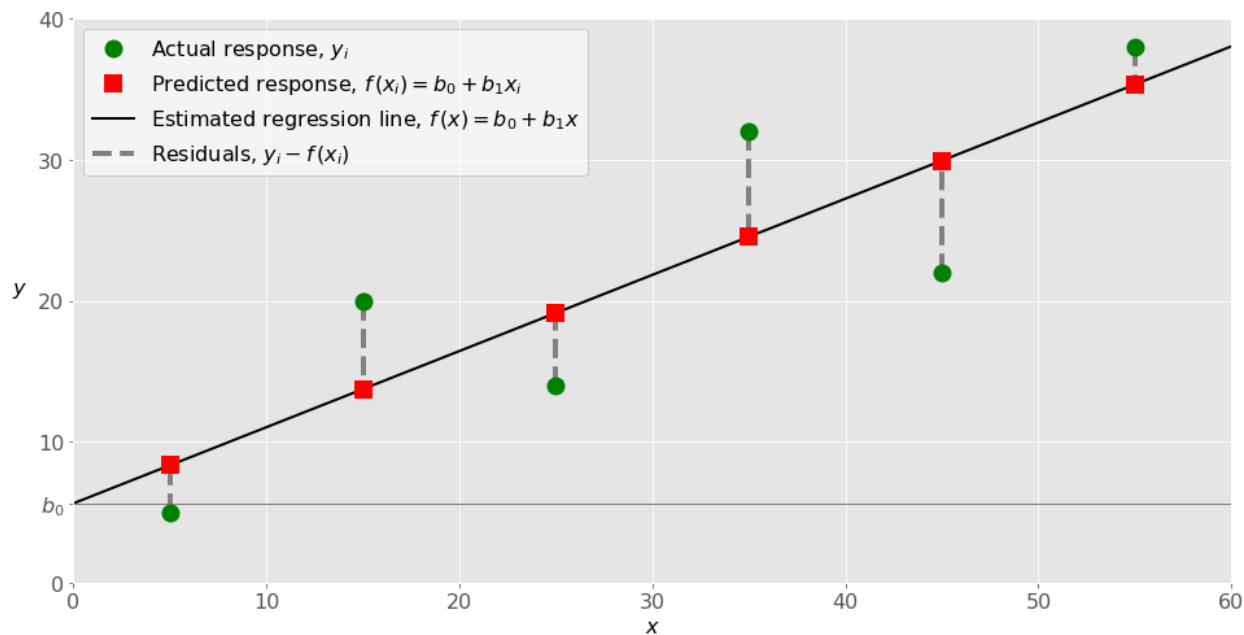
- Is the simplest case of linear regression with a single independent variable,  $\mathbf{x} = x$





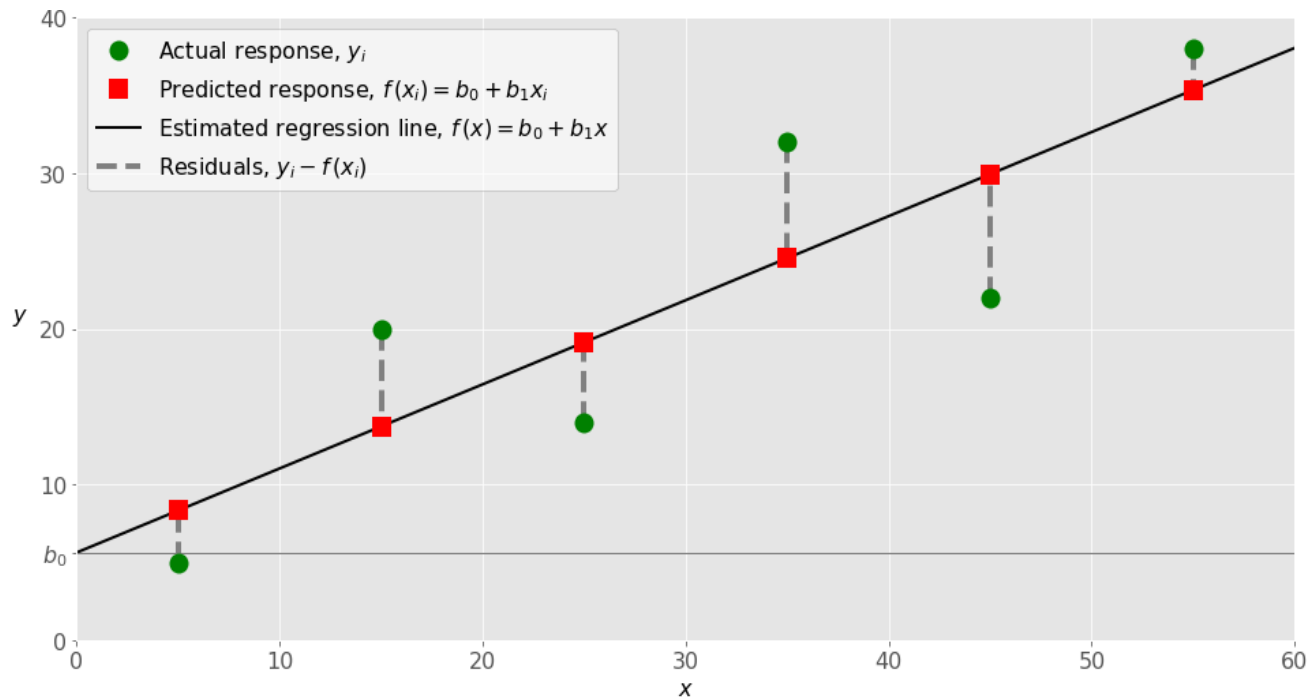
# Simple Linear Regression

- The estimated regression function (black line) is represented by  $f(x) = b_0 + b_1x$
- Goal is to calculate the optimal values of the predicted weights  $b_0$  and  $b_1$  that minimize residual
- The value of  $b_0$ , also called the **intercept**, shows the point where the estimated regression line crosses the  $y$  axis.
- The value of  $b_1$  determines the **slope** of the estimated regression line.



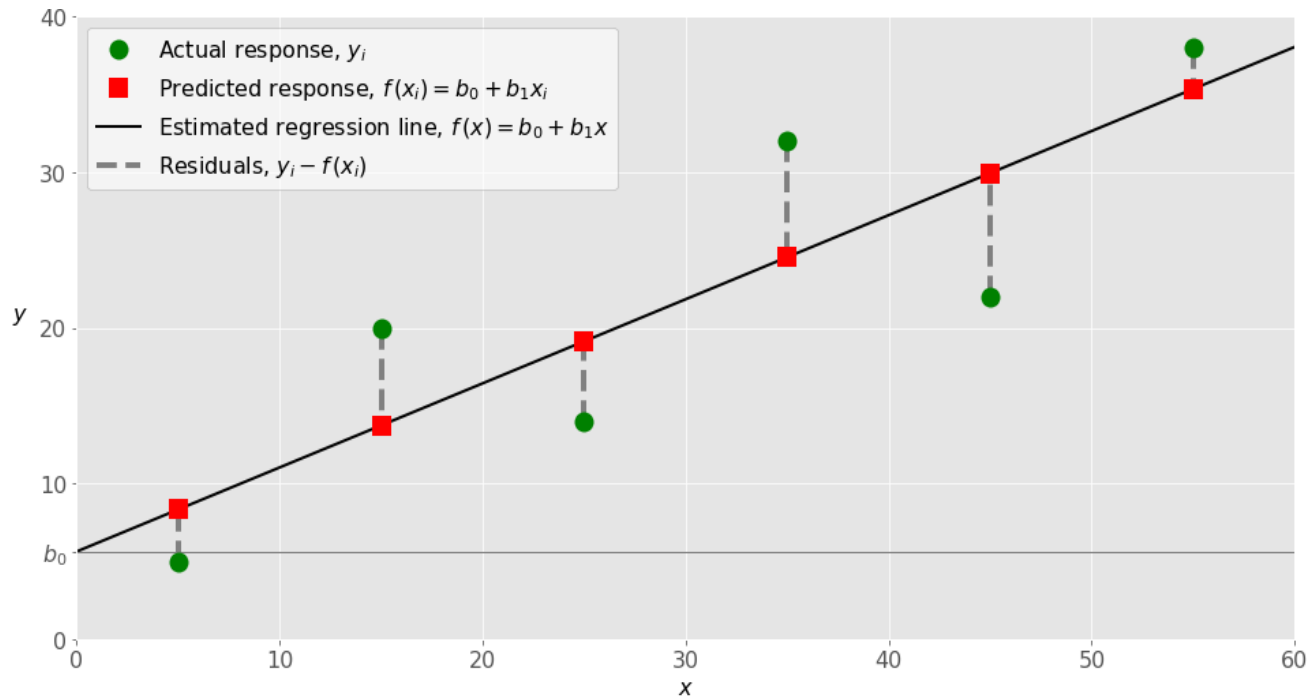
# Simple Linear Regression

- The predicted responses (red squares) are the points on the regression line that correspond to the input values
- For the input  $x = 5$ ,
- Predicted response is  $f(5) = 8.33$



# Simple Linear Regression

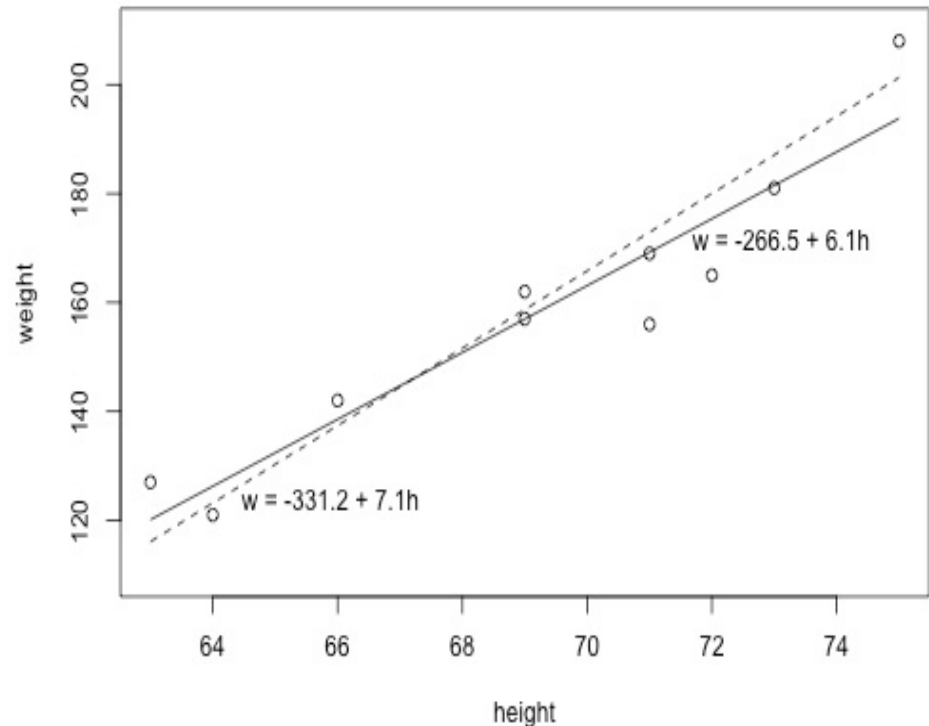
- Residuals (vertical dashed gray lines) can be calculated as  $y_i - f(\mathbf{x}_i) = y_i - (b_0 + b_1x_i)$  for  $i = 1, \dots, n$
- Residuals are the distances between the green circles and red squares
- Linear regression minimizes these distances and make the red squares as close to the predefined green circles as possible



# What is the "Best Fitting Line"?

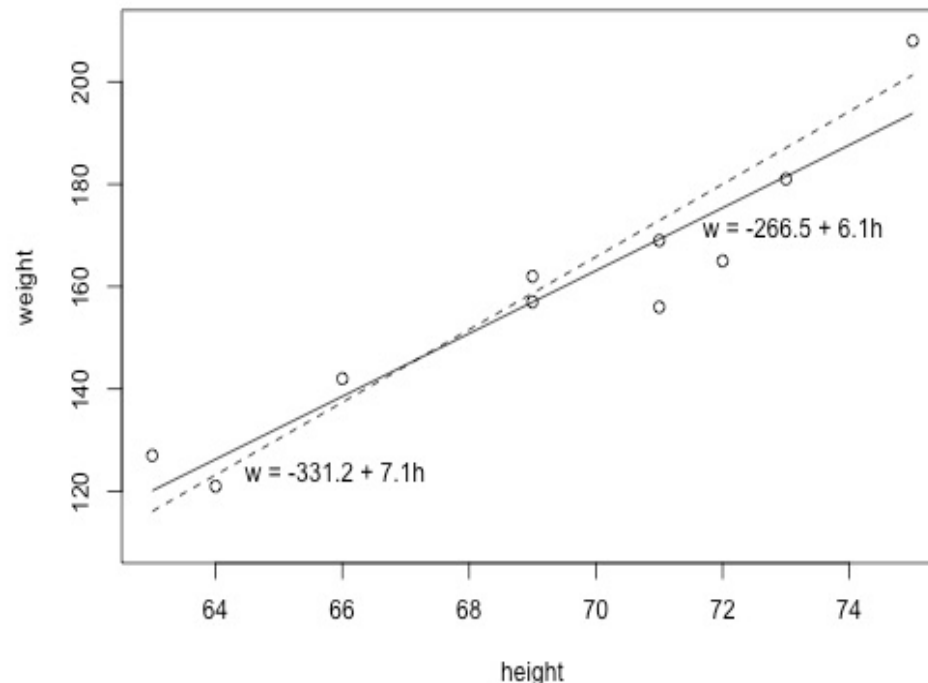
- Line which summarizes the trend between height and weight in the best way

$i$	$x_i$	$y_i$
1	63	127
2	64	121
3	66	142
4	69	157
5	69	162
6	71	156
7	71	169
8	72	165
9	73	181
10	75	208



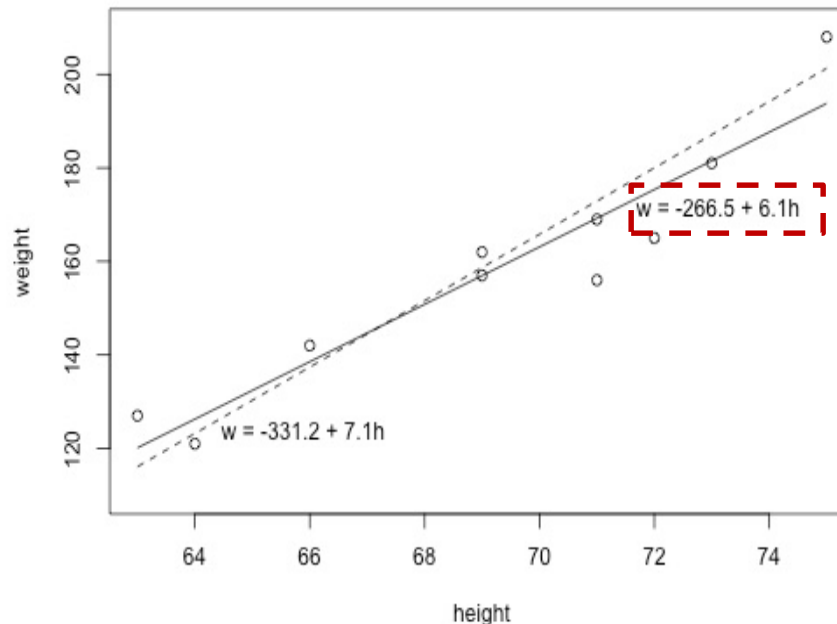
# What is the "Best Fitting Line"?

- $y_i$  denotes the weight for data point,  $i$
- $x_i$  denotes the height for data point,  $i$
- $\hat{y}_i$  is the predicted response (or fitted value) for a data point,  $i$
- The equation for the best fitting line is:  $\hat{y}_i = b_0 + b_1 x_i$
- Relation is summarized by a line  $w = -266.53 + 6.1376 h$



# What is the "Best Fitting Line"?

- $w = -266.53 + 6.1376 h$
- Given that a student is 63 inches tall and weighs 127 pounds
- Predicted student's weight is
$$\hat{y}_1 = 266.53 + 6.1376(63)$$
$$= 120.1$$
- Prediction is not perfectly correct
- **"prediction error" (or "residual error")** =  $127 - 120.1 = 6.9$  pounds



# What is the "Best Fitting Line"?

$$w = -266.53 + 6.1376 h$$

$i$	$x_i$	$y_i$	$\hat{y}_i$
1	63	127	120.1
2	64	121	126.3
3	66	142	138.5
4	69	157	157.0
5	69	162	157.0
6	71	156	169.2
7	71	169	169.2
8	72	165	175.4
9	73	181	181.5
10	75	208	193.8

- To predict response,  
 $\hat{y}_i = b_0 + b_1 x_i$
- Actual response is  $y_i$
- Prediction error (or residual error) is  
 $e_i = y_i - \hat{y}_i$

# What is the "Best Fitting Line"?

- A line that fits the data "best" is the one for which  
 $n$  prediction errors (one for each observed data point) are as small as possible in some overall sense
- least squares criterion is based on  
"minimize the sum of the squared prediction errors"



# What is the "Best Fitting Line"?

- Best fitting line is:  $\hat{y}_i = b_0 + b_1 x_i$
- Find the values  $b_0$  and  $b_1$  that make the sum of the squared prediction errors the smallest it can be
- That is

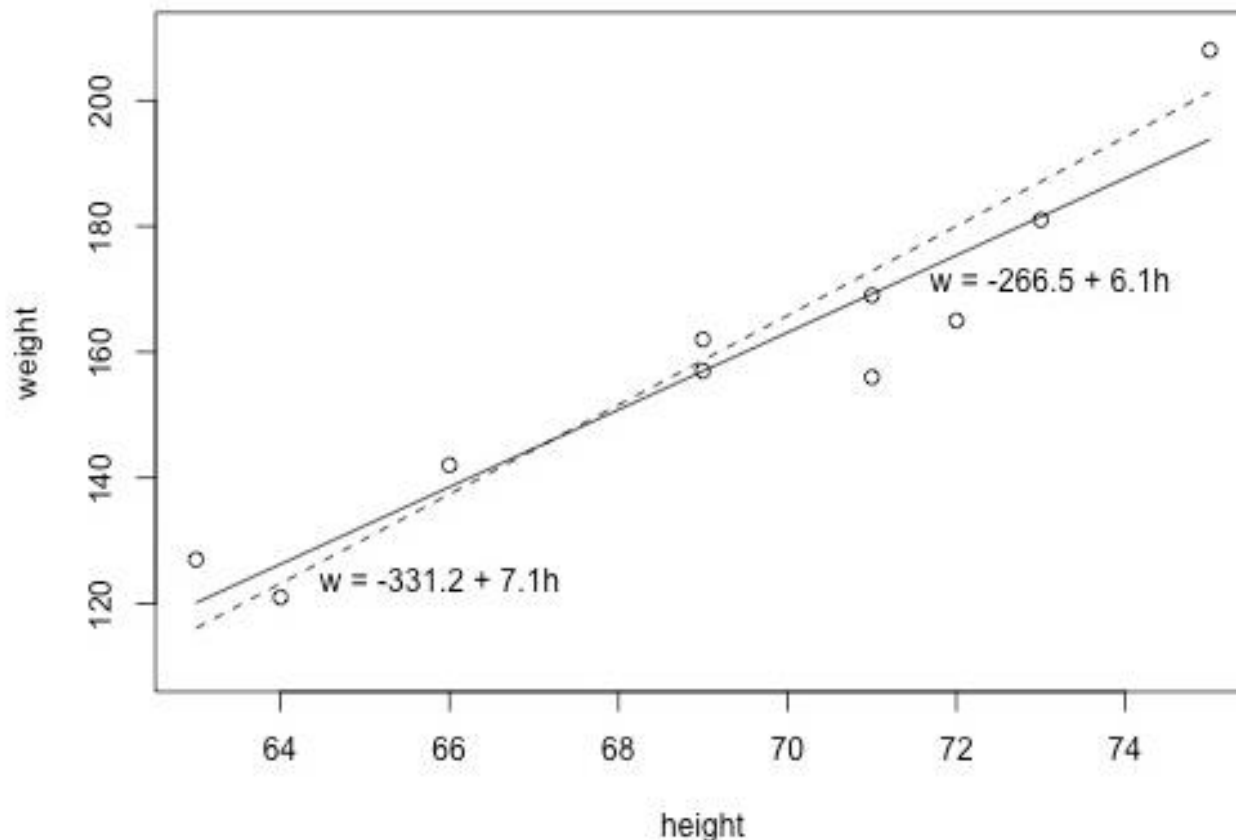
$$Q = \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

# What is the "Best Fitting Line"?

Two lines are

$w = -266.5 + 6.1h$  and  $w = -331.2 + 7.1h$

Determine the total error for each line



# What is the "Best Fitting Line"?

$w = -331.2 + 7.1 h$ (the dashed line)					
$i$	$x_i$	$y_i$	$\hat{y}_i$	$(y_i - \hat{y}_i)$	$(y_i - \hat{y}_i)^2$
1	63	127	116.1	10.9	118.81
2	64	121	123.2	-2.2	4.84
3	66	142	137.4	4.6	21.16
4	69	157	158.7	-1.7	2.89
5	69	162	158.7	3.3	10.89
6	71	156	172.9	-16.9	285.61
7	71	169	172.9	-3.9	15.21
8	72	165	180.0	-15.0	225.00
9	73	181	187.1	-6.1	37.21
10	75	208	201.3	6.7	44.89
					<u>766.5</u>

$w = -266.53 + 6.1376 h$ (the solid line)					
$i$	$x_i$	$y_i$	$\hat{y}_i$	$(y_i - \hat{y}_i)$	$(y_i - \hat{y}_i)^2$
1	63	127	120.139	6.8612	47.076
2	64	121	126.276	-5.2764	27.840
3	66	142	138.552	3.4484	11.891
4	69	157	156.964	0.0356	0.001
5	69	162	156.964	5.0356	25.357
6	71	156	169.240	-13.2396	175.287
7	71	169	169.240	-0.2396	0.057
8	72	165	175.377	-10.3772	107.686
9	73	181	181.515	-0.5148	0.265
10	75	208	193.790	14.2100	201.924
					<u>597.4</u>

- $w = -266.5 + 6.1376h$ , best summarizes the data
- Does not guarantee to be the best fitting line of all of the possible lines

# Determine $b_0$ and $b_1$ for Least Error

$$Q = \sum_{i=1}^n (y_i - (b_0 + b_1 x_i))^2$$

- For the **least squares error**  
 $b_0 = \bar{y} - b_1 \bar{x}$ ,  $\bar{y}$  and  $\bar{x}$  are mean values
- Least squares line passes through the point  $(\bar{x}, \bar{y})$ ,

$$b_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

- Or

$$b_1 = (\bar{x}\bar{y} - \overline{xy}) / (\bar{x}^2 - \overline{x^2})$$

# Example: least squares regression line

- 3 points are (1,2), (2,1), (4,3)
- plot them on axes
- find best fitting regression

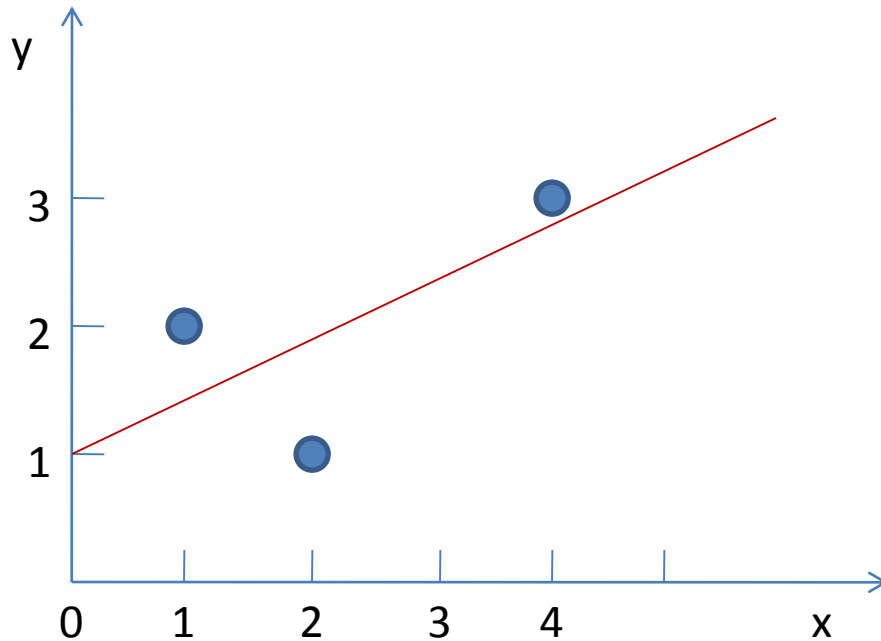
$$b_1 = (\bar{x}\bar{y} - \overline{xy}) / (\bar{x}^2 - \overline{x^2})$$

$$\bar{x} = \frac{1 + 2 + 4}{3} = \frac{7}{3}, \bar{y} = 2, \overline{xy} = \frac{16}{3}, \overline{x^2} = 7,$$

- $b_1 = \{2(7/3) - 16/3\} / \{(49/9) - 7\}$   
 $= 3/7$
- $b_0 = \bar{y} - b_1\bar{x}$   
 $= 1$
- best fitting line is  $y = (3/7)x + 1$

# Least Squares Regression Line

- 3 points are (1,2), (2,1), (4,3)
- best fitting line is  $y = (3/7)x + 1$



# Significance of $b_0$ and $b_1$

- Given relation between height and weight is  
 $w = -266.53 + 6.1376h$
- If a person is 0 inches tall then his predicted weight is 266.53 pounds!
- Scope of the model does not include  $x = 0$
- It is "extrapolated" beyond the "scope of the model"
- If the "scope of the model" includes  $x = 0$ , then  $b_0$  is the predicted mean response when  $x = 0$
- Otherwise,  $b_0$  is not meaningful
- Response increases or decreases by  $b_1$  units for every one unit increase in  $x$ .

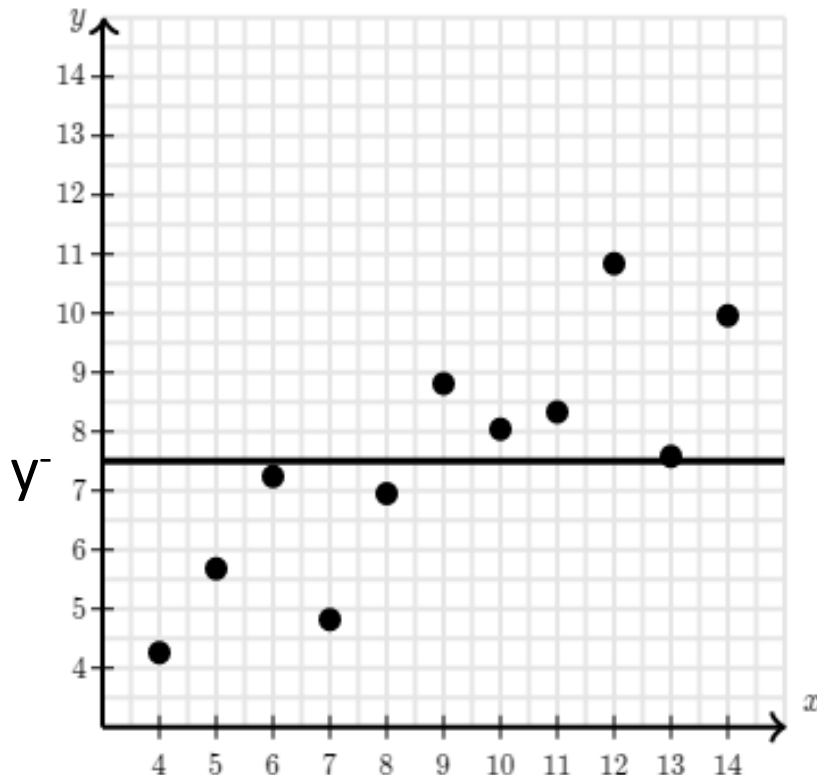
# Coefficient of Determination, r-squared ( $R^2$ )

- Linear regression is used to predict  $y$  given some value of  $x$ .
- Measures how much prediction error is eliminated when we use least-squares regression
- Larger  $R^2$  indicates a better fit and means that the model can better explain the variation of the output with different inputs
- The value  $R^2 = 1$  corresponds to **perfect fit**
- Then Total error = 0
- Perfect fit shows the values of predicted and actual responses fit completely to each other



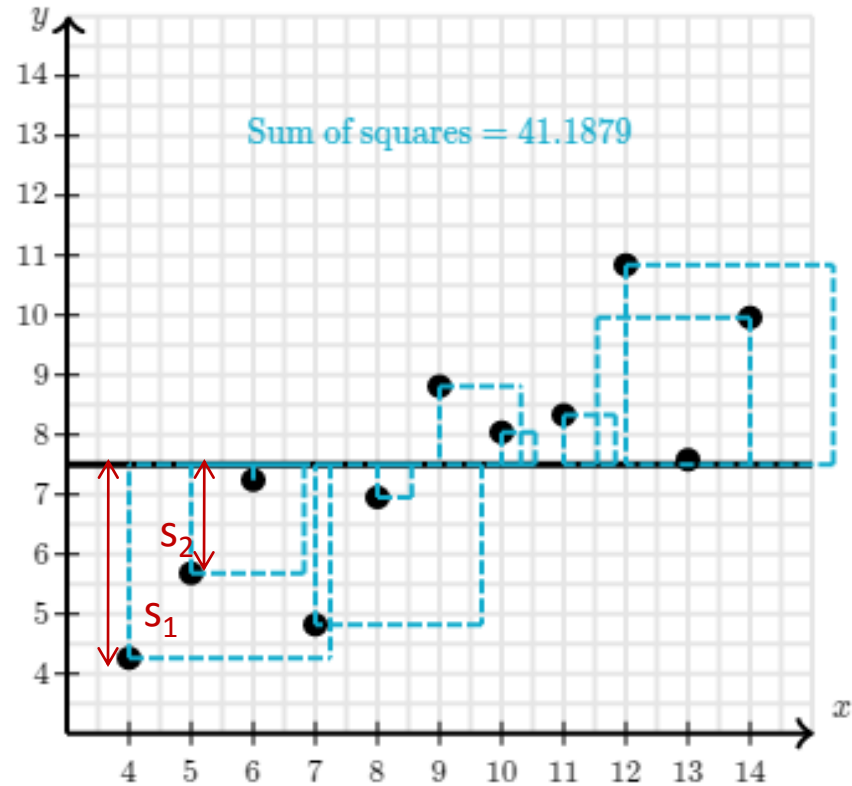
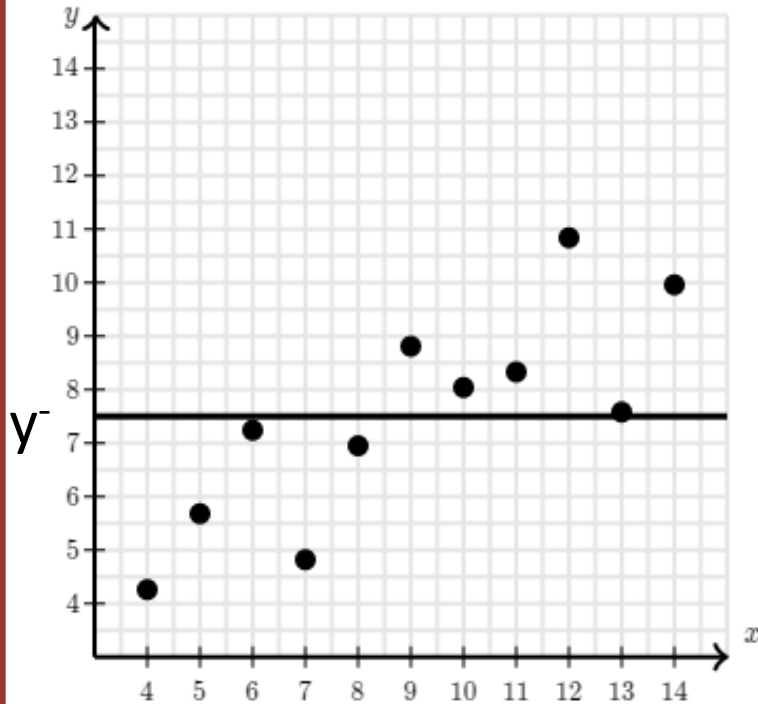
# Predicting without Regression

- Without using regression on the  $x$  variable, most reasonable estimate is to predict the average of the  $y$  values.
- Line shows average value of output



- line doesn't fit the data very well
- To measure the fit of the line
- Calculate the Sum of the Square Residuals (SSR)
- SSR gives an overall sense of how much prediction error a given model has

# Predicting without Regression

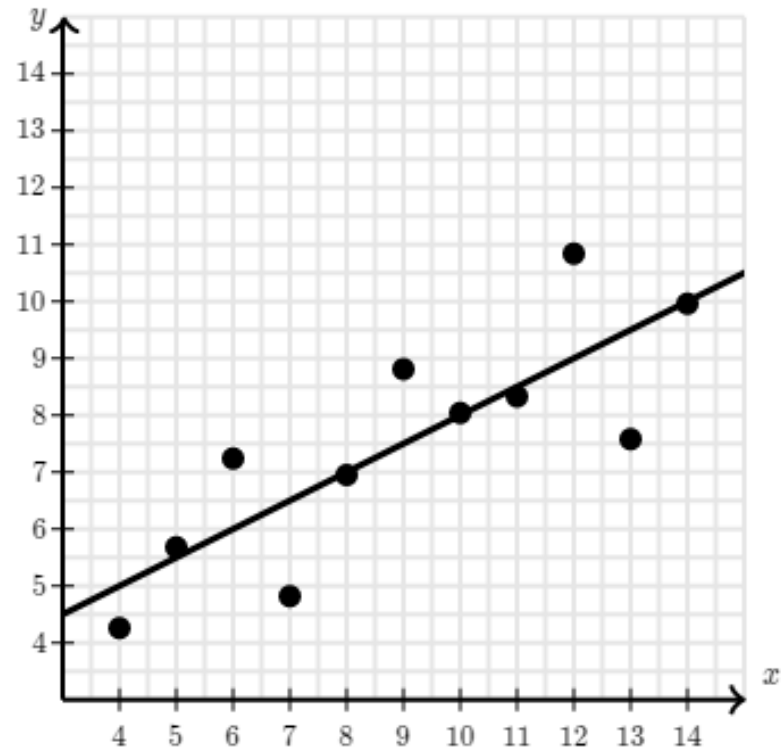
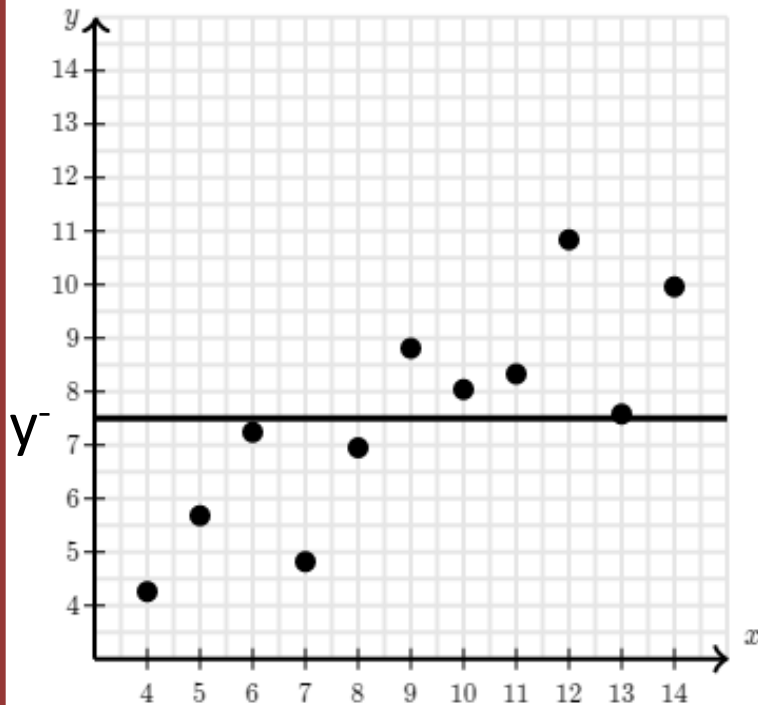


- Prediction Error is Sum of Square, SSR

$$S_1^2 + S_2^2 + \dots = 41.1879$$

# Predicting with Regression

- Regression line is  
$$\hat{y} = 0.5x + 1.5$$

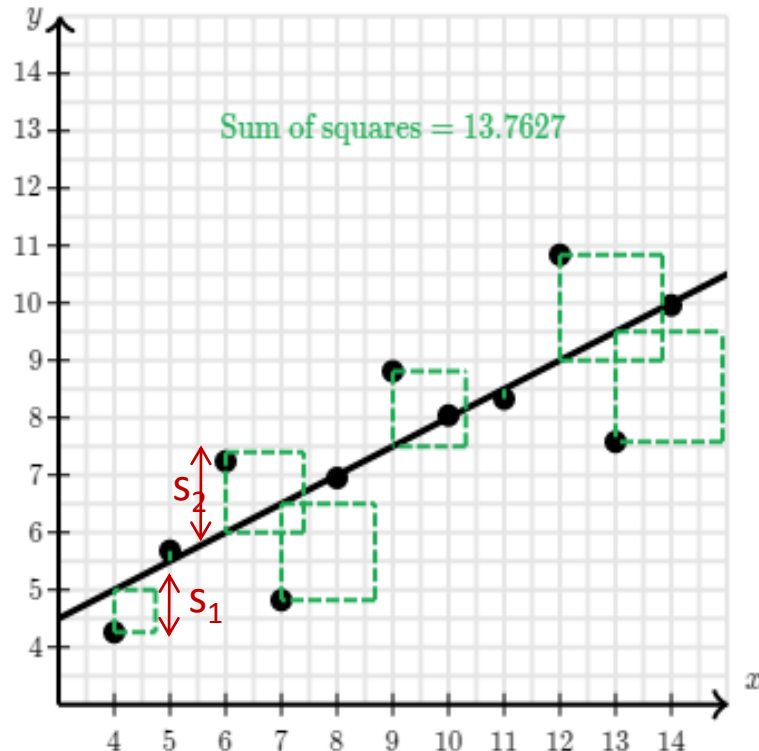
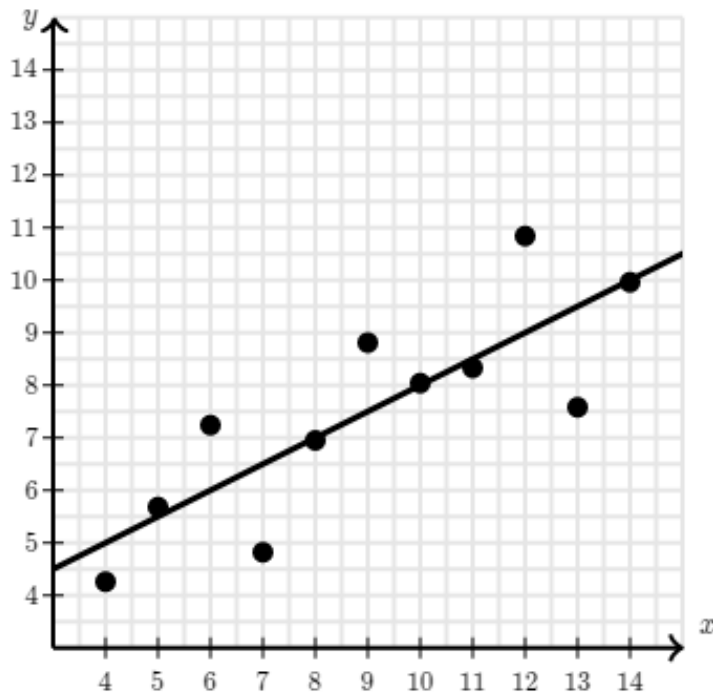


# Predicting with Regression

- Prediction Error is Sum of Square, SSR

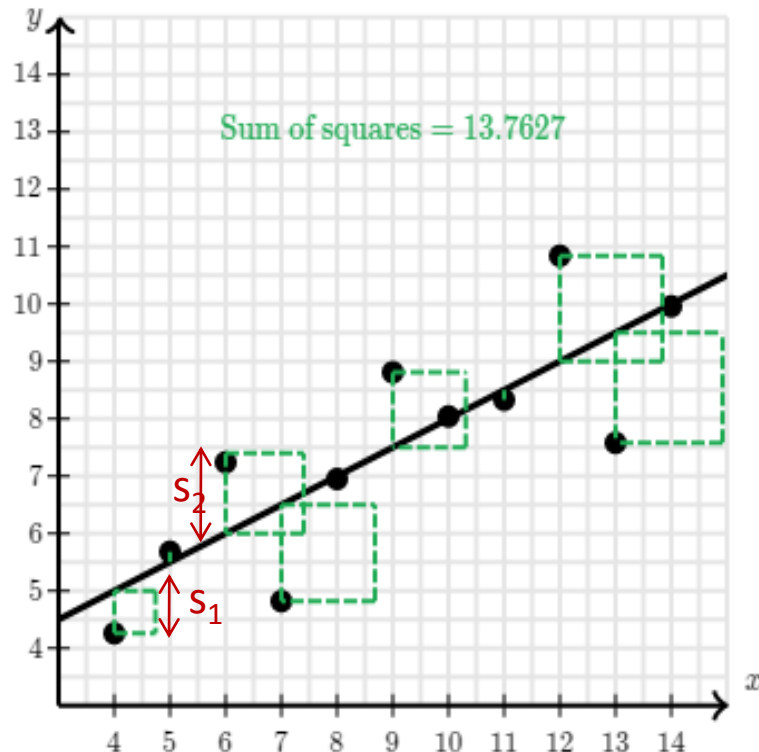
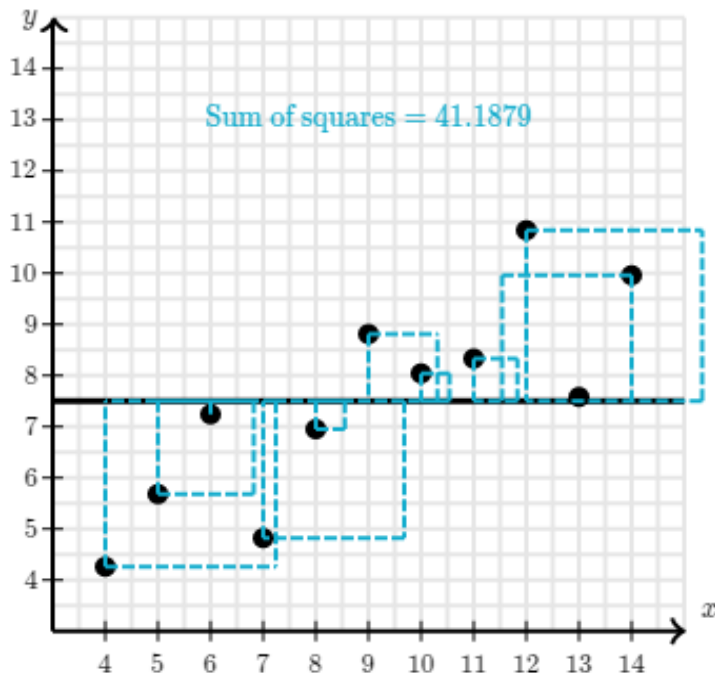
$$S_1^2 + S_2^2 + \dots = 13.7627$$

- Least-squares regression reduces the amount of prediction error



# Predicting with Regression

- Least-squares regression reduced the sum of the squared residuals from 41.1879 to 13.7627
- Reduction in prediction error is  $41.1879 - 13.7627 = 27.4252$
- R-squared measures how much prediction error is eliminated



# Coefficient of Determination, $R^2$ or $r^2$

- Reduction as a percentage of the original amount of prediction error is

$$\frac{41.1879 - 13.7627}{41.1879} = \frac{27.4252}{41.1879} \approx 66.59\%$$

- Coefficient of determination,  $r^2 = 0.6659$
- R-squared represents what percent of the prediction error in the y variable is eliminated when we use least-squares regression on the x variable

# Coefficient of Determination, $R^2$ or $r^2$

- $r^2$  represents the percent of the variability in the  $y$  variable by the regression on the  $x$  variable
- To determine  $r$ -square, SSR and SSE are computed
- SSR is the "regression sum of squares"
  - quantifies how far the estimated sloped regression line,  $\hat{y}_i$ , is from the horizontal "no relationship line," the sample mean or  $\bar{y}$
- SSE is the "error sum of squares"
  - quantifies how much the data points,  $y_i$ , vary around the estimated regression line,  $\hat{y}_i$
- SSTO is the "total sum of squares"
  - quantifies how much the data points,  $y_i$ , vary around their mean,  $\bar{y}$
- $SSTO = SSR + SSE$

# Example: r-squared (revisited)

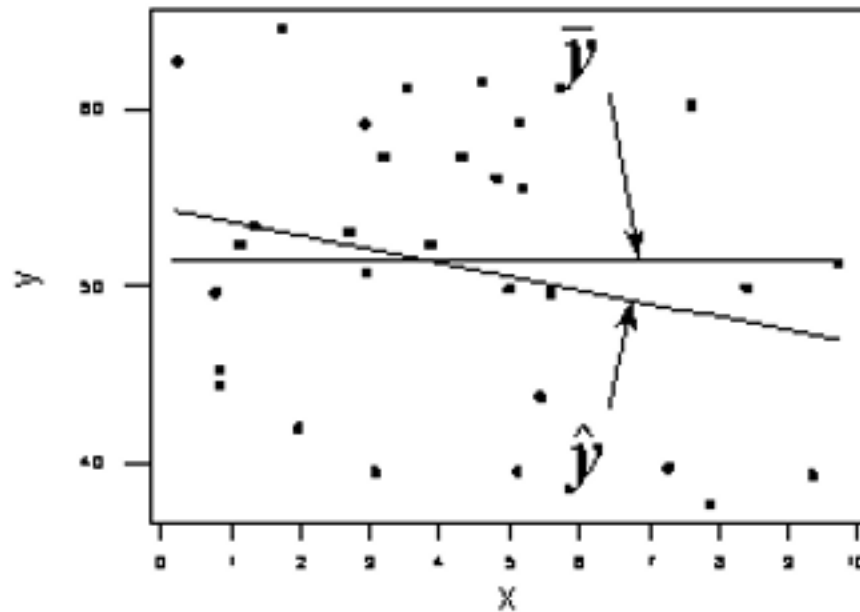
x	y	$\hat{y} = (41/42)x - (5/21)$	Squared error from line $(y - \hat{y})^2$	Squared error from mean $(\hat{y} - \bar{y})^2$
-2	-3	-2.1905	0.655328798	5.96
-1	-1	-1.2143	0.045918367	2.14
1	2	0.7381	1.592403628	0.24
4	3	3.66667	0.444444444	11.68
	$\bar{y} = 0.25$	Total	SSE = 2.738095238	SSR = 20.02

- SSE = 2.74, SSR = 20.02
- % of total variation not explained by the variation in x,  
 $SSE / SSR = 2.74/20.02 = 0.1369 = 13.69\%$
- % of total variation is explained by the variation in x,
- $r^2 = 1 - (SSE / SSR) = 1 - (2.74/20.02) = 0.8631 = 86.31\%$
- Percent is good. Therefore, most portion is explained



## Ex 2: coefficient of determination, $r^2$

- Relationship between the response  $y$  and the predictor  $x$  is very weak
- Lines are placed at the average response,  $\bar{y}$ , and estimated regression line,  $\hat{y}$
- Slope of the estimated regression line,  $\hat{y}$  is not very steep
- Suggesting that as the predictor  $x$  increases, there is not much of a change in the average response  $y$
- Data points are not close to the estimated regression line

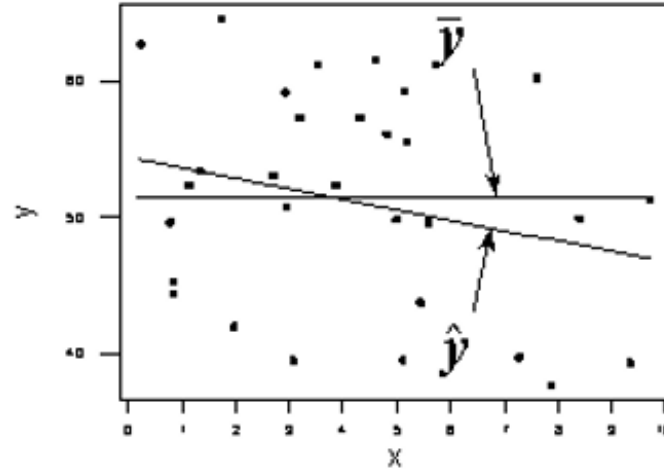


## Ex 2: coefficient of determination, $r^2$

$$SSR = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 = 119.1$$

$$SSE = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = 1708.5$$

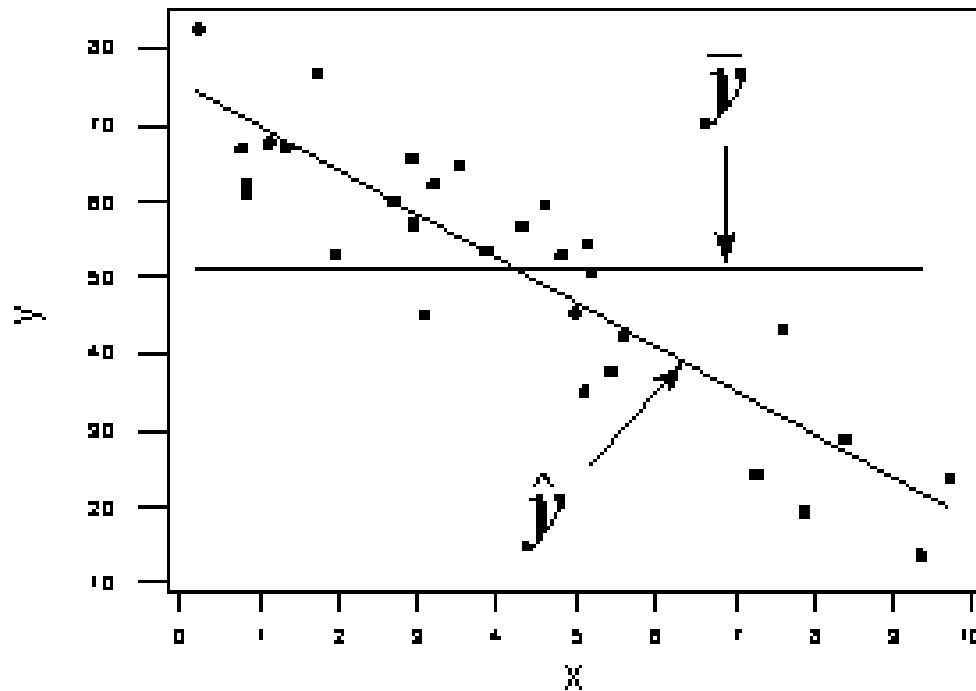
$$SSTO = \sum_{i=1}^n (y_i - \bar{y})^2 = 1827.6$$



- The sums of squares convey most of the information
- Represent most of the variation in the response  $y$  ( $SSTO = 1827.6$ ) is due to random variation ( $SSE = 1708.5$ ), not due to the regression of  $y$  on  $x$  ( $SSR = 119.1$ )
- And  $SSR/SSTO = 119.1/1827.6 = 0.065$
- $R^2 = 0.065$  or 6.5%

## Ex2: coefficient of determination, $r^2$

- Fairly convincing relationship between  $y$  and  $x$
- The slope of the estimated regression line is much steeper
- Suggesting that as the predictor  $x$  increases, there is a fairly substantial change (decrease) in the response  $y$
- Data points are close to estimated regression line

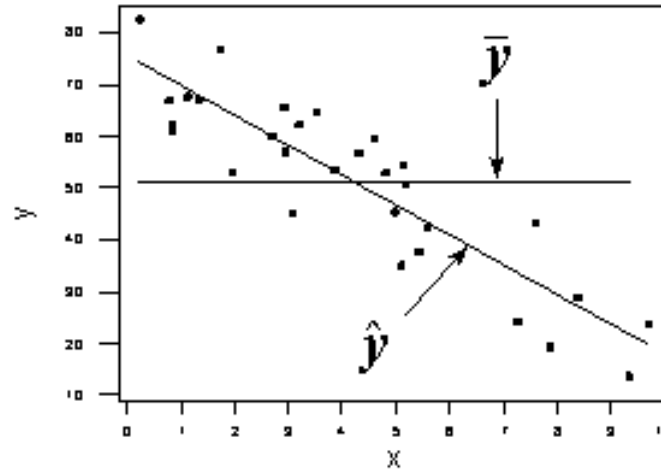


## Ex2: coefficient of determination, $r^2$

$$SSR = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 = 6679.3$$

$$SSE = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = 1708.5$$

$$SSTO = \sum_{i=1}^n (y_i - \bar{y})^2 = 8487.8$$



- Most of the variation in the response  $y$  ( $SSTO = 8487.8$ ) is
  - due to the regression of  $y$  on  $x$  ( $SSR = 6679.3$ )
  - not due to random error ( $SSE = 1708.5$ )
- And,  $SSR / SSTO = 6679.3 / 8487.8 = 0.799$
- $R^2 = 0.799 = 79.9 \%$

# Characteristics of coefficient of determination

$$r^2 = \frac{SSR}{SSTO} = 1 - \frac{SSE}{SSTO}$$

- The predictor  $x$  accounts for *all* of the variation in  $y$
- $0 \leq r^2 \leq 1$
- If  $r^2 = 1$ , all of the data points fall perfectly on the regression line
- If  $r^2 = 0$ , the estimated regression line is perfectly horizontal  
The predictor  $x$  accounts for *none* of the variation in  $y$ !
- " $r^2 \times 100$  percent of the variation in  $y$  is "explained by" the variation in predictor  $x$ "

# Which value is considered large for $r^2$ ?

- Depends on the application
- Social scientists who are often trying to learn something about the huge variation in human behavior find it very hard to get 25% or 30%
- For engineers, tend to study more exact systems 30% is unacceptable

# (Pearson) Correlation Coefficient $r$

- The correlation coefficient  $r$  is directly related to the coefficient of determination  $r^2$

$$r = \pm \sqrt{r^2}$$

- The sign of  $r$  depends on the sign of the estimated slope coefficient  $b_1$
- If  $b_1$  is negative, then  $r$  takes a negative sign
- If  $b_1$  is positive, then  $r$  takes a positive sign
- The estimated slope and the correlation coefficient,  $r$  share the same sign
- $r^2$  is always a number between 0 and 1, the correlation coefficient  $r$  is always a number between -1 and 1

# Alternative method for computation of r

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}}$$

- It is unitless
- Therefore correlation coefficients can be calculated on different data sets with different units
- Ex: x is height in inches and weight is in pounds



# One more method for computation of $r$

$$r = \frac{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}}{\sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}} \times b_1$$

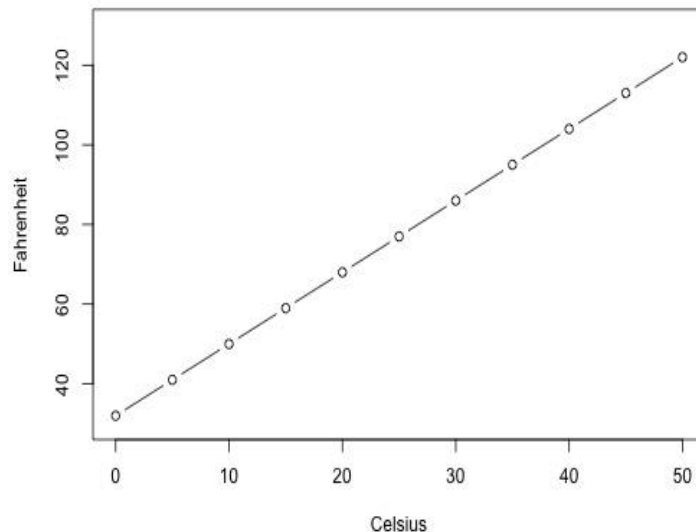
- The estimated slope  $b_1$  of the regression line and the correlation coefficient  $r$  always share the same sign
- If the estimated slope  $b_1$  of the regression line is 0, then the correlation coefficient  $r$  must also be 0
- If  $r = -1$ , then there is a perfect negative linear relationship between  $x$  and  $y$
- The closer  $r$  is to -1, the stronger the negative linear relationship
- If  $r = 1$ , then there is a perfect positive linear relationship between  $x$  and  $y$ .
- If  $r = 0$ , then there is no linear relationship between  $x$  and  $y$
- The closer  $r$  is to 0, the weaker the linear relationship

# Example: skin cancer

- Correlation between skin cancer mortality and latitude,  $r = -0.825$
- The relationship between mortality and latitude is quite strong (value is pretty close to -1)
- The relationship is negative
- As the latitude increases, the skin cancer mortality rate decreases (linearly)

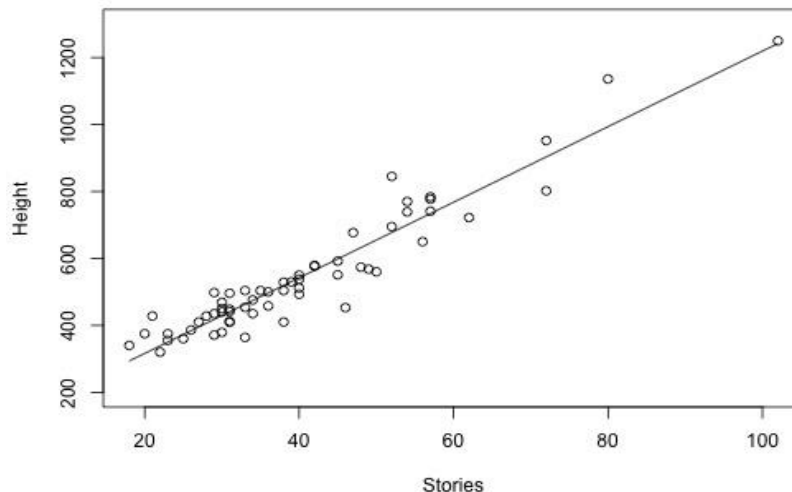
# Example 1: r-square and r

- How strong is the linear relationship between temperatures in Celsius and temperatures in Fahrenheit?
- For estimated regression equation,  $r^2 = 100\%$  and  $r = 1.000$
- There is a perfect linear relationship between temperature in degrees Celsius and temperature in degrees Fahrenheit
- $r^2$  tells us that 100% of the variation in temperatures in Fahrenheit is explained by the temperature in Celsius



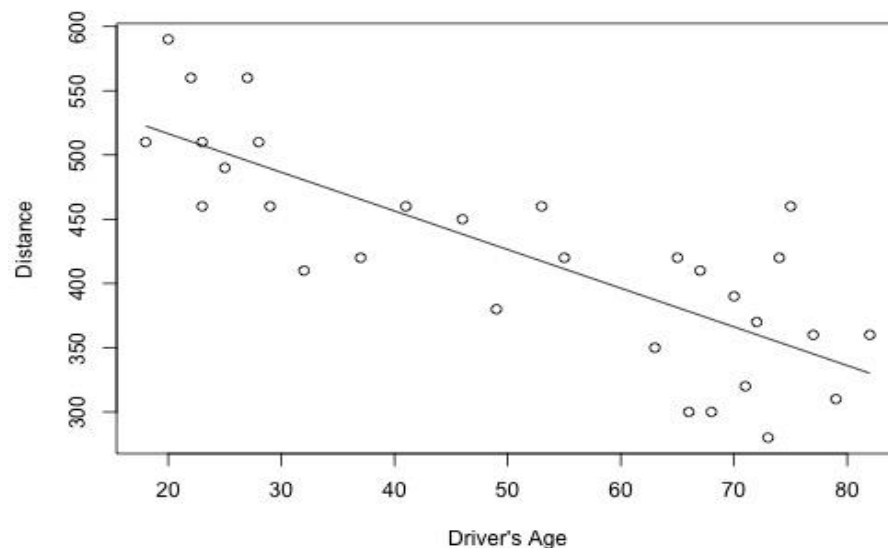
# Example 2 $r$ -square and $r$

- How strong is the linear relationship between the number of stories a building has and its height?
- As the number of stories increases, the height would increase, but not perfectly
- $r^2 = 90.4\%$  and  $r = 0.951$
- The positive sign of  $r$  tells us that the relationship is positive
- Because  $r$  is close to 1, it tells us that the linear relationship is very strong, but not perfect.
- The  $r^2$  value tells us that 90.4% of the variation in the height of the building is explained by the number of stories in the building



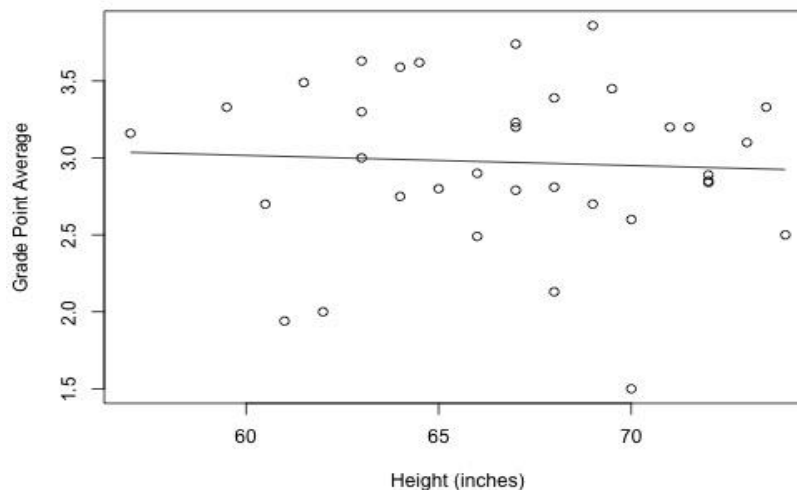
# Example 3: r-square and r

- How strong is the linear relationship between the age of a driver and the distance the driver can see?
- Probably the relationship is negative — as age increases, the distance decreases
- Statistical software reports that  $r^2 = 64.2\%$  and  $r = -0.801$
- Because  $r$  is fairly close to -1, it tells us that the linear relationship is fairly strong, but not perfect.
- The  $r^2$  value tells us that 64.2% of the variation in the seeing distance is reduced by taking into account the age of the driver



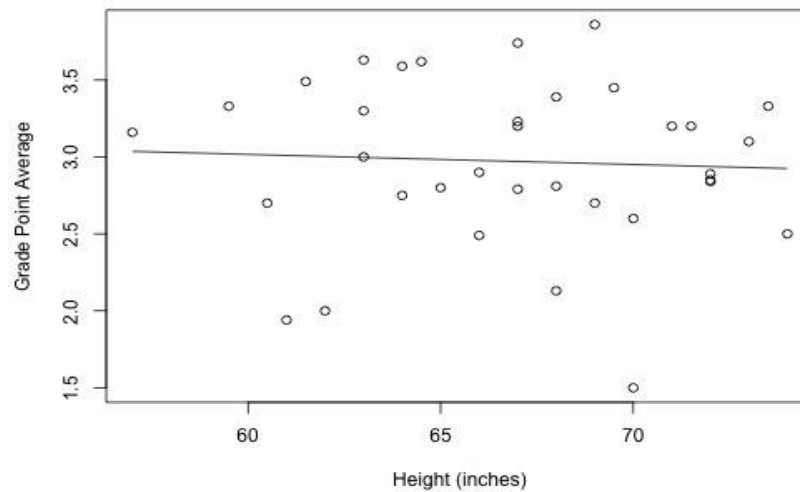
# Example 4

- How strong is the linear relationship between the height of a student and his or her grade point average?
- Data were collected on a random sample of  $n = 35$  students in a statistics course at Penn State University
- Statistical software reports that  $r^2 = 0.3\%$  and  $r = -0.053$



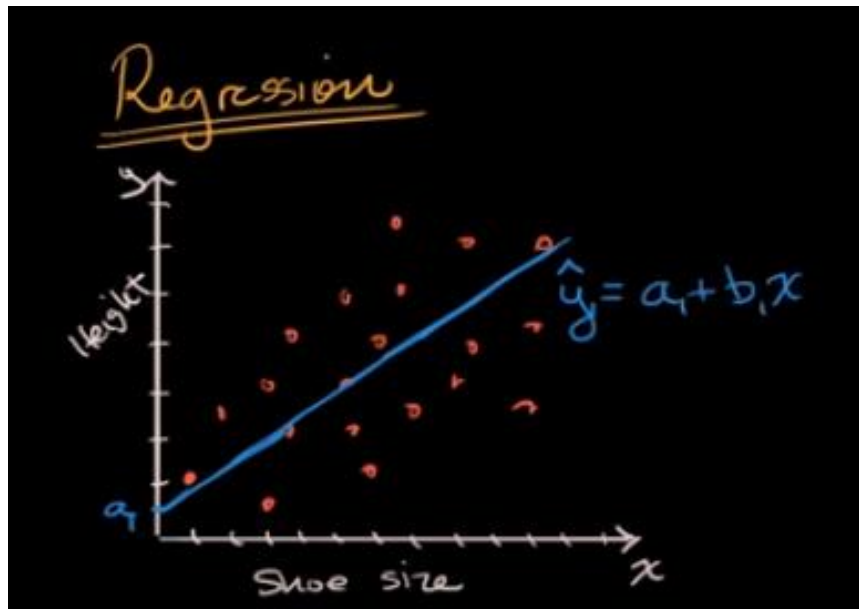
# Example 4

- Since  $r$  is quite close to 0, there is next to no linear relationship between height and grade point average
- The  $r^2$  value tells us that only 0.3% of the variation in the grade point averages of the students in the sample can be explained by their height.
- need to identify another more important variable, such as number of hours studied, if predicting a student's grade point average is important to us.



# Inference about slope

- Regression line for 20 samples

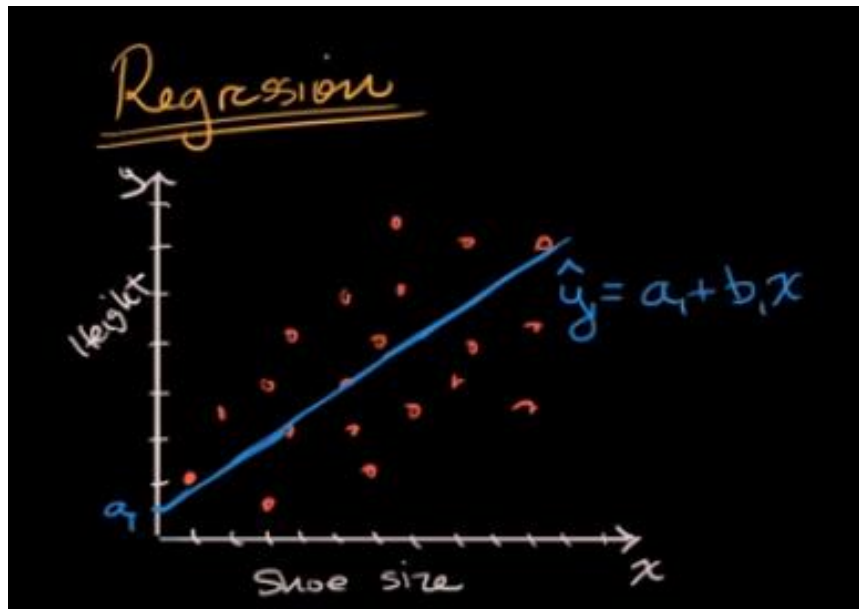


20 samples



# Inference about slope

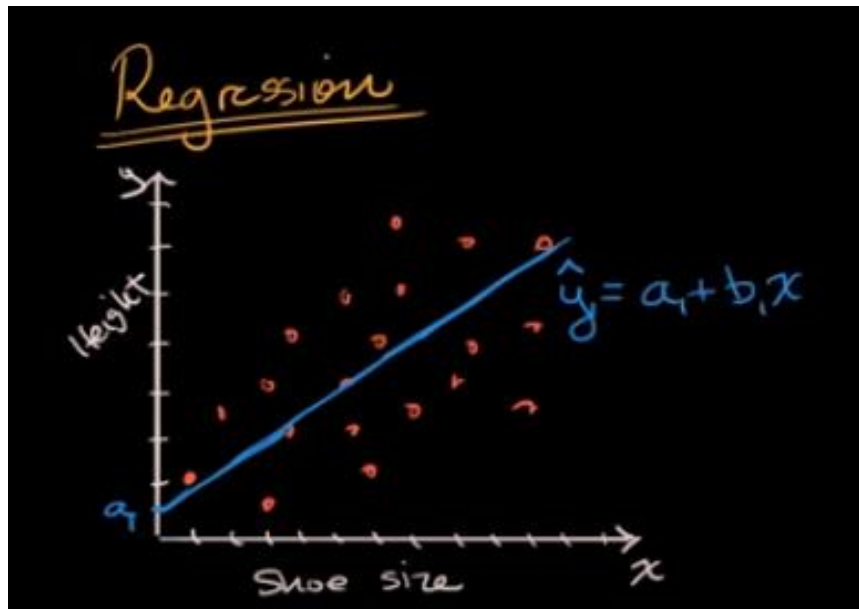
- Regression line for 20 samples
- After adding 20 more samples regression line changes



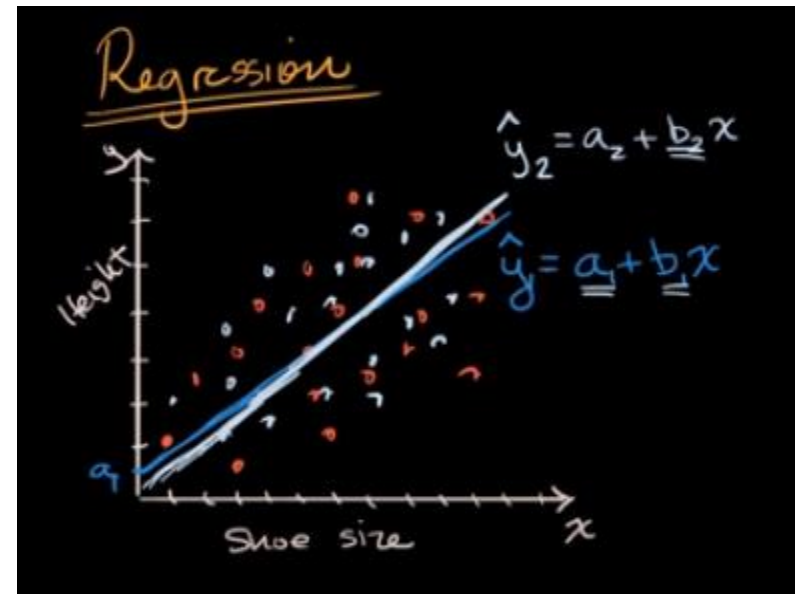
20 samples

# Inference about slope

- Regression line for 20 samples
- After adding 20 more samples regression line changes



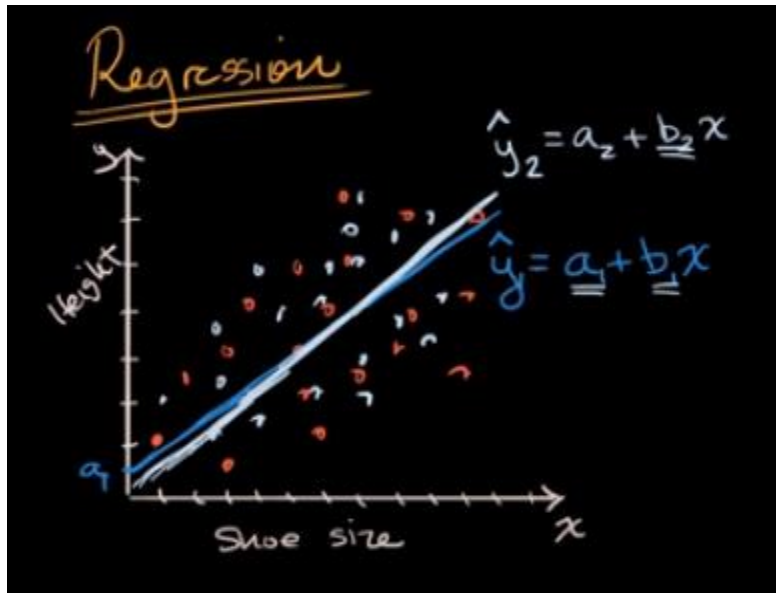
20 samples



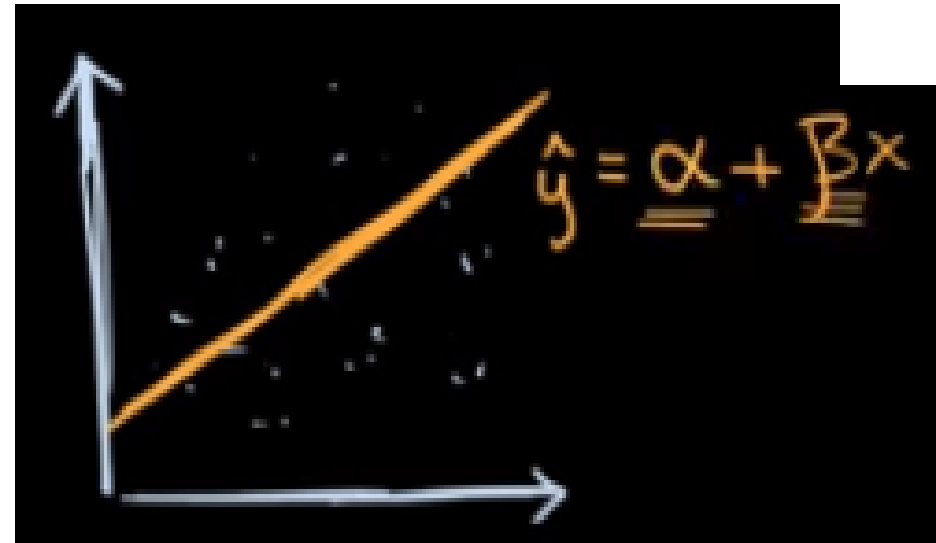
40 samples

# Inference about slope

- Actual values of slope and intercept are alpha and beta
- Require confidence level for the estimated values



inferences based on actual samples



Actual values

# Confidence Interval

- Create a confidence interval in order to get the variations from true parameters
- confidence interval,  $C = b_1 \pm t \text{SSE}_b$
- $b_1$  is slope and  $\text{SSE}_b$  is standard error
- t-value decides the confidence interval
- Or determine t value for the given confidence interval

# t-test

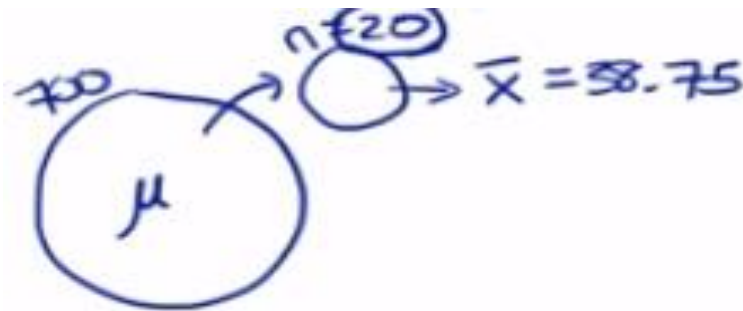
- Population (entire world, complete information) is given
- Take n samples from it and calculate sample mean,  $\bar{x}$  and sample standard deviation, s
- Confidence interval will be

$$\bar{x} \pm t^* \frac{s}{\sqrt{n}}$$

- For other set of n samples the value of confidence remains the same
- If C = 95%, then
- 95% of the time above interval will contain true mean
- This is called t-statistics

# Example: t-test

- Reena wanted to estimate age of the faculty at her university
- She collects data of 20 of the approximately 700 faculty
- The data was skewed to the right with a sample mean of  $\bar{x} = 38.75$ .
- She can use this data to make a confidence interval to estimate mean age of faculty members at her university
- Build a confidence interval to carry out inference on a mean



$$\bar{x} \pm t^* \frac{s}{\sqrt{n}}$$

# Conditions for inference on a mean

- For the accuracy of methods three conditions should be met
  - Otherwise the calculations and conclusions may not be correct
- 1. Random**
    - A random sample or randomized experiment should be used to obtain the data
  - 2. Normal**
    - The sampling distribution of the sample mean needs to be approximately normal
    - This is true if our parent population is normal
    - or if sample size is reasonably large ( $n \geq 30$ )
  - 3. Independent**
    - Individual observations need to be independent
    - If sampling is done without replacement, then sample size shouldn't be more than 10% of the population

# 1. The random condition

- Random samples give us unbiased data from a population
- Ex: a bag of ping pong balls individually numbered from 0 to 30 and population mean of the bag is 15
- Take random samples of balls from the bag and calculate the mean from each sample
- Some samples would have a mean higher than 15 and some would be lower
- On average, the mean of each sample will be 15 which holds true as long as samples are random
- Biased samples can lead to inaccurate results



## 2. The normal condition

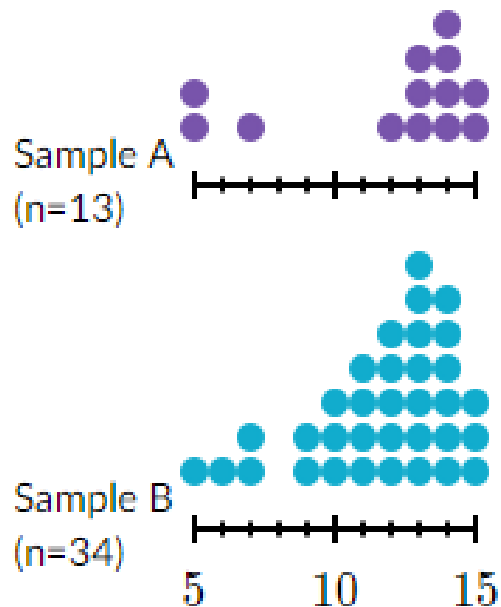
- The sampling distribution of  $\bar{x}$ , (a sample mean) should be approximately normal
- The shape of the sampling distribution of  $\bar{x}$ , mostly depends on the shape of the parent population and the sample size,  $n$
- If parent population has normal distribution and sample size,  $n > 30$
- then  $\bar{x}$  is normally distributed regardless of the shape of the sample data or its population

## 2. The normal condition

- When sample size is smaller than 30, plot data to check distribution
  - If the data shows skew or outliers then parent population may not be approximately normal
  - As long as the sample data looks roughly symmetric with no outliers, the sampling distribution of  $\bar{x}$  will be approximately normal

# Example

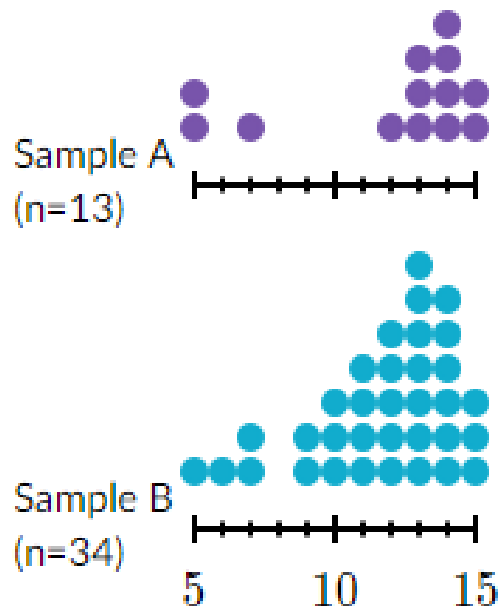
- Two different samples are drawn from two different populations



- Which sample satisfies the normal condition for constructing a t interval?
- Sample A fails the normal condition because of the small sample size and low outliers

# Example

- Two different samples are drawn from two different populations



- Sample B has a large enough sample size ( $n=34$ ) it passes the normal condition
- Sample B does not have normal distribution
- It will be approximately normal due to the sample size ( $n>30$ )

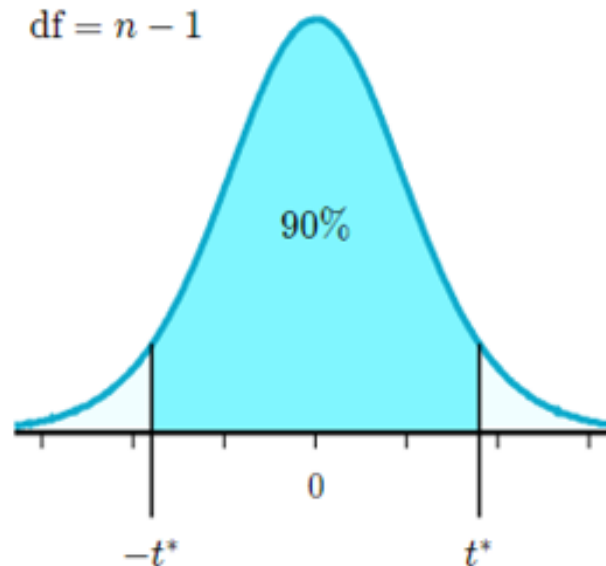
# 3. The Independence condition

- Individual observations should be independent
- Individual observations aren't technically independent since removing each observation changes the population
- However the 10% condition says that
- If 10% percent or less of the population is sampled then individual observations can be treated as independent
- This is because removing an observation doesn't change the population
- Ex: Sample size is  $n=30$

There should to be at least  $N=300$  members in the population for the sample to meet the independence condition

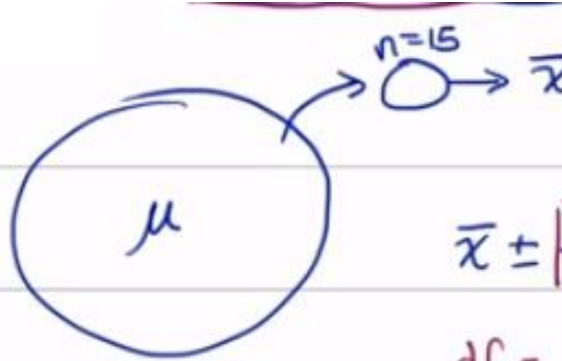
# Critical Value of t ( $t^*$ )

- There is a different t-distribution for each sample size,  $n$
- Use t-distribution with degrees of freedom,  $df = n - 1$
- The critical value  $t^*$  for 90% confidence is the distance that tells us  
how far we must go above and below the center of a t-distribution to obtain an area of 90%



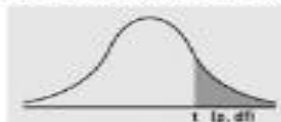
# Critical value, $t^*$

- What is the critical value,  $t^*$  to achieve 98% confidence interval for a mean from a sample size of  $n=15$  observations?
- Assume that all the three conditions are satisfied
- degree of freedom,  $df = n-1 = 14$
- t-table is available for different types of distributions


$$\bar{x} \pm \boxed{t^*} \cdot \frac{s}{\sqrt{n}}$$
$$df = n - 1$$
$$df = 14$$

# t-table

Numbers in each row of the table are values on a  $t$ -distribution with ( $df$ ) degrees of freedom for selected right-tail (greater-than) probabilities ( $p$ ).



df/p	0.40	0.25	0.10	0.05	0.025	0.01	0.005	0.0005
1	0.324920	1.000000	3.077684	6.313752	12.70620	31.82052	63.65674	636.6192
2	0.288675	0.816497	1.885618	2.919986	4.30265	6.96456	9.92484	31.5991
3	0.276671	0.764892	1.637744	2.353363	3.18245	4.54070	5.84091	12.9240
4	0.270722	0.740697	1.533206	2.131847	2.77645	3.74695	4.60409	8.6103
5	0.267181	0.726687	1.475884	2.015048	2.57058	3.36493	4.03214	6.8688
6	0.264835	0.717558	1.439756	1.943180	2.44691	3.14267	3.70743	5.9588
7	0.263167	0.711142	1.414924	1.894579	2.36462	2.99795	3.49948	5.4079
8	0.261921	0.706387	1.396815	1.859548	2.30600	2.89646	3.35539	5.0413
9	0.260955	0.702722	1.383029	1.833113	2.26216	2.82144	3.24984	4.7809
10	0.260185	0.699812	1.372184	1.812461	2.22814	2.76377	3.16927	4.5869
11	0.259556	0.697445	1.363430	1.795885	2.20099	2.71808	3.10581	4.4370
12	0.259033	0.695483	1.356217	1.782288	2.17881	2.68100	3.05454	4.3178
13	0.258591	0.693829	1.350171	1.770933	2.16037	2.65031	3.01228	4.2208
14	0.258213	0.692417	1.345030	1.761310	2.14479	2.62449	2.97684	4.1405
15	0.257885	0.691197	1.340606	1.753050	2.13145	2.60248	2.94671	4.0728
16	0.257599	0.690132	1.336757	1.745884	2.11991	2.58349	2.92078	4.0150
17	0.257347	0.689195	1.333379	1.739607	2.10982	2.56693	2.89823	3.9651
18	0.257123	0.688364	1.330391	1.734064	2.10092	2.55238	2.87844	3.9216
19	0.256923	0.687621	1.327728	1.729133	2.09302	2.53948	2.86093	3.8834
20	0.256743	0.686954	1.325341	1.724718	2.08596	2.52798	2.84534	3.8495
21	0.256580	0.686352	1.323188	1.720743	2.07961	2.51765	2.83136	3.8193
22	0.256432	0.685805	1.321237	1.717144	2.07387	2.50832	2.81876	3.7921
23	0.256297	0.685306	1.319460	1.713872	2.06866	2.49987	2.80734	3.7676
24	0.256173	0.684850	1.317836	1.710882	2.06390	2.49216	2.79694	3.7454
25	0.256060	0.684430	1.316345	1.708141	2.05954	2.48511	2.78744	3.7251
26	0.255955	0.684043	1.314972	1.705618	2.05553	2.47863	2.77871	3.7066
27	0.255858	0.683685	1.313703	1.703288	2.05183	2.47266	2.77068	3.6896
28	0.255768	0.683353	1.312527	1.701131	2.04841	2.46714	2.76326	3.6739
29	0.255684	0.683044	1.311434	1.699127	2.04523	2.46202	2.75639	3.6594
30	0.255605	0.682756	1.310415	1.697261	2.04227	2.45726	2.75000	3.6460
z	0.253347	0.674490	1.281552	1.844854	1.95996	2.32635	2.57583	3.2905
CI	—	—	80%	90%	95%	98%	99%	99.9%



# Critical value, $t^*$

- What is the critical value,  $t^*$  to achieve 98% confidence interval for a mean from a sample size of  $n=15$  observations?
- Therefore,  $t^*=2.624$

df	Tail probability $p$											
	.25	.20	.15	.10	.05	.025	.02	.01	.005	.0025	.001	.0005
1	1.000	1.376	1.963	3.078	6.314	12.71	15.89	31.82	63.66	127.3	318.3	636.6
2	.816	1.061	1.386	1.886	2.920	4.303	4.849	6.965	9.925	14.09	22.33	31.60
3	.765	.978	1.250	1.638	2.353	3.182	3.482	4.541	5.841	7.453	10.21	12.92
4	.741	.941	1.190	1.533	2.132	2.776	2.999	3.747	4.604	5.598	7.173	8.610
5	.727	.920	1.156	1.476	2.015	2.571	2.757	3.365	4.032	4.773	5.893	6.869
6	.718	.906	1.134	1.440	1.943	2.447	2.612	3.143	3.707	4.317	5.208	5.959
7	.711	.896	1.119	1.415	1.895	2.365	2.517	2.998	3.499	4.029	4.785	5.408
8	.706	.889	1.108	1.397	1.860	2.306	2.449	2.896	3.355	3.833	4.501	5.041
9	.703	.883	1.100	1.383	1.833	2.262	2.398	2.821	3.250	3.690	4.297	4.781
10	.700	.879	1.093	1.372	1.812	2.228	2.359	2.764	3.169	3.581	4.144	4.587
11	.697	.876	1.088	1.363	1.796	2.201	2.328	2.718	3.106	3.497	4.025	4.437
12	.695	.873	1.083	1.356	1.782	2.179	2.303	2.681	3.055	3.428	3.930	4.318
13	.694	.870	1.079	1.350	1.771	2.160	2.282	2.650	3.012	3.372	3.852	4.221
14	.692	.868	1.076	1.345	1.761	2.145	2.264	2.624	2.977	3.326	3.787	4.140
15	.691	.866	1.074	1.341	1.753	2.131	2.249	2.602	2.947	3.286	3.733	4.073
26	.684	.856	1.058	1.315	1.706	2.056	2.162	2.479	2.779	3.067	3.435	3.707
27	.684	.855	1.057	1.314	1.703	2.052	2.158	2.473	2.771	3.057	3.421	3.690
28	.683	.855	1.056	1.313	1.701	2.048	2.154	2.467	2.763	3.047	3.408	3.674
29	.683	.854	1.055	1.311	1.699	2.045	2.150	2.462	2.756	3.038	3.396	3.659
30	.683	.854	1.055	1.310	1.697	2.042	2.147	2.457	2.750	3.030	3.385	3.646
40	.681	.851	1.050	1.303	1.684	2.021	2.123	2.423	2.704	2.971	3.307	3.551
50	.679	.849	1.047	1.299	1.676	2.009	2.109	2.403	2.678	2.937	3.261	3.496
60	.679	.848	1.045	1.296	1.671	2.000	2.099	2.390	2.660	2.915	3.232	3.460
80	.678	.846	1.043	1.292	1.664	1.990	2.088	2.374	2.639	2.887	3.195	3.416
100	.677	.845	1.042	1.290	1.660	1.984	2.081	2.364	2.626	2.871	3.174	3.390
1000	.675	.842	1.037	1.282	1.646	1.962	2.056	2.330	2.581	2.813	3.098	3.300
$\infty$	.674	.841	1.036	1.282	1.645	1.960	2.054	2.326	2.576	2.807	3.091	3.291

Copy link



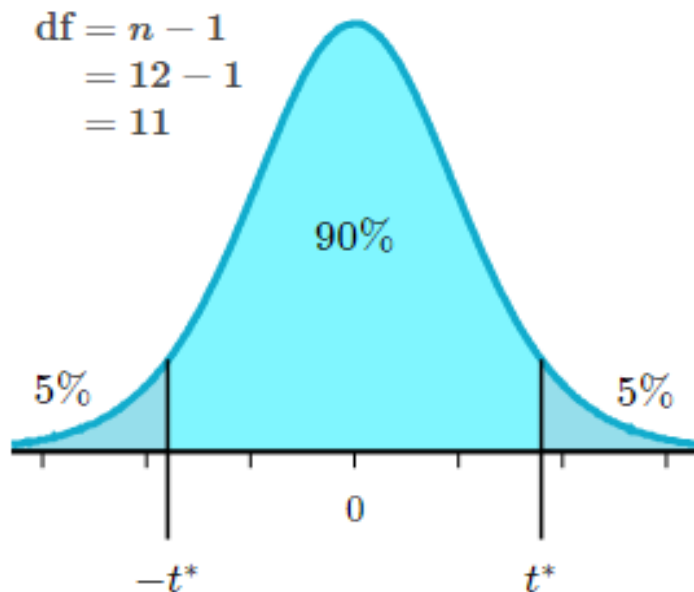
C = 50% 60% 70% 80% 90% 95% 96% 98% 99% 99.5% 99.8% 99.9%

# Ex: Critical Value of $t$ ( $=t^*$ )

- Ruchi took a random sample of  $n=12$  octopus and tracked them to calculate their mean lifespan
- These life spans are roughly symmetric with a mean of  $\bar{x} = 4$  years and standard deviation of  $\sigma = 0.5$  years
- She wants to use this data to construct a  $t$ -interval for the mean lifespan with 90% confidence

# Strategy to find $t^*$

- Determine area remaining in the tails in a t-distribution with  $df = 12 - 1$
- Remaining area,  $p = 100\% - 90\% = 10\%$
- $10\%/2 = 5\%$  per tail



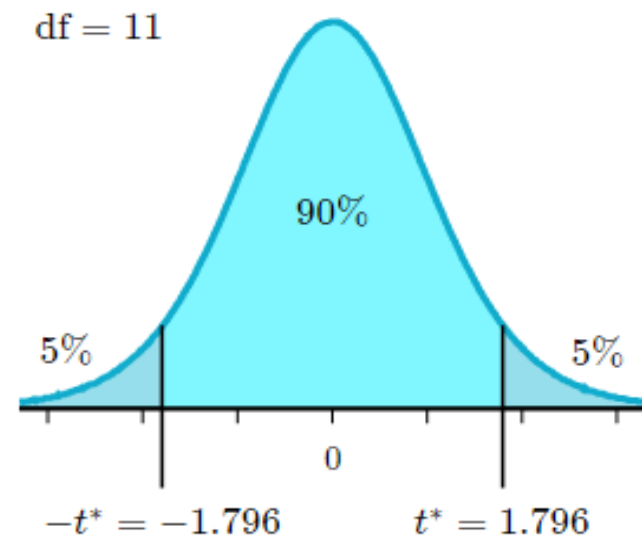
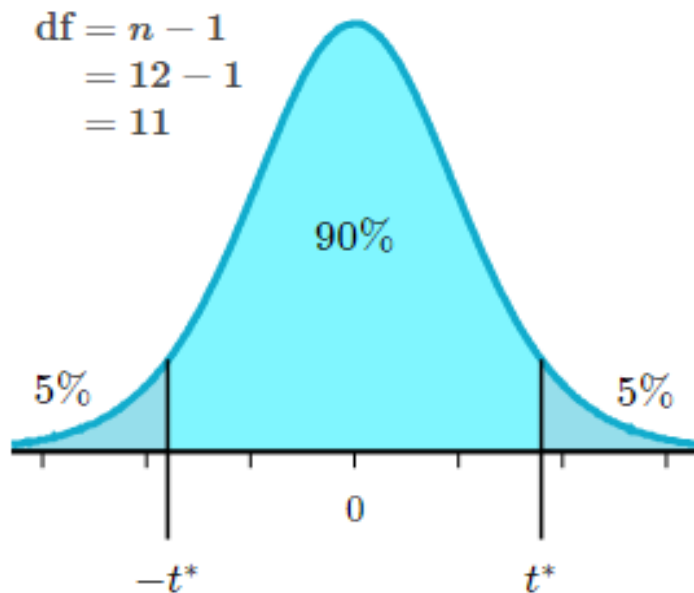
t-distribution

$p$ (1-tail)	0.1	0.05	0.025
$p$ (2-tail)	0.2	0.1	0.05
df			
10	1.372	1.812	2.228
11	1.363	1.796	2.201
12	1.356	1.782	2.179

t- table

# Strategy to find $t^*$

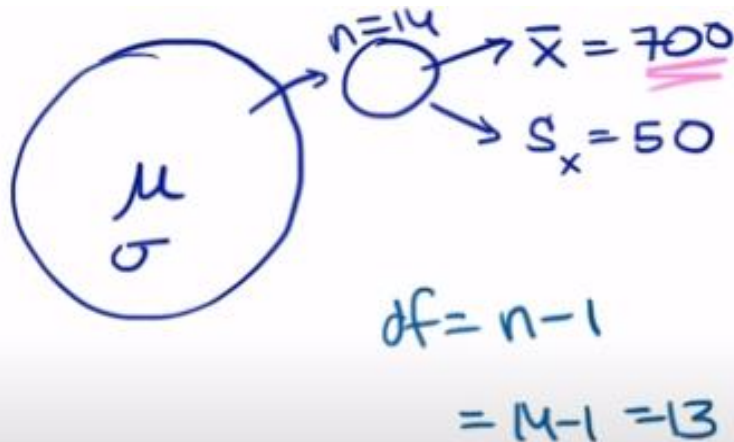
- Determine area remaining in the tails in a t-distribution with  $df = 12 - 1$
- Remaining area =  $100\% - 90\% = 10\%$
- $10\%/2 = 5\%$  per tail



So  $t^* = 1.796$

# Example: C-interval for a mean

- A nutritionist wants to estimate the average caloric content of 14 pizzas
- Sample data is roughly symmetric with a mean of 700 calories and a standard deviation of 50 calories
- Determine 95% confidence interval for the mean of caloric content of pizzas



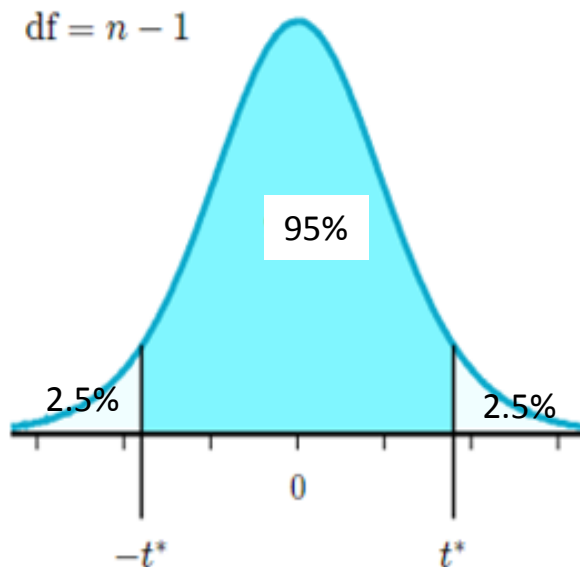
Confidence interval,  
 $C = \bar{x} \pm (t^*s)/\text{sqrt}(n)$

# Example: C-interval for a mean

- Confidence level is 95%
- Remaining area=  $100\% - 95\% = 5\%$
- $5\%/2 = 2.5\%$  per tail
- $df = n-1 = 14-1 = 13$
- Critical value,  $t^* = 2.160$

# Example: C-interval for a mean

- Confidence level is 95%
- Remaining area= 100% - 95% = 5%
- 5%/2 = 2.5% per tail
- df = 13
- Critical value,  $t^* = 2.160$



df	Tail probability p											
	.25	.20	.15	.10	.05	.025	.02	.01	.005	.0025	.001	.0005
1	1.000	1.376	1.963	3.078	6.314	12.71	15.89	31.82	63.66	127.3	318.3	636.6
2	.816	1.061	1.386	1.886	2.920	4.303	4.849	6.965	9.925	14.09	22.33	31.82
3	.765	.978	1.250	1.638	2.353	3.182	3.482	4.541	5.841	7.453	10.21	12.94
4	.741	.941	1.190	1.533	2.132	2.776	2.999	3.747	4.604	5.598	7.173	8.610
5	.727	.920	1.156	1.476	2.015	2.571	2.757	3.365	4.032	4.773	5.893	6.860
6	.718	.906	1.134	1.440	1.943	2.447	2.612	3.143	3.707	4.317	5.208	5.959
7	.711	.896	1.119	1.415	1.895	2.365	2.517	2.998	3.499	4.029	4.785	5.401
8	.706	.889	1.108	1.397	1.860	2.306	2.449	2.896	3.355	3.833	4.501	5.041
9	.703	.883	1.100	1.383	1.833	2.262	2.398	2.821	3.250	3.690	4.297	4.781
10	.700	.879	1.093	1.372	1.812	2.228	2.359	2.764	3.169	3.581	4.144	4.581
11	.697	.876	1.088	1.363	1.796	2.201	2.328	2.718	3.106	3.497	4.025	4.437
12	.695	.873	1.083	1.356	1.782	2.179	2.303	2.681	3.055	3.428	3.930	4.315
13	.694	.870	1.079	1.350	1.771	2.160	2.282	2.650	3.012	3.372	3.852	4.221
14	.692	.868	1.076	1.345	1.761	2.145	2.264	2.624	2.977	3.326	3.787	4.140
15	.691	.866	1.074	1.341	1.753	2.131	2.249	2.602	2.947	3.286	3.733	4.076
16	.690	.865	1.071	1.337	1.746	2.120	2.235	2.583	2.921	3.252	3.686	4.019
17	.689	.863	1.069	1.333	1.740	2.110	2.224	2.567	2.898	3.222	3.646	3.970
18	.688	.862	1.067	1.330	1.734	2.101	2.214	2.552	2.878	3.197	3.611	3.929
19	.688	.861	1.066	1.328	1.729	2.093	2.205	2.539	2.861	3.174	3.579	3.889
20	.687	.860	1.064	1.325	1.725	2.086	2.197	2.528	2.845	3.153	3.552	3.858

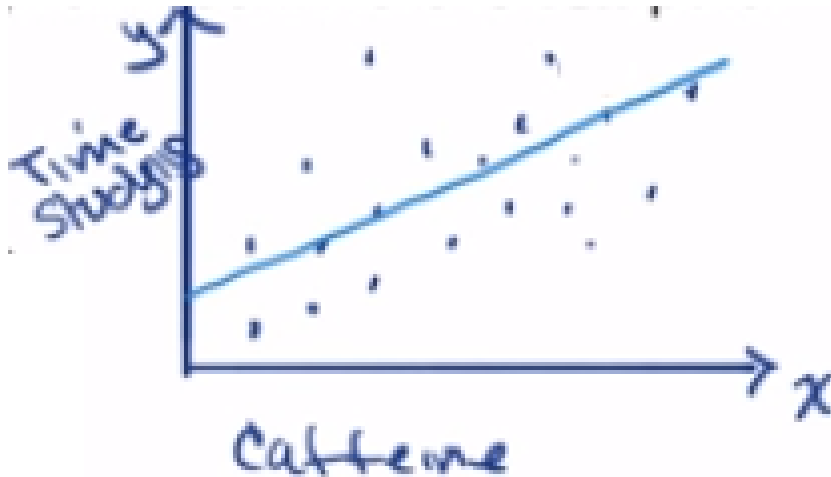
# Example: C-interval for a mean

- Confidence interval,
- $C = \bar{x} \pm (t*s)/\sqrt{n}$   
 $= 700 \pm (2.160*50)/\sqrt{14}$   
 $= 700 \pm 28.9$   
 $= 671.1 \text{ to } 728.9$
- Confidence interval for the mean of a regression line is 671.1 to 728.9



# Example: C-interval for slope

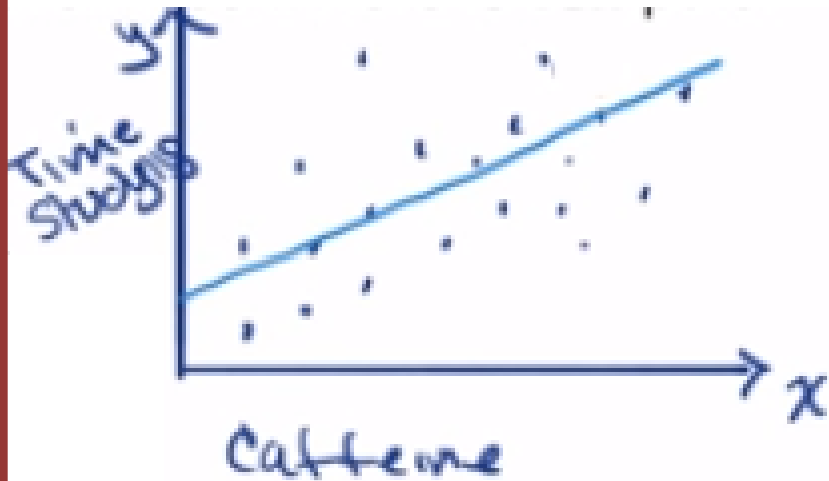
- Anushka randomly selects 20 students at her school and records their caffeine intake (mg) and the amount of time spent studying in a given week
- What is the confidence interval for the slope of least squares regression line?



Given:

1. Plot has 20 points
2. y-intercept,  $b=2.544$  and slope is 0.164 for least square regression line

# Example: C-interval for slope



## Formula

Here's the formula for a  $t$  interval estimating slope:

$$(\text{statistic}) \pm \left( \begin{array}{c} \text{critical} \\ \text{value} \end{array} \right) \left( \begin{array}{c} \text{standard deviation} \\ \text{of statistic} \end{array} \right)$$

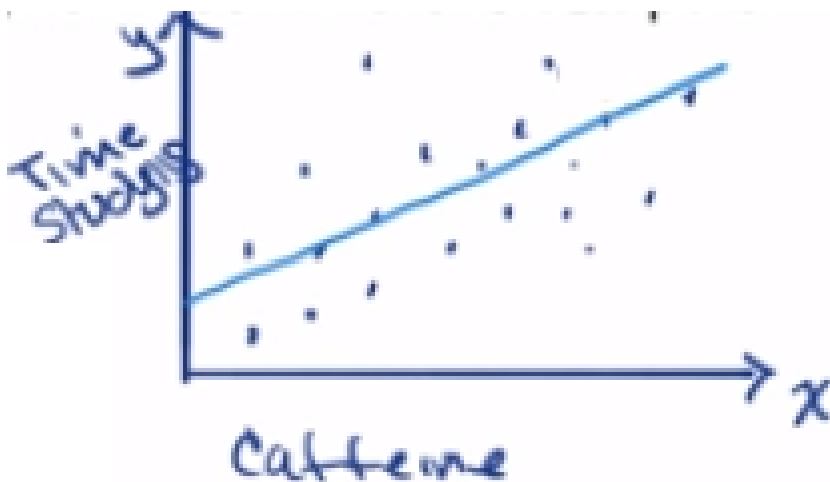
$$b \pm t_{n-2}^* (SE_b)$$

Given:

1. Plot has 20 points
2. y-intercept,  $b=2.544$  and slope is 0.164 for least square regression line

# Example: C-interval for slope

- R-sq represents how much of the variance in y-variable is explainable by x-variable
- S is standard deviation of error (residual)
- SE Coef shows Standard error for coefficients
- $C = \text{slope} \pm t^* \text{ SE Coef}$   
 $= 0.164 \pm t^* 0.057$



Computer output

Predictor	Coef	SE Coef	T	P
Constant	2.544	0.134	18.955	0.000
Caffeine	0.164	0.057	2.862	0.010
S = 1.532 R-sq = 60.0%				

Assume that all conditions for inference have been met.

# Example: C-interval for slope

- SE Coef for caffeine = 0.057
- Number of data points = 20  
df = 20-2 =18
- For confidence level of 95%

$$\begin{aligned} C &= \text{slope} \pm t^* \text{ SE coef} \\ &= 0.164 \pm t^* 0.057 \\ &= 0.164 \pm 2.101 \times 0.057 \\ &= 0.044 \text{ to } 0.284 \end{aligned}$$

[illegible]

# t statistic for slope of regression line

- Cerro Negro is an active volcano located in Nicaragua. Data of the date and volume (in thousand cubic meters) of Cerro Negro's 23 recent eruptions is available
- Regression output on the sample data (years are counted as number of years since 1850) is given below

Regression: volume vs. year

Predictor	Coef	SE Coef
Constant	-3984	13,390
Year	1198	131

S = 22,744   R-sq = 79.84%

Assume that all conditions for inference have been met.

# t statistic for slope of regression line

- For a t test about slope,
- $t = b_1 / SE_b$
- Where  $b_1$  is slope and  $SE_b$  is Standard Error for slope
- $SE_b = \sqrt{ \Sigma(y_i - \hat{y}_i)^2 / (n - 2) } / \sqrt{ \Sigma(x - \bar{x})^2 }$
- $t = 1198 / 131 = 9.145$

Computer output

Regression: volume vs. year

Predictor	Coef	SE Coef
Constant	-3984	13390
Year	1198	131
S = 22744    R-sq = 79.84%		

# Hypothesis

- A claim that is to be tested
- Take samples, analyse them and then test for the claim
- Null hypothesis, ( $H_0$ ) : things are happening as expected
- Alternative hypothesis, ( $H_a$ ): things are not happening as expected

# Example: Hypothesis

- It is believed that a candy machine makes chocolate bars that are on average 5 gm. A worker claims that the machine after maintenance no longer makes 5 gm bars. Write  $H_o$  and  $H_a$ .
- $H_o : \mu = 5\text{gm}$
- $H_a : \mu \neq 5\text{gm}$
- $H_o$  and  $H_a$  are mathematically opposite
- Possible outcome of this test is
  - Reject null hypothesis
  - Fail to reject null hypothesis
- Test statistic: Use sample data to test null hypothesis



# Example: Hypothesis

- Sample of 50 bars are given
- One person checked average weight of 50 bars and gets 5.12 gm on Monday
  - May not reject  $H_0$
  - Machine can be assumed to be normal
- Another person goes on Tuesday and gets average weight of 5.72 gm
  - May rethink before rejecting null hypothesis
  - May say that we have checked only 50 bars. May try more bars
- Third person on Thursday gets 7.23 gm
  - definitely reject  $H_0$
  - Can conclude that machine is not functioning
- These cases are subjective
  - Each person may have different opinion
  - Even second case can be rejected

# Example: Hypothesis

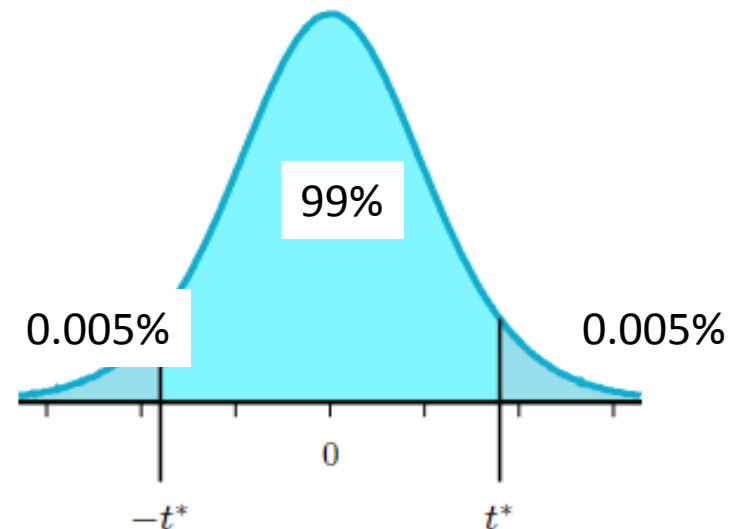
- Concrete way is to be used to decide when to reject a hypothesis
- 'C' represents level of confidence in taking decision
- C can be 95%, 99%, 50%
- 95% and 99% (high values) show that  $H_0$  can be rejected
- If C is 50% does not represent confidence
- Level of significance,  $\alpha = 1 - C = 0.05$

# Example: Hypothesis

- A company has stated that their straw machine makes straws that are 4 mm diameter. A worker believes the machine no longer makes such straws
- Perform a hypothesis test with 99% confidence
- $H_o : \mu = 4 \text{ mm}$
- $H_a : \mu \neq 4 \text{ mm}$
- $n=100$
- $C = 99\%$
- $\alpha = 1-C = 0.01$  (for 2 tails)

# Example: Hypothesis

- A company has stated that their straw machine makes straws that are 4 mm diameter. A worker believes the machine no longer makes such straws
- Perform a hypothesis test with 99% confidence
- $H_o : \mu = 4 \text{ mm}$
- $H_a : \mu \neq 4 \text{ mm}$
- $n=100$
- $C = 99\%$
- $\alpha = 1-C = 0.01$  (for 2 tails)

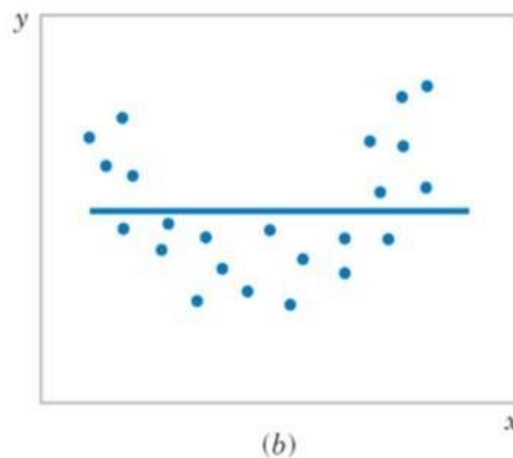
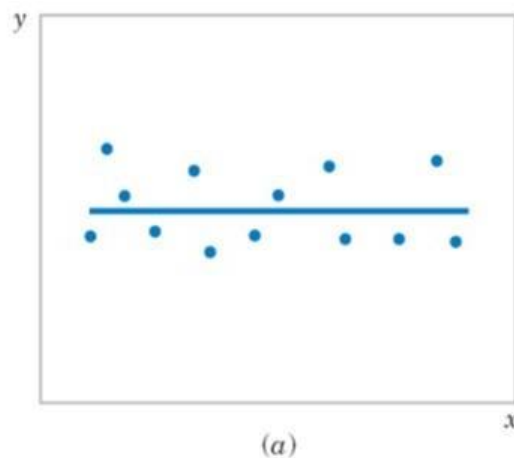


# Hypothesis for slope of Linear Regression

- Regression line for the entire population is
$$\hat{y} = \beta_0 + \beta_1 x$$
- Null Hypothesis represents zero relationship between independent and dependent variable
- Or slope of true population regression line is 0
- $H_0: \beta_1 = 0$
- Alternative hypothesis, slope of true relationship is not 0
- $H_a: \beta_1 \neq 0$

# Hypothesis for slope of Linear Regression

Figure The hypothesis  $H_0: \beta_1 = 0$  is not rejected.



# Hypothesis for slope of Linear Regression

Figure The hypothesis  $H_0: \beta_1 = 0$  is not rejected.

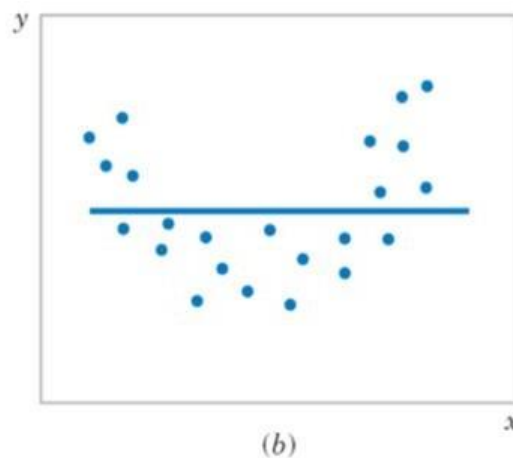
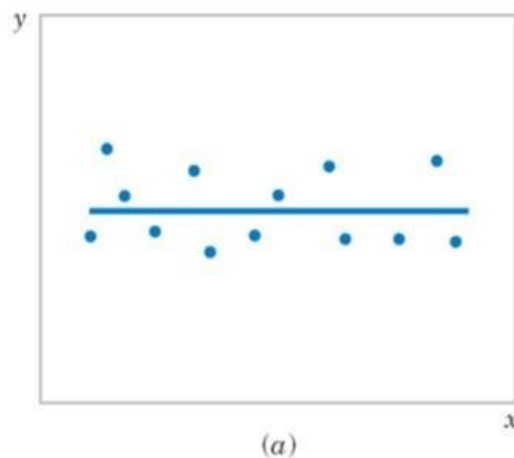
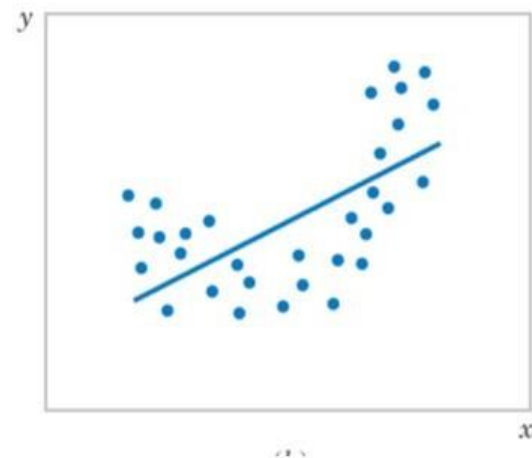


Figure The hypothesis  $H_0: \beta_1 = 0$  is rejected.



# Hypothesis for slope of Linear Regression

- $H_0: \beta_1 = 0$  versus  $H_a: \beta_1 \neq 0$
- Methods to test are
  1. The t-test for the slope
  2. Analysis of variance (ANOVA) or F-test

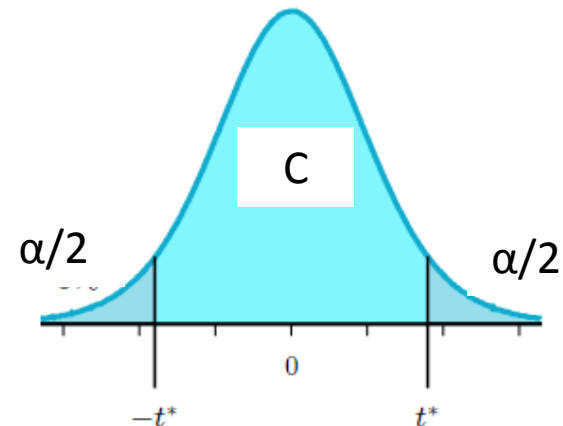
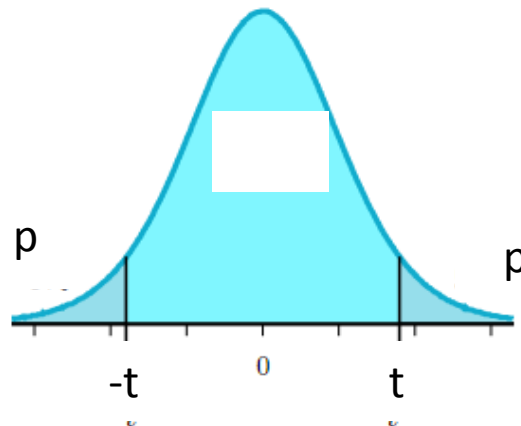


# Steps to Test Hypothesis


- Given significance level,  $\alpha$
- The value of  $\alpha$  corresponds to  $t^*$
- Compute  $t$ -value for the given samples

$$t = b_1 / SE_b$$

- The  $P$ -value is determined by referring to a  $t$ -distribution table with  $n-2$  degrees of freedom

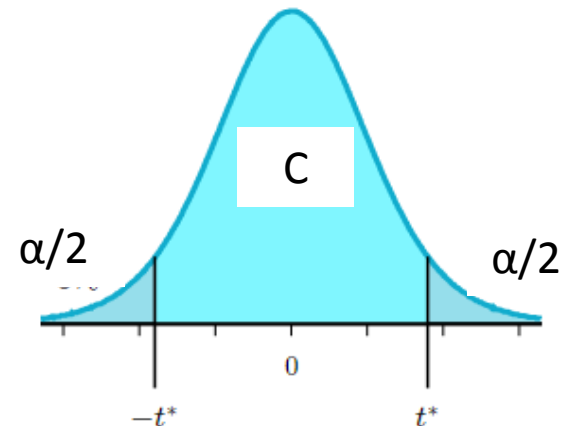
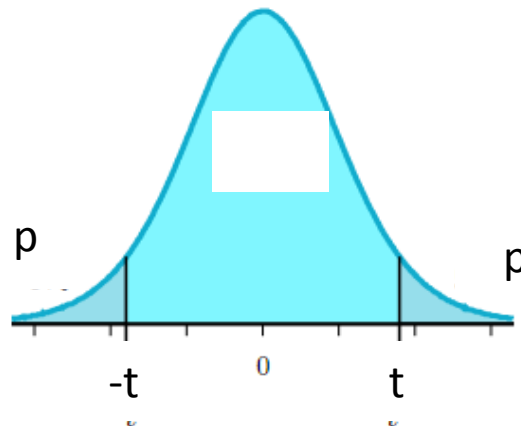


# T-distribution Table for t and p values

df	Tail probability p 											
	.25	.20	.15	.10	.05	.025	.02	.01	.005	.0025	.001	.0005
1	1.000	1.376	1.963	3.078	6.314	12.71	15.89	31.82	63.66	127.3	318.3	636
2	.816	1.061	1.386	1.886	2.920	4.303	4.849	6.965	9.925	14.09	22.33	31.0
3	.765	.978	1.250	1.638	2.353	3.182	3.482	4.541	5.841	7.453	10.21	12.9
4	.741	.941	1.190	1.533	2.132	2.776	2.999	3.747	4.604	5.598	7.173	8.6
5	.727	.920	1.156	1.476	2.015	2.571	2.757	3.365	4.032	4.773	5.893	6.86
6	.718	.906	1.134	1.440	1.943	2.447	2.612	3.143	3.707	4.317	5.208	5.95
7	.711	.896	1.119	1.415	1.895	2.365	2.517	2.998	3.499	4.029	4.785	5.40
8	.706	.889	1.108	1.397	1.860	2.306	2.449	2.896	3.355	3.833	4.501	5.04
9	.703	.883	1.100	1.383	1.833	2.262	2.398	2.821	3.250	3.690	4.297	4.78
10	.700	.879	1.093	1.372	1.812	2.228	2.359	2.764	3.169	3.581	4.144	4.58
11	.697	.876	1.088	1.363	1.796	2.201	2.328	2.718	3.106	3.497	4.025	4.43
12	.695	.873	1.083	1.356	1.782	2.179	2.303	2.681	3.055	3.428	3.930	4.3
13	.694	.870	1.079	1.350	1.771	2.160	2.282	2.650	3.012	3.372	3.852	4.22
14	.692	.868	1.076	1.345	1.761	2.145	2.264	2.624	2.977	3.326	3.787	4.14
15	.691	.866	1.074	1.341	1.753	2.131	2.249	2.602	2.947	3.286	3.733	4.07
16	.690	.865	1.071	1.337	1.746	2.120	2.235	2.583	2.921	3.252	3.686	4.0
17	.689	.863	1.069	1.333	1.740	2.110	2.224	2.567	2.898	3.222	3.646	3.96
18	.688	.862	1.067	1.330	1.734	2.101	2.214	2.552	2.878	3.197	3.611	3.92
19	.688	.861	1.066	1.328	1.729	2.093	2.205	2.539	2.861	3.174	3.579	3.88
20	.687	.860	1.064	1.325	1.725	2.086	2.197	2.528	2.845	3.153	3.552	3.84

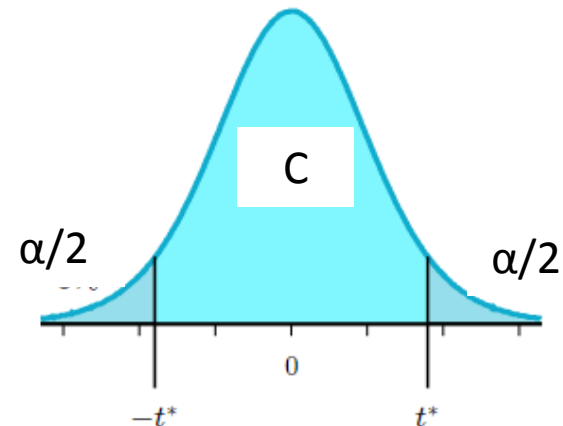
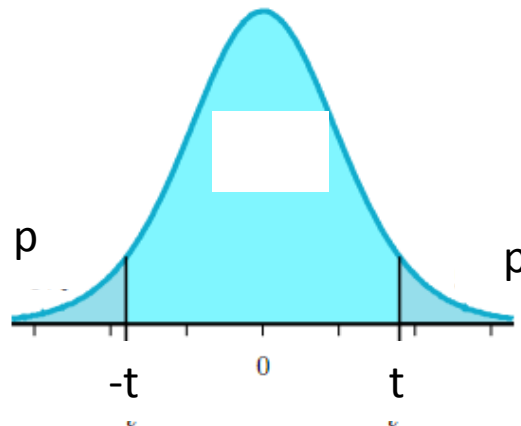
# Steps to Test Hypothesis

- Use t-value to determine p-value
- The P-value is the probability corresponding to t-value
- If the  $P$ -value is smaller than the significance level  $\alpha$ ,
  - reject the null hypothesis in favor of the alternative
  - there is sufficient evidence at the  $\alpha$  level to conclude that  
there is a linear relationship in the population  
between the predictor  $x$  and response  $y$



# Steps to Test Hypothesis

- Use t-value to determine p-value
- The P-value is the probability corresponding to t-value
- If the  $P$ -value is larger than the significance level  $\alpha$ 
  - Do not reject the null hypothesis
  - Enough evidence at the  $\alpha$  level is not available to conclude that  
there is a linear relationship in the population  
between the predictor  $x$  and response  $y$



# Ex 1: Hypothesis Test

- The response variable  $y$  is the mortality rate (number of deaths per 10 million people) due to a disease
- The predictor variable  $x$  is the latitude (degrees North) at the center of each state in the United States

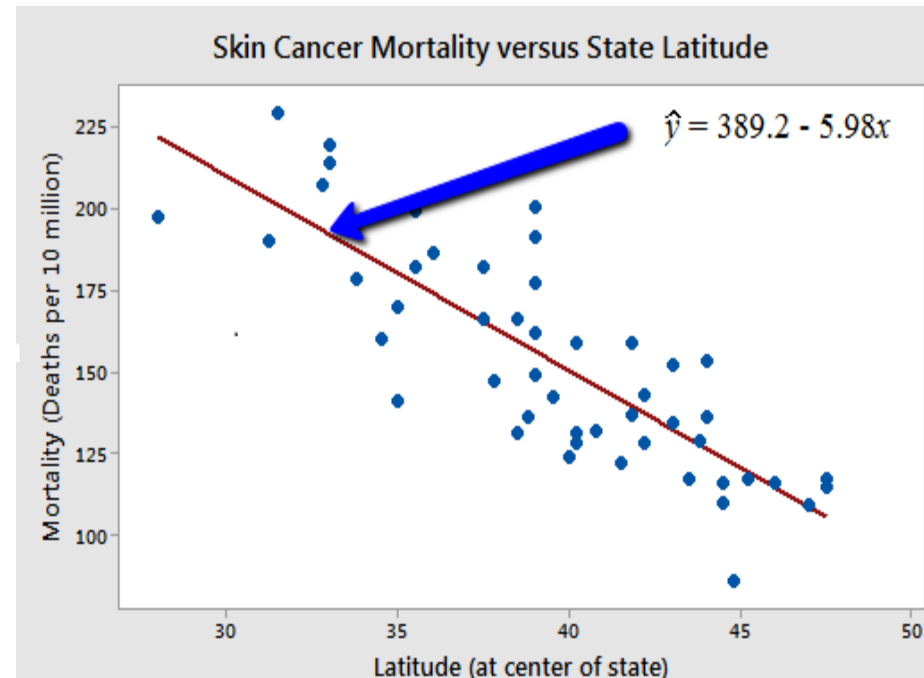
# Ex 1: Hypothesis Test

- The response variable  $y$  is the mortality rate (number of deaths per 10 million people) due to a disease
- The predictor variable  $x$  is the latitude (degrees North) at the center of each state in the United States

subset of data

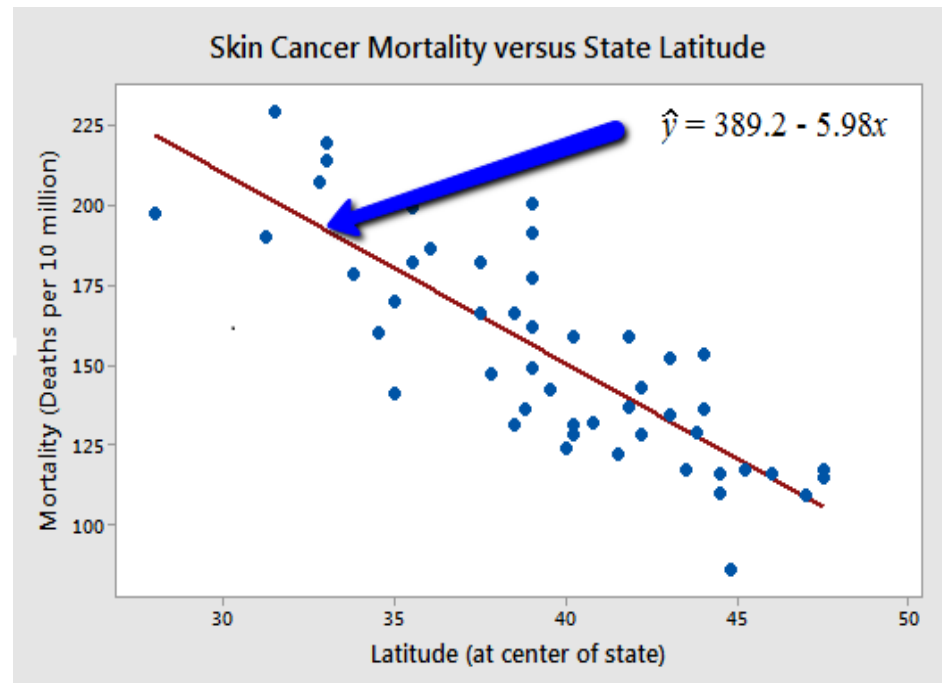
#	State	Latitude	Mortality
1	Alabama	33.0	219
2	Arizona	34.5	160
3	Arkansas	35.0	170
4	California	37.5	182
5	Colorado	39.0	149
---	---	---	---
49	Wyoming	43.0	134

Regression Line



# Ex 1: Hypothesis Test

- There is a relationship between state latitude and skin cancer mortality for 49 samples
- Because the estimated slope of the line,  $b_1$ , is -5.98, not 0
- There may not be the same relationship in the entire population of latitudes and skin cancer mortality rates
- Determine whether the population slope  $\beta_1$  is likely to be 0



# Ex 1: Hypothesis Test

- Regression analysis output for skin cancer mortality and latitude
- Estimated slope coefficient  $b_1$ , under the column, **Coef**, is -**5.9776**
- Estimated standard error of  $b_1$ , under **SE Coef** for "standard error of the coefficient," is **0.5984**.

The regression equation is Mort = 389 - 5.98 Lat

Predictor	Coef	SE Coef	T	P
Constant	389.19	23.81	16.34	0.000
Lat	-5.9776	0.5984	-9.99	0.000

S = 19.12

R-Sq = 68.0%

R-Sq(adj) = 67.3%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	36464	36464	99.80	0.000
Residual Error	47	17173	365		
Total	48	53637			

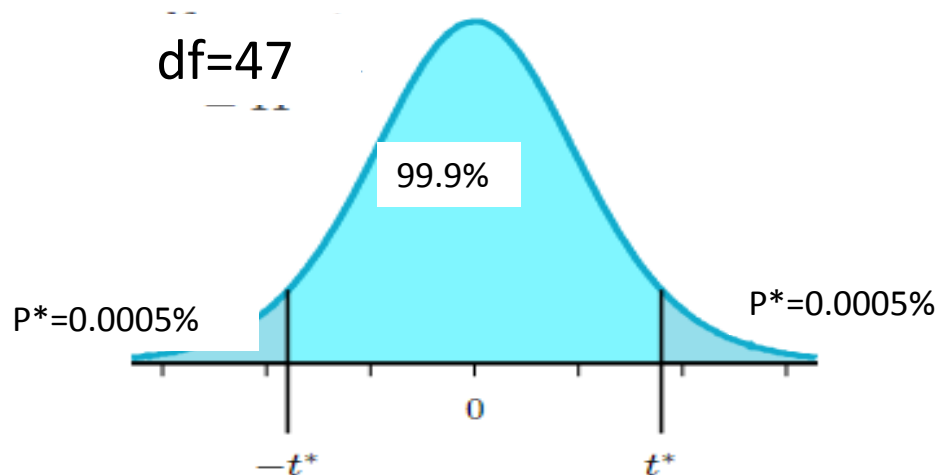


# Ex 1: Hypothesis Test

- Test statistics,  $t = b_1 / SE_{b1}$
- $t = -5.9776 / 0.5984 = -9.99$
- Generally confidence level,  $C = 99.9\%$
- Therefore, significance value,  
 $\alpha = 1 - C = 1 - 0.999 = 0.001$

# Ex 1: Hypothesis Test

- Degree of freedom for p-value =  $n - 2$   
 $= 49 - 2 = 47$
- $\alpha$ -value represents probability that a  $t$ -value with 47 degrees of freedom falls in upper tail



# Ex 1: t-test for slope and hypothesis

- P-value corresponding to t-value is less than  $\alpha$ -value (= 0.001)
- Therefore reject the null hypothesis and conclude that  $\beta_1 \neq 0$

The regression equation is Mort = 389 - 5.98 Lat

Predictor	Coef	SE Coef	T	P
Constant	389.19	23.81	16.34	0.000
Lat	-5.9776	0.5984	-9.99	0.000

S = 19.12      R-Sq = 68.0%      R-Sq(adj) = 67.3%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	36464	36464	99.80	0.000
Residual Error	47	17173	365		
Total	48	53637			

# Ex 2: Hypothesis Test

- The local utility company surveys 101 randomly selected customers. For each survey participant, the company collects the following: annual electric bill (in dollars) and home size (in square feet). Use a 0.05 level of significance.
- Output from a regression analysis appears below.

Computer output

Regression equation:				
Annual bill = 0.55 * Home size + 15				
Predictor	Coef	SE Coef	T	P
Constant	15	3	5.0	0.00
Home size	0.55	0.24	2.29	0.01

Is there a significant linear relationship between annual bill and home size?

# Ex 2: Hypothesis Test

- Slope ( $b_1$ ) and the standard error (SE) from the regression output  
 $b_1 = 0.55$        $SE = 0.24$
- Degrees of freedom (df) for the number of observations in the sample (n)  
 $DF = n - 2 = 101 - 2 = 99$
- For t-statistic,  $t = b_1/SE = 0.55/0.24 = 2.29$

## Regression equation:

$$\text{Annual bill} = 0.55 * \text{Home size} + 15$$

Predictor	Coef	SE Coef	T	P
Constant	15	3	5.0	0.00
Home size	0.55	0.24	2.29	0.01

# Ex 2: Hypothesis Test

- Based on the t statistic and the degrees of freedom, determine the P-value
- The P-value is the probability that a t-value for 99 degrees of freedom is more than 2.29 or less than -2.29
- Using t Distribution table,
  - $P(t > 2.29) = 0.0121$
  - $P(t < -2.29) = 0.0121$
- Therefore, the P-value is  $0.0121 + 0.0121$  or 0.0242

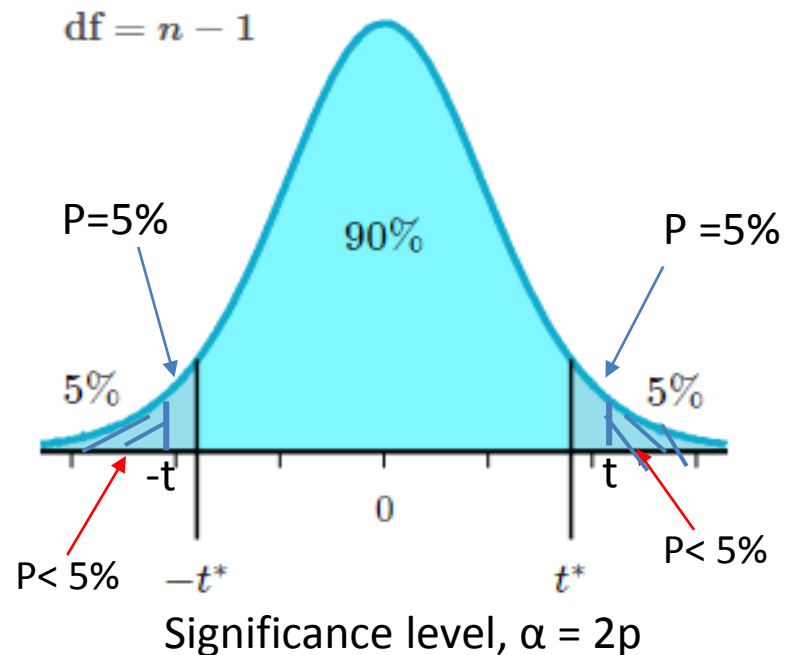
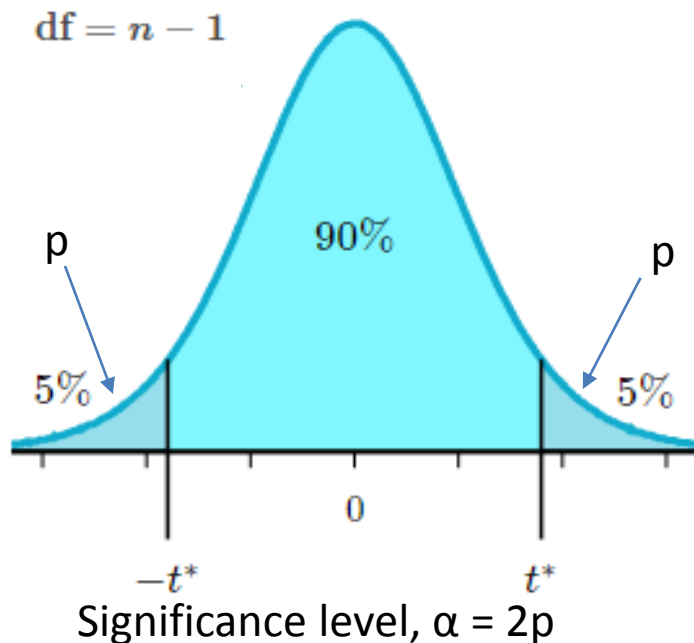
## 4. Interpret results

Since the P-value (0.0242) is less than the significance level (0.05)

Do not accept the null hypothesis

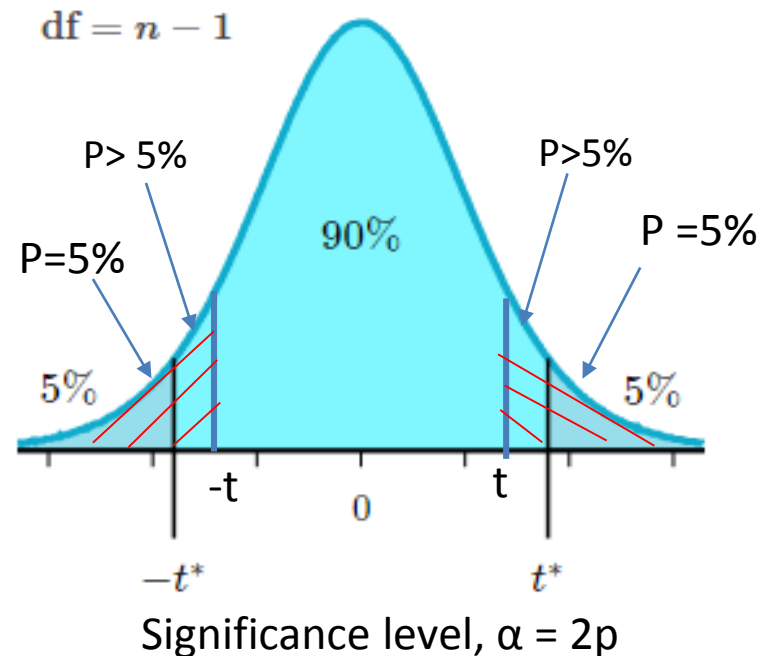
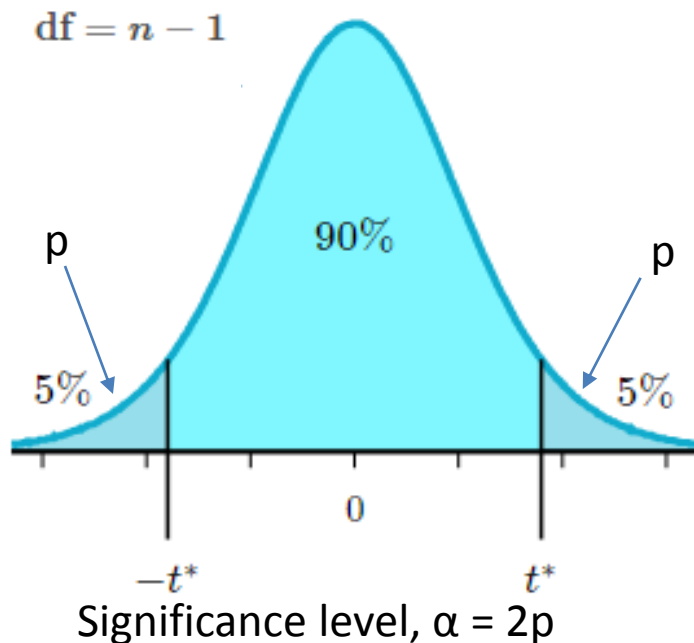
# Significance of p-value and $t^*$

- Given  $C = 90\%$  or  $0.9$
- Significance value,  $\alpha = 1 - C = 2p$
- Therefore  $p = 5\%$  or  $0.005$
- For the given value of  $C$  (or  $t^*$ ), p-value corresponding to denotes the probability that slope is not within  $-t^*$  to  $t^*$
- If 'p-value corresponding to calculated t-value' is less then 'p-value corresponding to  $t^*$ ' then
  - Slope does not fall in confidence interval for slope 0
  - Therefore null hypothesis can be relected



# Significance of p-value and $t^*$

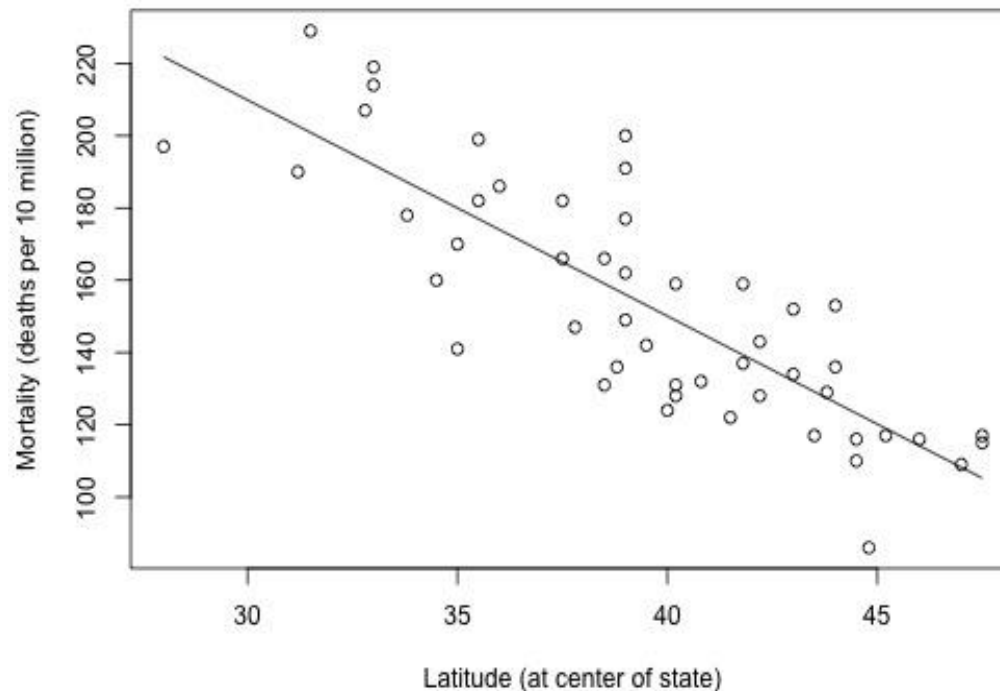
- If 'p-value corresponding to calculated t-value' is more than 'p-value corresponding to  $t^*$ ' then
  - Slope falls in confidence interval for slope 0
  - Therefore null hypothesis can be accepted





# Ex 3: t-test for hypothesis

- Is there a (linear) relationship between skin cancer mortality and latitude?”
- Linear relationship looks fairly strong
- Estimated slope is negative, not equal to 0.



# Ex 3: t-test for hypothesis

- $H_0: \beta_1 = 0$  and  $H_a: \beta_1 \neq 0$
- Significance level is 0.001
- $P$ -value of the  $t$ -test for "Lat" is less than 0.001
- There is enough statistical evidence to conclude that the slope is not 0
- that is, that there is a linear relationship between skin cancer mortality and latitude

The regression equation is Mort = 389 - 5.98 Lat

Predictor	Coef	SE Coef	T	P
Constant	389.19	23.81	16.34	0.000
Lat	-5.9776	0.5984	-9.99	0.000

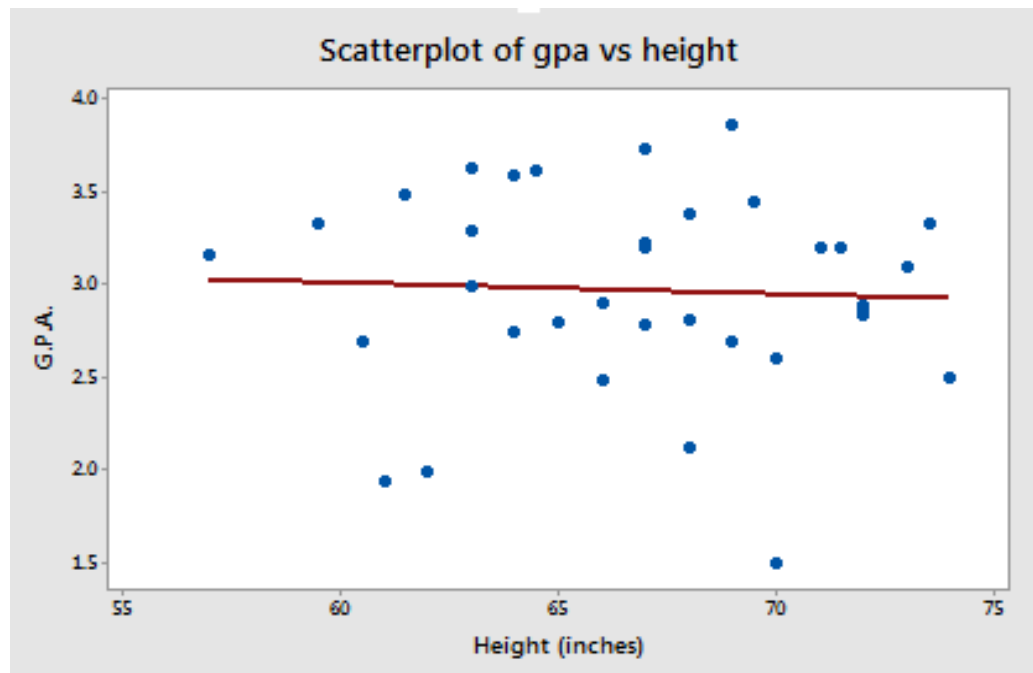
S = 19.12      R-Sq = 68.0%      R-Sq(adj) = 67.3%

## Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	36464	36464	99.80	0.000
Residual Error	47	17173	365		
Total	48	53637			

# Ex 3: t-test for hypothesis

- "Is there a (linear) relationship between height and grade point average?"
- There is almost no relationship
- The estimated slope is almost 0



# Ex 3: t-test for hypothesis

- Null hypothesis  $H_0: \beta_1 = 0$ ,  $H_a: \beta_1 \neq 0$
- $P$ -value of the  $t$ -test for "height" is 0.761
- Therefore, there is no linear relationship between height and grade point average

The regression equation is `gpa = 3.41 - 0.0066 height`

Predictor	Coef	SE Coef	T	P
Constant	3.410	1.435	2.38	0.023
height	-0.00656	0.02143	-0.31	0.761

`S = 0.5423`      `R-Sq = 0.38`      `R-Sq(adj) = 0.08`

## Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	0.0276	0.0276	0.09	0.761
Residual Error	33	9.7055	0.2941		
Total	34	9.7331			

statistical software output

# Analysis of Variance (ANOVA)

- Divide total variation in  $y$  ("total sum of squares") into two components:
  - due to the change in  $x$  ("regression sum of squares")
  - due to random error ("error sum of squares")

Where,

- $x_i$  and  $y_i$  are given data points
- $\bar{y}$  is mean value of  $y$
- $\hat{y}_i = b_0 - b_1x$ , for linear regression line
- $n$  is number of samples

$$\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

**SSTO**                      **SSR**                      **SSE**  
Total sum of squares      Regression sum of squares      Error sum of squares

$$SSTO = SSR + SSE$$

# Analysis of Variance (ANOVA)

- If the regression sum of squares is a "large" component of the total sum of squares  
it suggests that there *is* a linear association between the predictor  $x$  and the response  $y$

$$\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

**SSTO**                      **SSR**                      **SSE**  
Total sum of squares      Regression sum of squares      Error sum of squares

$$SSTO = SSR + SSE$$

# Analysis of Variance (ANOVA)

- $SSTO = SSR + SSE$
- Degrees of freedom is how many data points are independent
- The degrees of freedom associated with each of these sums of squares follow a similar decomposition
- That is  
 $df \text{ of } SSTO = df \text{ of } SSR + df \text{ of } SSE$

$(n - 1)$	=	$(1)$	+	$(n - 2)$
degrees of freedom associated with SSTO		degrees of freedom associated with SSR		degrees of freedom associated with SSE

# Parameters for ANOVA

Source of Variation	DF	SS	MS	F
Regression	1	$SSR = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$	$MSR = \frac{SSR}{1}$	$\frac{MSR}{MSE}$
Residual error	$n-2$	$SSE = \sum_{i=1}^n (y_i - \hat{y}_i)^2$	$MSE = \frac{SSE}{n-2}$	
Total	$n-1$	$SSTO = \sum_{i=1}^n (y_i - \bar{y})^2$		

- $F = [SSR/(\text{df of SSR})]/[SSE/\text{df of SSE}]$   
 $= [ssr/(1)][SSE/(n-2)]$
- Df of SSR is called numerator df (or  $d_1$ )
- Df of SSE is called denominator df (or  $d_2$ )
- F is often referred to as the analysis of variance F-test



# Example (revisited)

x	y	$\hat{y} = (41/42)x - (5/21)$	Squared error from line $(y - \hat{y})^2$	Squared error from mean $(\hat{y} - \bar{y})^2$
-2	-3	-2.1905	0.655328798	5.96
-1	-1	-1.2143	0.045918367	2.14
1	2	0.7381	1.592403628	0.24
4	3	3.66667	0.444444444	11.68
	$\bar{y} = 0.25$	Total	SSE = 2.738095238	SSR = 20.02

- $SSE = 2.74$ ,  $SSR = 20.02$
- % of total variation not explained by the variation in x,  
 $SSE / SSR = 2.74/20.02 = 0.1369 = 13.69\%$
- % of total variation is explained by the variation in x,
- $r^2 = 1 - (SSE / SSR) = 1 - (2.74/20.02) = 0.8631 = 86.31\%$

# Example: F- statistics

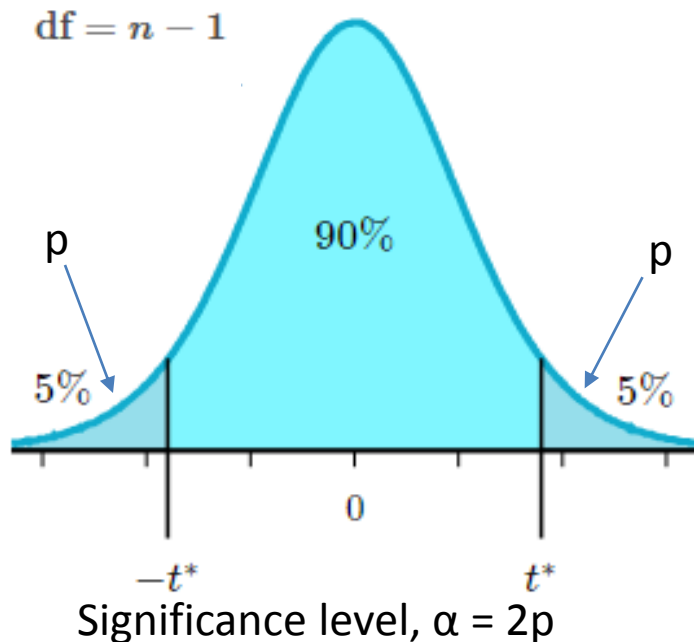
x	y	$\hat{y} = (41/42)x - (5/21)$	Squared error from line $(y - \hat{y})^2$	Squared error from mean $(\hat{y} - \bar{y})^2$
-2	-3	-2.1905	0.655328798	5.96
-1	-1	-1.2143	0.045918367	2.14
1	2	0.7381	1.592403628	0.24
4	3	3.66667	0.444444444	11.68
	$\bar{y} = 0.25$	Total	SSE = 2.738095238	SSR = 20.02

- $n = 4$
- $F = [SSR/(1)]/[SSE/(n-2)]$   
 $= 20.02 \times 2 / 2.73$   
 $= 14.66$

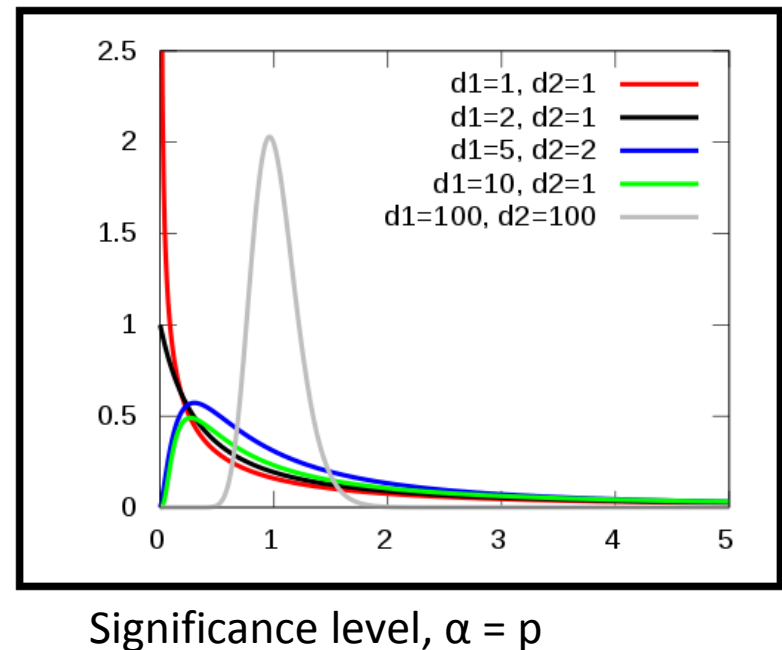
# T-distribution and F-distribution

- F-distribution curve is dependent on degrees of freedom for numerator and denominator
- F- Distribution is positive sided
- Therefore F- distribution does not have two tails
- T- distribution is positive and negative sided, has two tails

T- distribution

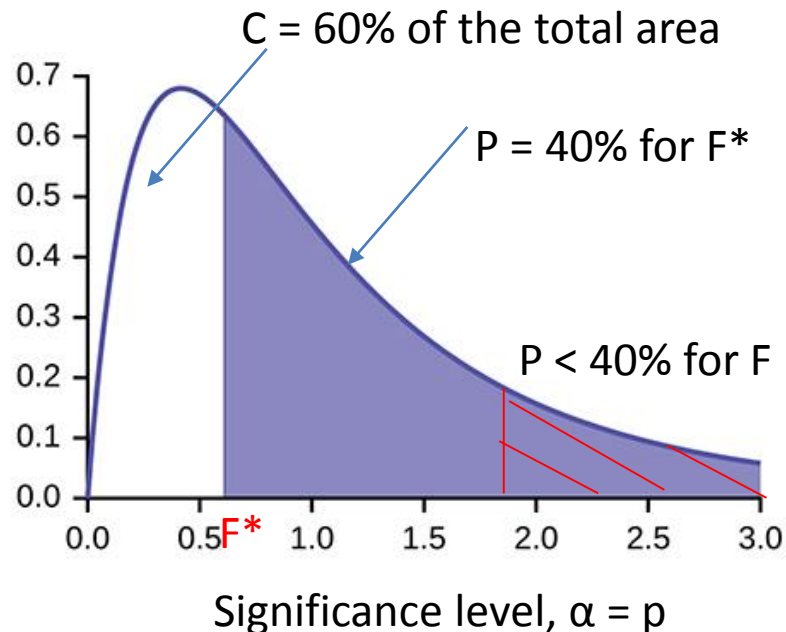


T- distribution



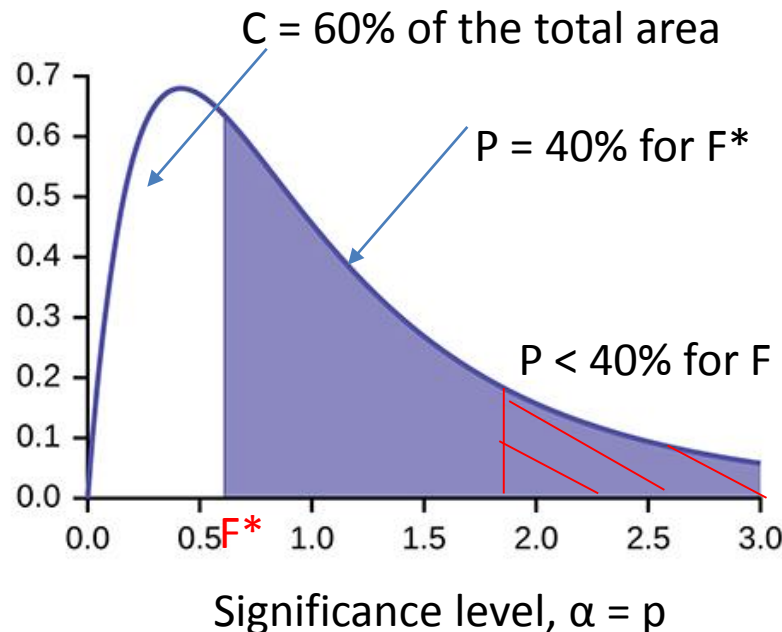
# Significance of p-value

- Confidence interval,  $C$  is for slope 0 (null hypothesis)
- For the given value of  $C$  (or  $F^*$ ), p-value corresponding to  $F^*$  denotes the probability that slope is not within  $F^*$
- Ex:  $C = 60\%$ ,  $\alpha = 1 - C = 40\% = p$



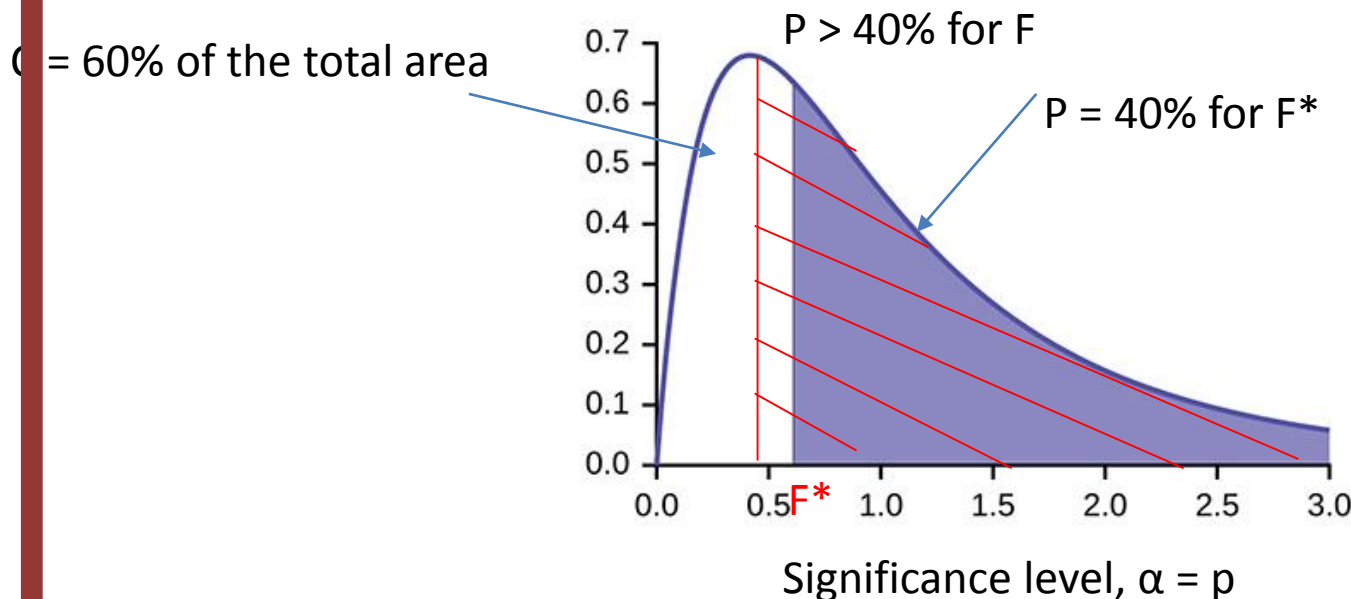
# Significance of p-value

- If 'p-value corresponding to calculated F-value' is less than 'p-value corresponding to  $F^*$ ' then
- Actual slope does not fall in confidence interval for slope 0
- Therefore null hypothesis can be rejected



# Significance of p-value

- If 'p-value corresponding to calculated F-value' is more than 'p-value corresponding to  $F^*$ ' then
- Actual slope fall in confidence interval for slope 0
- Therefore null hypothesis can be accepted



# Ex: ANOVA table and the F-test

- Relation between skin cancer mortality and latitude
- There were 49 states in the data set (49 samples)
- DF associated with  $SSR = 1$  for the simple linear regression model
- DF associated with  $SSTO = n-1 = 49-1 = 48$
- DF associated with  $SSE = n-2 = 49-2 = 47$
- Total degrees of freedom:  $1 + 47 = 48$

The regression equation is  $Mort = 389 - 5.98 \text{ Lat}$

Predictor	Coef	SE Coef	T	P
Constant	389.19	23.81	16.34	0.000
Lat	-5.9776	0.5984	-9.99	0.000

$S = 19.12$        $R\text{-Sq} = 68.0\%$        $R\text{-Sq(adj)} = 67.3\%$

## Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	36464	36464	99.80	0.000
Residual Error	47	17173	365		
Total	48	53637			

# ANOVA table and the F-test

- The sums of squares add up:  $SSTO = SSR + SSE$   
 $53637 = 36464 + 17173$
- $F = [SSR/(1)]/[SSE/(n-2)]$   
 $= 36464 \times 47 / 17173$   
 $= 99.97$

The regression equation is  $Mort = 389 - 5.98 \text{ Lat}$

Predictor	Coef	SE Coef	T	P
Constant	389.19	23.81	16.34	0.000
Lat	-5.9776	0.5984	-9.99	0.000

$S = 19.12$        $R\text{-Sq} = 68.0\%$        $R\text{-Sq}(\text{adj}) = 67.3\%$

## Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	36464	36464	99.80	0.000
Residual Error	47	17173	365		
Total	48	53637			



# ANOVA table and the F-test

- Assume  $C = 99\%$
- Therefore  $\alpha = 1\% = 0.001$
- For the data, calculated p-value is 0.000
- Which is less than 0.001
- Therefore null hypothesis can be rejected

The regression equation is  $\text{Mort} = 389 - 5.98 \text{ Lat}$

Predictor	Coef	SE Coef	T	P
Constant	389.19	23.81	16.34	0.000
Lat	-5.9776	0.5984	-9.99	0.000

$S = 19.12$        $R\text{-Sq} = 68.0\%$        $R\text{-Sq}(\text{adj}) = 67.3\%$

## Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	36464	36464	99.80	0.000
Residual Error	47	17173	365		
Total	48	53637			

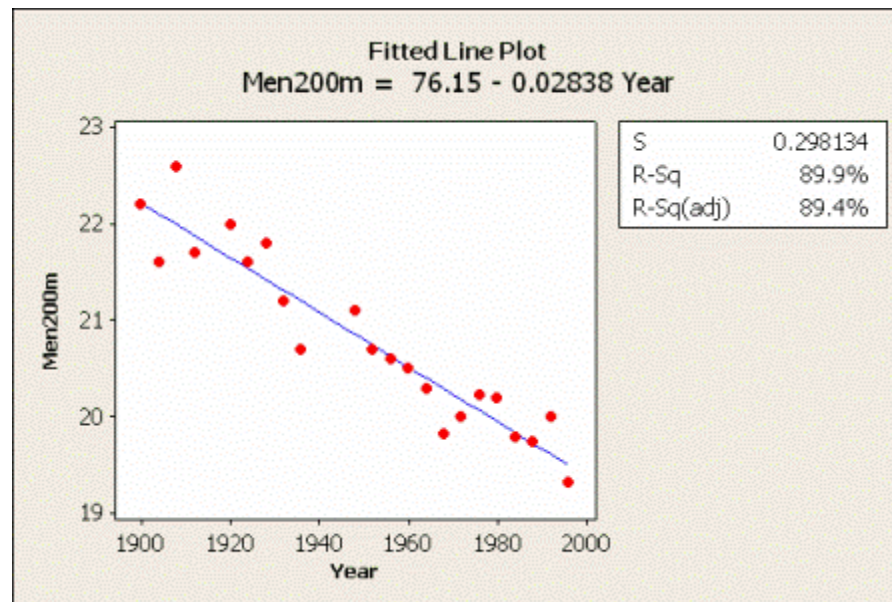
# Hypothesis Using ANOVA

- For computation of F-statistics, numerator df and denominator degree of freedom are considered
- In excel '=finv( $\alpha$ ,df1,df2)' shows F value
- Same can be used for '=tinv(p,df1)'
- Sample table

F Table for $\alpha = 0.10$											
	df <sub>1</sub> =1	2	3	4	5	6	7	8	9	10	12
df <sub>2</sub> =1	19.96346	49.50000	53.58324	55.83296	57.24008	58.20442	58.90595	59.43898	59.85759	60.19498	60.701
2	8.52832	9.00000	9.18179	9.24342	9.29283	9.32353	9.34908	9.36877	9.38514	9.39857	9.408
3	5.53832	5.48238	5.39877	5.34264	5.30918	5.28473	5.26819	5.25187	5.24000	5.23041	5.21
4	4.54477	4.32456	4.19088	4.10725	4.05058	4.00975	3.97897	3.95494	3.93567	3.91988	3.89
5	4.06042	3.77972	3.62948	3.52020	3.45298	3.40451	3.36790	3.33928	3.31828	3.29740	3.26
6	3.77595	3.48330	3.28876	3.18076	3.10751	3.05455	3.01448	2.98304	2.95774	2.93683	2.90

# Ex 1: SLR Evaluation using ANOVA

- Given data set contains the winning times (in seconds) of the 22 men's 200 meter Olympic sprints held between 1900 and 1996
- Is there a linear relationship between year and the winning times?
- Are Sprinters Getting Faster?



# Conduct the formal $F$ -test

- Null hypothesis  $H_0: \beta_1 = 0$
- Alternative hypothesis  $H_a: \beta_1 \neq 0$
- Consider P-value, which is 0.000 (to three decimal places),  
That is, the P-value is less than 0.001
- Therefore, we reject the null hypothesis  $H_0: \beta_1 = 0$  in favor of the alternative hypothesis  $H_A: \beta_1 \neq 0$

Predictor	Coef	SE Coef	T	P
Constant	76.153	4.152	18.34	0.000
Year	-0.0284	0.00213	-13.33	0.000

## Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	15.796	15.796	177.7	0.000
Residual Error	20	1.778	0.089		
Total	21	17.574			

# Equivalence of ANOVA $F$ -test and $t$ -test

$$(t_{(n-2)}^*)^2 = F_{(1,n-2)}^*$$

$$(-13.33)^2 = 177.7$$

Predictor	Coef	SE Coef	T	P
Constant	76.153	4.152	18.34	0.000
Year	-0.0284	0.00213	-13.33	0.000

## Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	15.796	15.796	177.7	0.000
Residual Error	20	1.778	0.089		
Total	21	17.574			

# Equivalence of ANOVA $F$ -test and $t$ -test

- For a given significance level  $\alpha$ , the  $F$ -test of  $\beta_1 = 0$  versus  $\beta_1 \neq 0$  is algebraically equivalent to the two-tailed  $t$ -test.
- We get the same  $P$ -values,
- If  $F$ -test rejects  $H_0$  then  $t$ -test also rejects it
- Same is for  $H_a$
- The  $F$ -test is appropriate for testing that the slope differs from 0 ( $\beta_1 \neq 0$ ).
- Use the  $t$ -test to test that the slope is positive ( $\beta_1 > 0$ ) or negative ( $\beta_1 < 0$ )
- The  $F$ -test is more useful for the multiple regression model for which more than one slope parameter are to be tested

# Equivalence of ANOVA $F$ -test and $t$ -test

- $P$ -value associated with the  $t$ -test is the same as the  $P$ -value associated with the analysis of variance  $F$ -test
- This is always true for the simple linear regression model
- Both  $P$ -values are 0.000 (to three decimal places)

Predictor	Coef	SE Coef	T	P
Constant	76.153	4.152	18.34	0.000
Year	-0.0284	0.00213	-13.33	0.000

Analysis of Variance					
Source	DF	SS	MS	F	P
Regression	1	15.796	15.796	177.7	0.000
Residual Error	20	1.778	0.089		
Total	21	17.574			

# Ex: ANOVA for multiple input variables

- Three types of different food are given to people who are taking tests. Their scores are mentioned in each column
- Data points are grouped into three sets of samples
- In general there are m groups (columns) with n data points each (row)
- Determine mean score for each type of food

Exam scores ↓	Food 1 ( $X_1$ )	Food 2 ( $X_2$ )	Food 3 ( $X_3$ )
	3	5	5
	2	3	6
	1	4	7
	$\bar{X}_1 =$	$\bar{X}_2 =$	$\bar{X}_3 =$



# Ex: ANOVA

Food 1 ( $X_1$ )	Food 2 ( $X_2$ )	Food 3 ( $X_3$ )
3	5	5
2	3	6
1	4	7
$\bar{X}_1=2$	$\bar{X}_2=4$	$\bar{X}_3=6$

↓ Exam scores

- Mean of group 3 is best among all
- Mean of population for each food is different
- That shows that type of food has effect on score
- Mean of entire population,  
$$\bar{X} = (3+2+1+5+3+4+5+6+7)/9 = 4$$

# Ex: ANOVA

- Mean of entire population,  $\bar{X} = 4$
- $SST = \sum (X_i - \bar{X})^2$   
 $= (3-4)^2 + (2-4)^2 + (1-4)^2 + (5-4)^2 + (3-4)^2$   
 $+ (4-4)^2 + (5-4)^2 + (6-4)^2 + (7-4)^2 = 30$

Food 1 ( $X_1$ )	Food 2 ( $X_2$ )	Food 3 ( $X_3$ )
3	5	5
2	3	6
1	4	7
$\bar{X}_1=2$	$\bar{X}_2=4$	$\bar{X}_3=6$



Exam  
scores

# Ex: ANOVA

- SSW (Sum of Squares for variation within each group) represents how far each data point is from mean of that group
- $$\begin{aligned} \text{SSW} &= \sum (X_i - \bar{X}_1)^2 + \sum (X_i - \bar{X}_2)^2 + \sum (X_i - \bar{X}_3)^2 \\ &= (3-2)^2 + (2-2)^2 + (1-2)^2 \\ &\quad + (5-4)^2 + (3-4)^2 + (4-4)^2 \\ &\quad + (5-6)^2 + (6-6)^2 + (6-7)^2 \\ &= 6 \end{aligned}$$

Food 1 ( $X_1$ )	Food 2 ( $X_2$ )	Food 3 ( $X_3$ )
3	5	5
2	3	6
1	4	7
$\bar{X}_1=2$	$\bar{X}_2=4$	$\bar{X}_3=6$

↓  
Exam  
scores

# Ex: ANOVA

- Sum of Squares Between (SSB) is due to variation between the mean of a group and mean of means.
- $$\begin{aligned} \text{SSB} &= \sum (X_1 - \bar{X})^2 + (X_2 - \bar{X})^2 + (X_3 - \bar{X})^2 \\ &= (2-4)^2 + (2-4)^2 + (2-4)^2 + (4-4)^2 + (4-4)^2 + (4-4)^2 + (6-4)^2 + (6-4)^2 + (6-4)^2 \\ &= 24 \end{aligned}$$

Variation	Degree of Freedom
SST = 30	$mn-1 = 8$
SSW = 6	$m(n-1) = 6$
SSB = 24	$m - 1$ for each data point = 2

- $\text{SST} = \text{SSW} + \text{SSB}$
- Total variation is due to within and between ( $24+6=30$ )
- Same is applicable to degrees of freedom ( $6+2=8$ )
- That is  $mn-1=m(n-1) + m-1$

# Hypothesis Using ANOVA

- $H_0$ : food does not make a difference
- $H_1$ : it does make a difference
- F-statistics =  $\{SSB/(m-1)\}/\{SSW/m(n-1)\}$ 
  - If numerator is much bigger than the denominator then
    - variation in the data is mostly due to the difference between actual means
    - and is less due to actual variations within the means
    - that shows that there is a difference in the true population
  - If denominator is large then variation within samples makes up the total variation
- Therefore large value of F-statistics says that variation in the output depends variation in the input samples
- Therefore there is a less probability that null hypothesis is correct


# Hypothesis Using ANOVA

- F-statistics =  $\{SSB/(m-1)\}/\{SSW/m(n-1)\}$   
=  $\{24/(3-1)\}/\{6/3(3-1)\} = 12$
- F-statistics in this example is a high number
- Each hypothesis test has some significance level,  $\alpha$
- Given,  $\alpha = 10\% = 0.1$
- Use F table to determine critical value of F given  $\alpha$  and df
- If calculated F-statistics > critical F-statistics
  - then reject null hypothesis
  - else do not reject it

# Hypothesis Using ANOVA

- Use F-table to determine critical value of F for the given  $\alpha$  and df
- F-statistics =  $\{SSB/(m-1)\}/\{SSW/m(n-1)\}$
- For computation of F-statistics, numerator df is 2 and denominator degree of freedom is 6

F Table for  $\alpha = 0.10$



	df <sub>1</sub> =1	2	3	4	5	6	7	8	9	10	12
df <sub>2</sub> =1	19.86346	49.50000	33.58334	25.83296	21.24008	18.20442	16.05995	14.43898	13.15759	12.19498	11.401
2	8.52432	9.00000	8.38179	7.74342	7.25263	6.82553	6.44908	6.11677	5.82054	5.55837	5.301
3	5.53833	5.46238	5.39077	5.34264	5.30016	5.26473	5.23419	5.20867	5.18700	5.16841	5.151
4	4.54477	4.52456	4.50888	4.49725	4.48808	4.48075	4.47497	4.46944	4.46417	4.45908	4.454
5	4.06042	4.07792	4.09448	4.10720	4.11798	4.12651	4.13370	4.13928	4.14328	4.14580	4.147
6	3.77595	3.46330	3.28876	3.18076	3.10751	3.05455	3.01448	2.98304	2.95774	2.93693	2.919

F critical,  
 $F^* = 3.46$   
 Calculated  
 $F = 12$

# Hypothesis Using ANOVA

- F- value = 12
- $F^*$  with 10% significance value = 3.46
- Since F-statistics  $> F^*$   
probability that null hypothesis is true is very low
- therefore we can reject null hypothesis
- It can be concluded that score in exam depends on the type of food



# Limitations of Statistical Model

- Regression model is selected in order to approximate the true population
- Simple Linear Regression model has two parameters, intercept and slope

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

- Simple Multiple regression model uses more than one independent variables

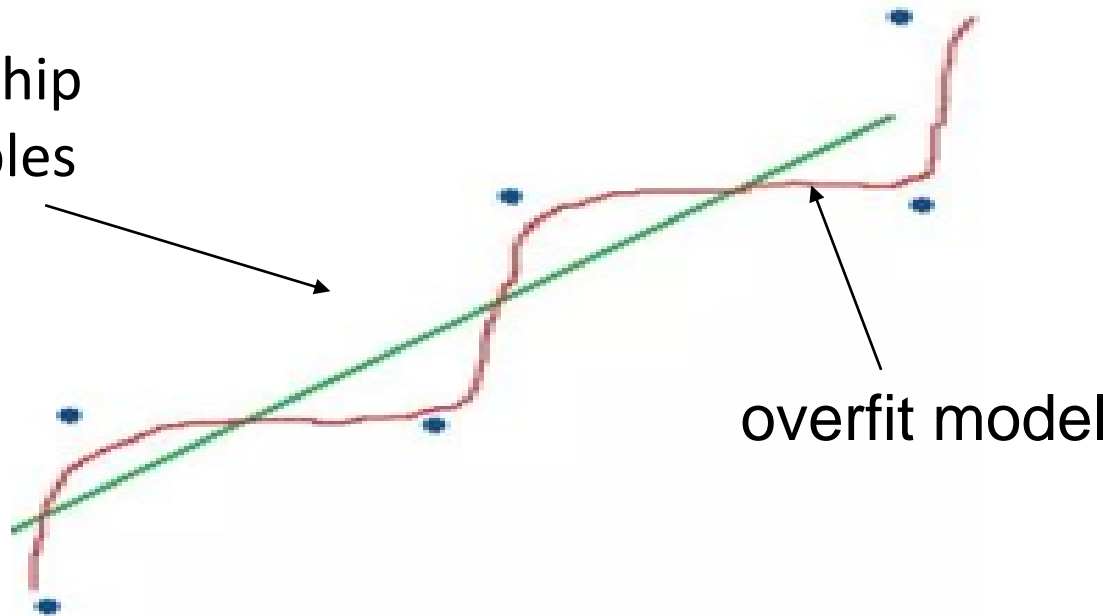
$$\hat{y} = b_0 + b_1x + b_2x_2 + \dots + b_nx_n$$

- And tries to fit all data points

# Simple Non Linear Regression Model

- May lead to overfitting and may cause random error
- Overfit regression models have too many terms for the number of observations
- Which results in noise coefficients rather than actual relationships

actual relationship  
between variables

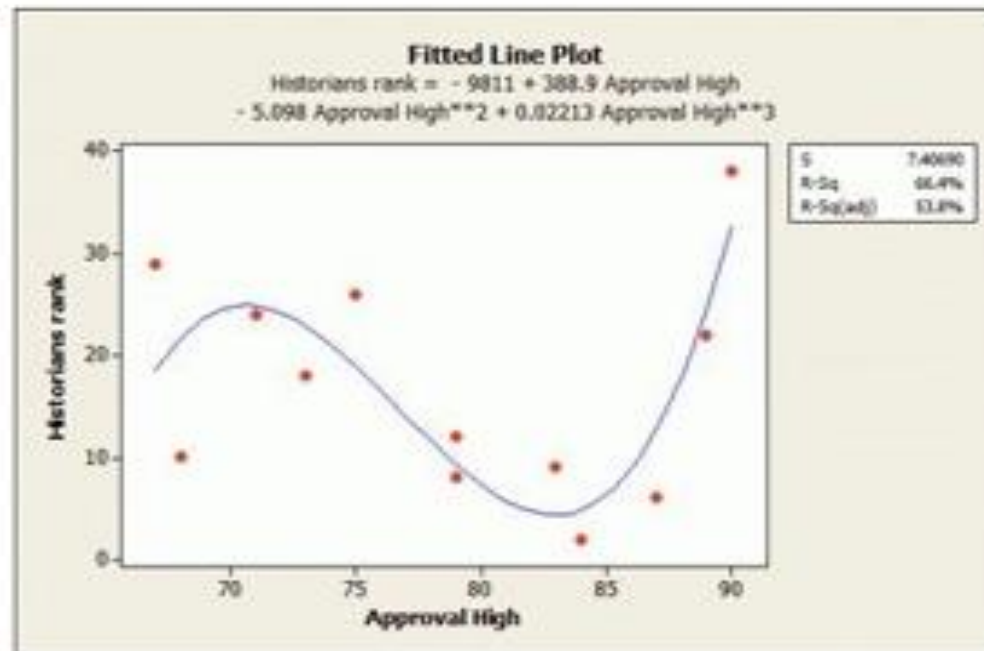


# Over-fitting

- A statistical model begins to describe the random error in the data rather than the relationships between variables
- R-squared is a popular measure of quality of fit in regression
- However it does not offer significant information about how well a given regression model can predict future values

# Over-fitting

- R-square value is high and explains good proportion of dependent variable
- It does not show the effect of overfitting
- Overfitting lead to erroneous R-squared, regression coefficients and p-values in the population



# Detecting over-fit models: Cross validation

- Used to estimate the behaviour of the large data set based on the a small part of data set
- Evaluate machine learning models on a limited data sample
- Use k number of groups to split the dataset
- Called k-fold cross validation
- Randomly split the dataset into k fold/groups of equal size
- First fold is considered as validation set and is verified against the remaining k-1 folds

# K-fold cross validation

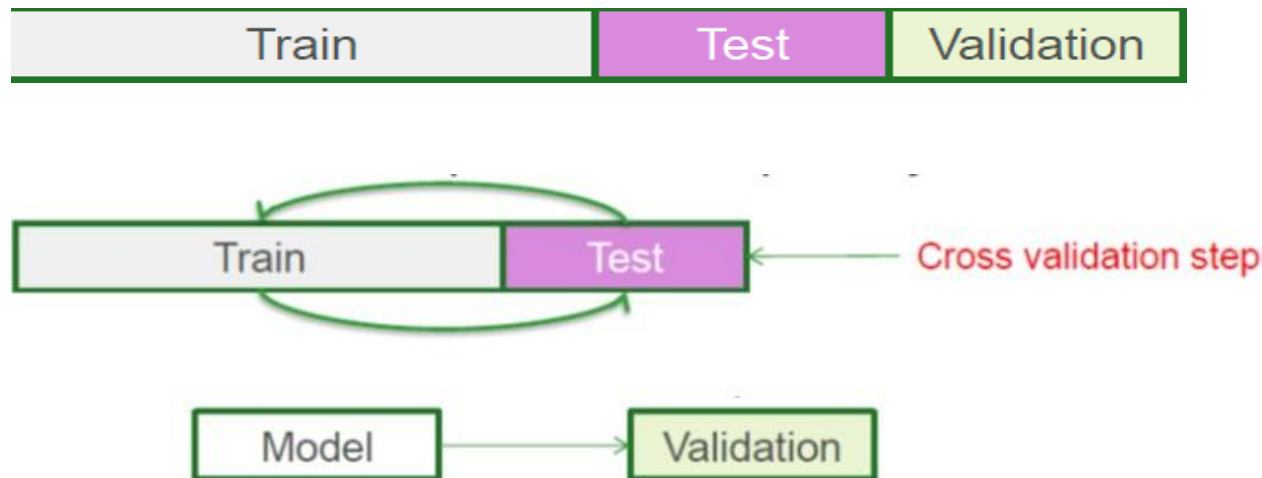
- Choosing the right value of  $k$  is quite complex
- Behaviour of the model is dependent on the dataset
- Some ways of choosing the value of  $k$  is
- Each train test group of data should be large enough to statistically representative
- $K=10$ , experimentally proven to be optimum
- $K=n$ , where  $n$  is size of data set such that each sample is given equal opportunity
- This is called Leave One Out Cross Validation (LOOCV)

# Ex: k-fold cross validation

- Data samples: [1, 2, 3, 4, 5, 6]
- $K=3$
- Fold1 = [5,2], Fold2 = [1,3], Fold3 = [4,6]
- Model1: trained on fold1 + fold2, tested on fold3
- Model2: trained on fold2 + fold3, tested on fold1
- Model3: trained on fold1 + fold3, tested on fold2
- Minimum number of samples can be 1

# Cross validation: The ideal procedure

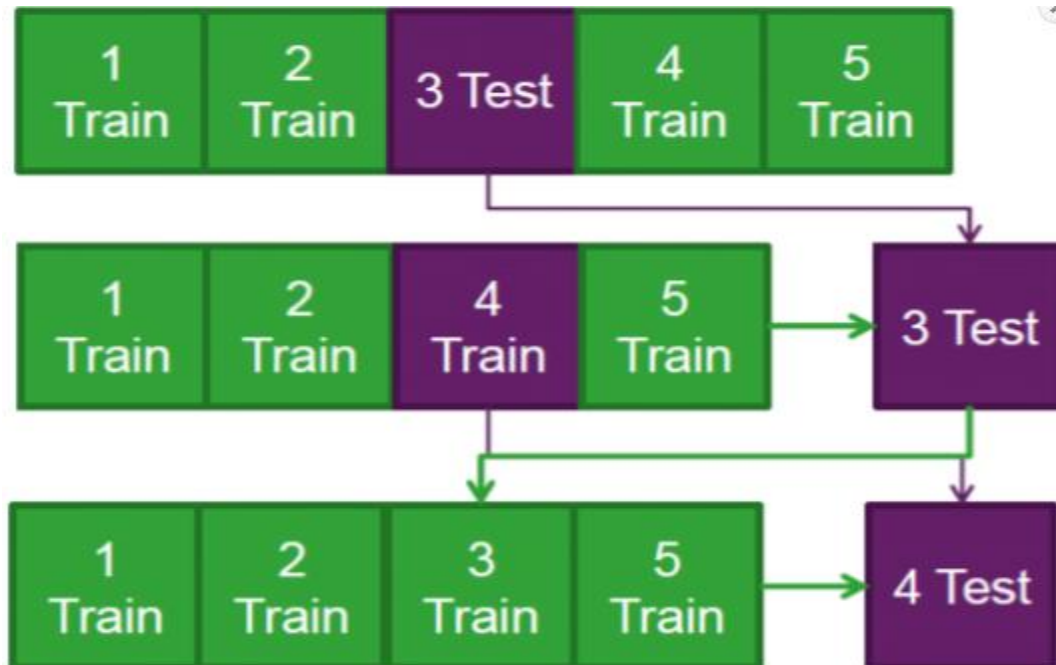
- Divide data into three sets, training, validation and test sets
- Parameters of regression model are calculated based on training data
- And accuracy is measured for new data
- The validation error gives an unbiased estimate of the predictive power of a model.





# K- fold Cross validation

- Split data into 5 samples
- Fit a model to the training sample
- Use test sample to determine Cross Validation Metric
- Repeat the process for next sample



# Logistic Regression

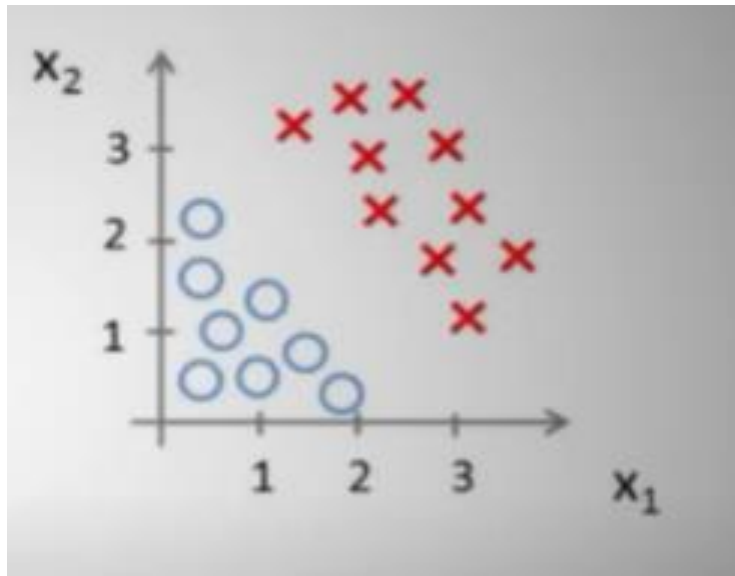
- For the given data points linear regression uses least square method to find optimum values for slope and intercept of a line to fit data
- Logistic Regression is a classification algorithm
- Used to classify data into the different classes

# Logistic Regression

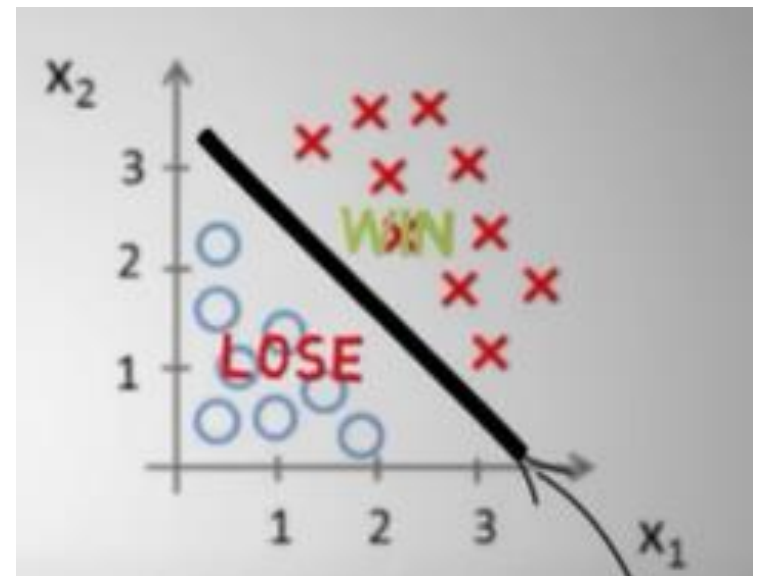
- Examples of classification
  - Email: Spam/ Not Spam
  - Online transaction: Fraudulent (Yes/ No)
  - Tumor: Malignant / Benign
- Types of logistic regression
  - Binary Classification  
Ex: Tumor Malignant or Benign  
With label '1' or '0' for each class
  - Multi-linear Class
    - Ex: Cats, dogs or Sheep
    - Labels are '0', '1', '2' for each class

# Logistic Regression

- Same for linear and logistic regression
- Here  $x_1$  and  $x_2$  are two features
- Classification is based on these features
- They are clustered together and can be separated by a classifying line



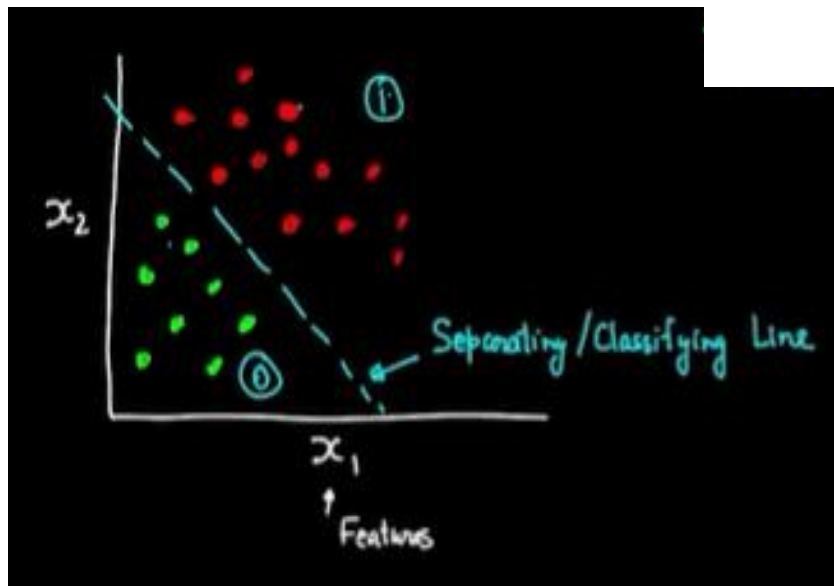
logistic regression



linear regression

# Logistic Regression

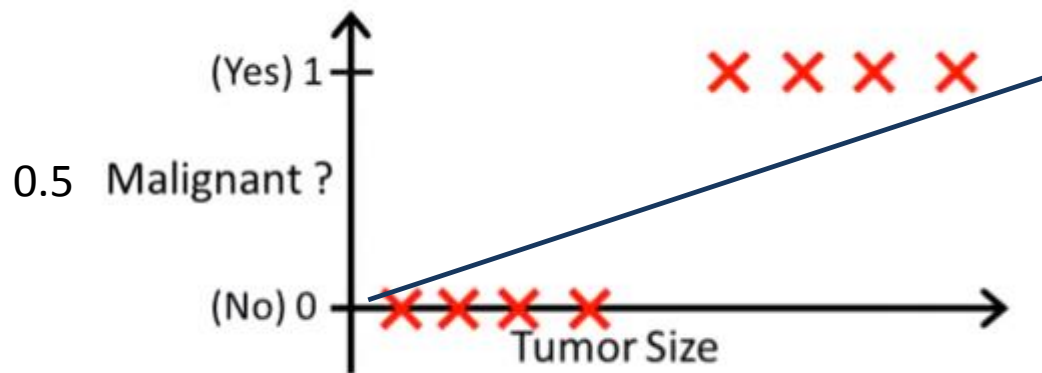
- Model should predict correct label



Data point	$x_1, x_2$ (input)	$Y$ (output)	$Y^{\wedge}$
1	$x^{(1)}_1, x^{(1)}_2$	'0' or '1'	'0'
2	$x^{(2)}_1, x^{(2)}_2$	'0' or '1'	'1'
:	:	:	
m	$x^{(m)}_1, x^{(m)}_2$	'0' or '1'	'1'

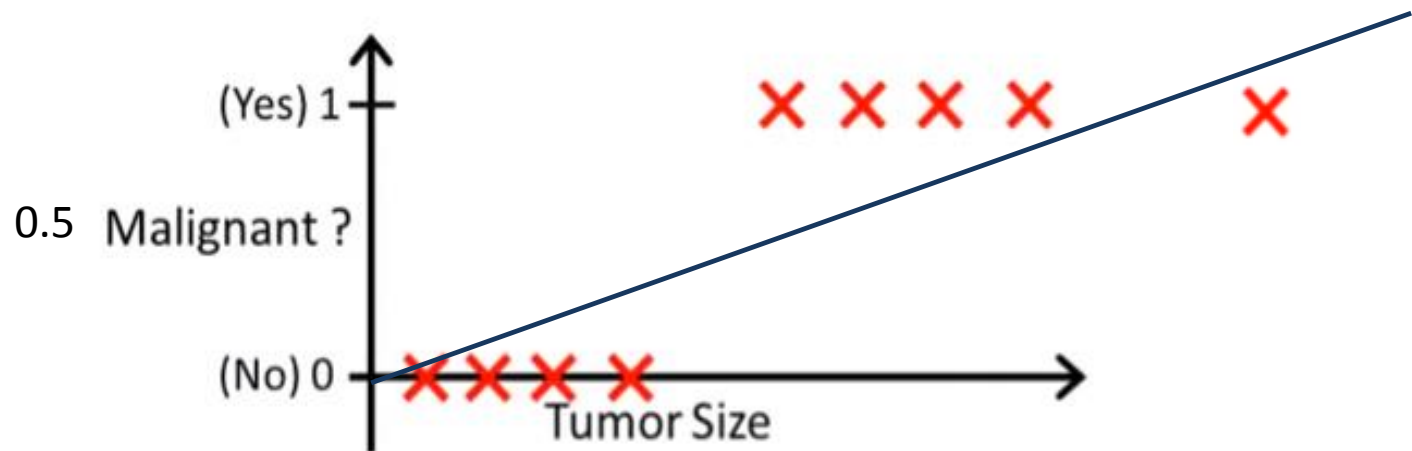
# Logistic Regression

- Output,  $y = 0$  Negative class (benign tumor)  
           $= 1$  Positive class (malignant tumor)
- Predictions of linear and logistic regression are same



# Logistic Regression

- If additional data point is added
- then linear regression shows incorrect results
- Predicted value of linear regression may exceed value, 1
- therefore linear regression can not be used for classification problems
- Requirement is to range of output to 0-1



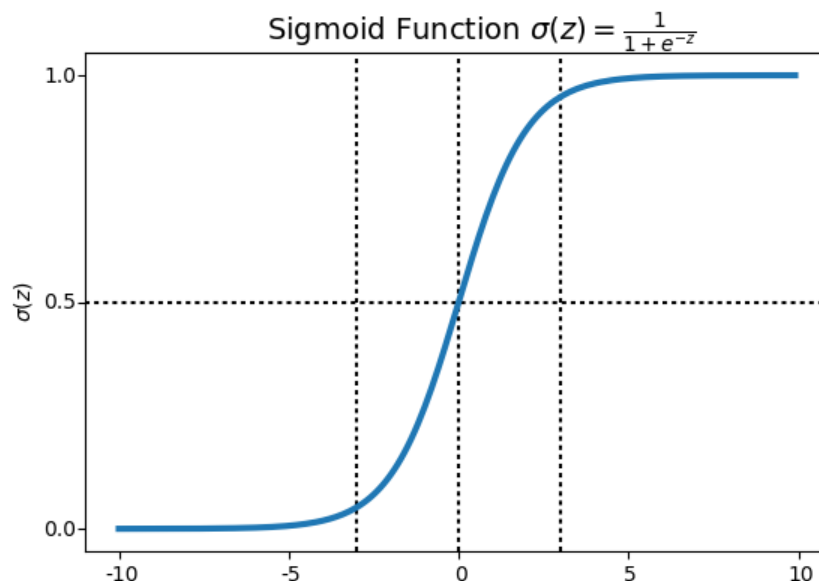
# What is the Sigmoid Function?

- Maps any real value into another value between 0 and 1
- Sigmoid function,  $\sigma(z) = 1/(1+e^{-z})$

As  $z \rightarrow \infty$ ,  $\sigma(z) \rightarrow 1$

As  $z \rightarrow -\infty$ ,  $\sigma(z) \rightarrow 0$

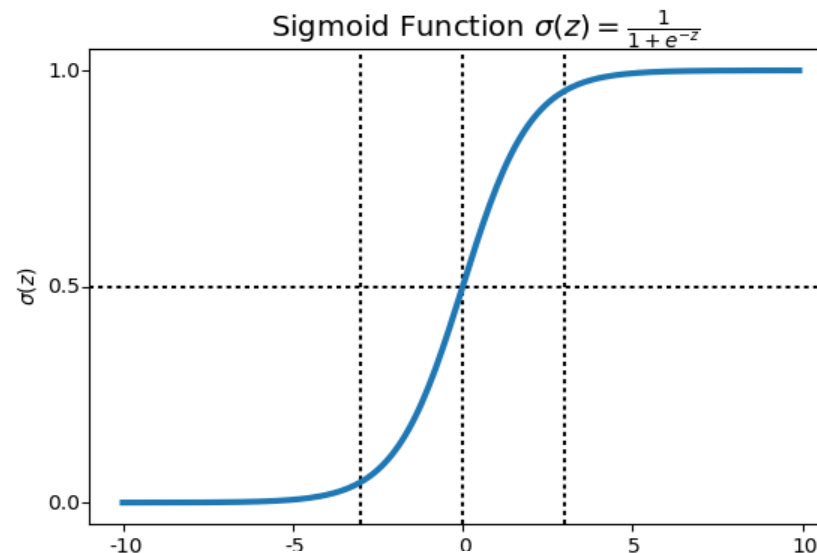
At  $z = 0$ ,  $\sigma(z) = 0.5$



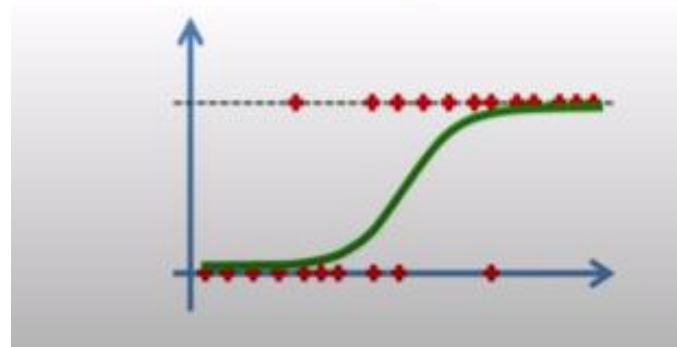
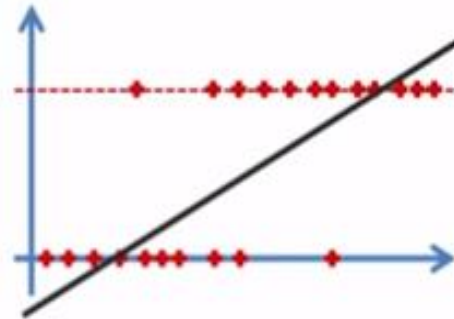


# What is the Sigmoid Function?

- $\sigma(\hat{y}) = \sigma(b_0 + b_1 x_1)$
- Therefore predicted value is from 0 to 1

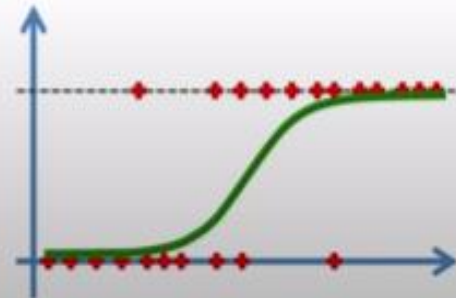
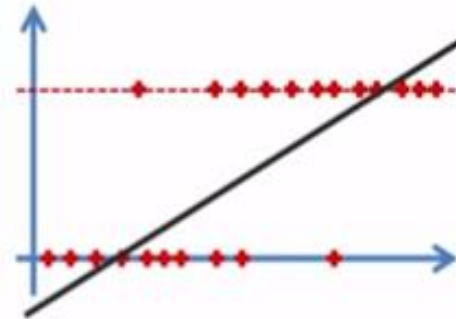


# Linear Regression and Logistic Regression



# Linear Regression and Logistic Regression

$$y = b_0 + b_1 * x$$

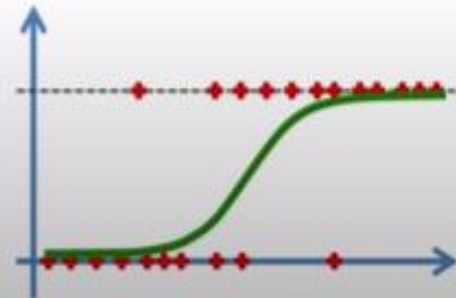
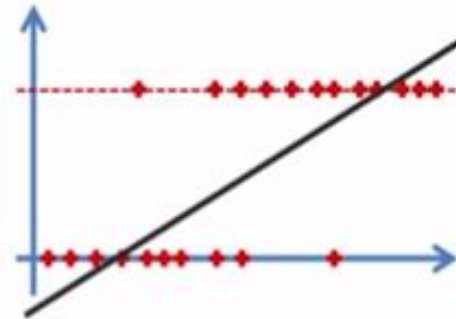


# Linear Regression and Logistic Regression

$$y = b_0 + b_1 * x$$

Sigmoid Function

$$p = \frac{1}{1 + e^{-y}}$$



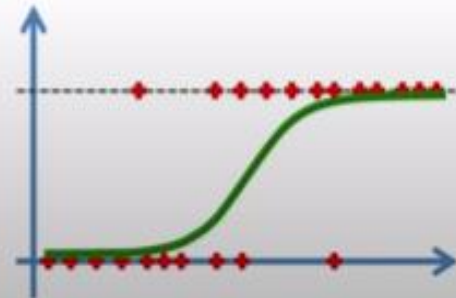
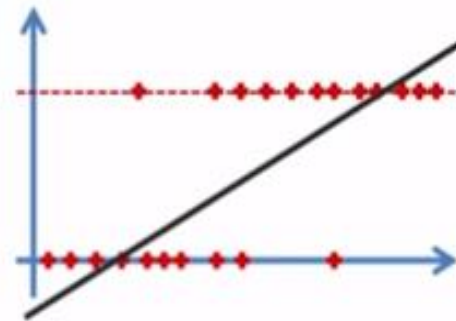
# Linear Regression and Logistic Regression

$$y = b_0 + b_1 * x$$

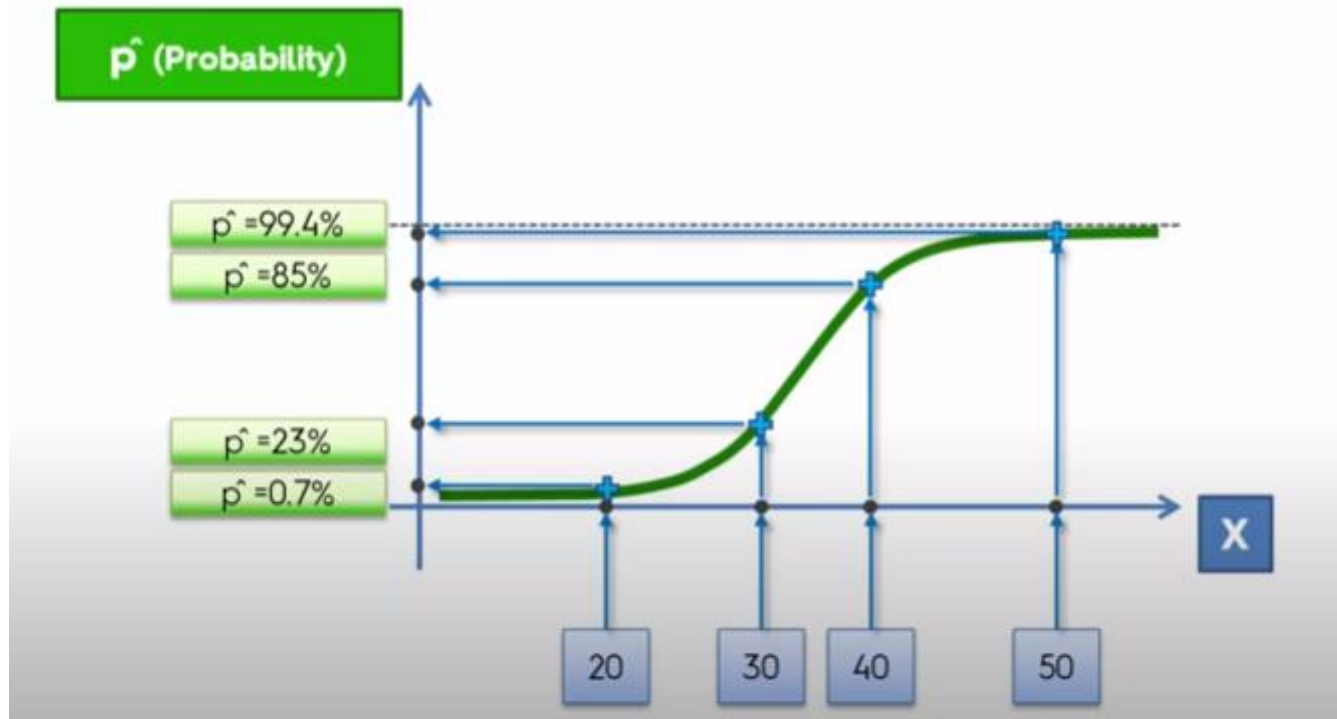
Sigmoid Function

$$p = \frac{1}{1 + e^{-y}}$$

$$\ln \left( \frac{p}{1 - p} \right) = b_0 + b_1 * x$$

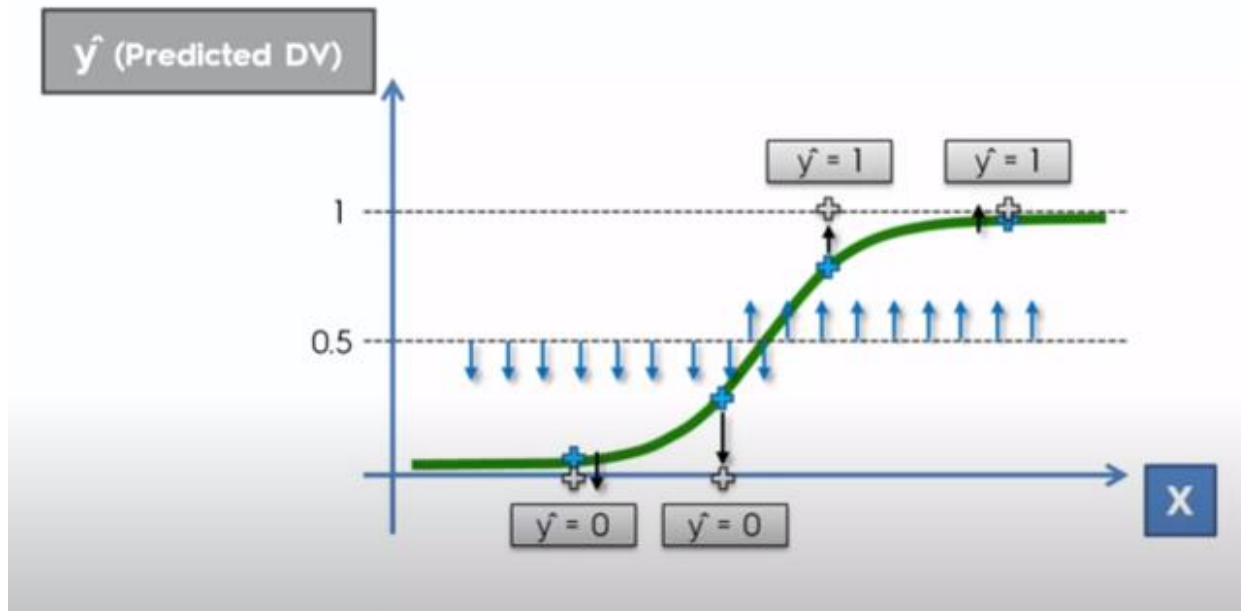


# Sigmoid function for prediction of Probability



If  $x = 0$ ,

# Sigmoid function for prediction of Probability



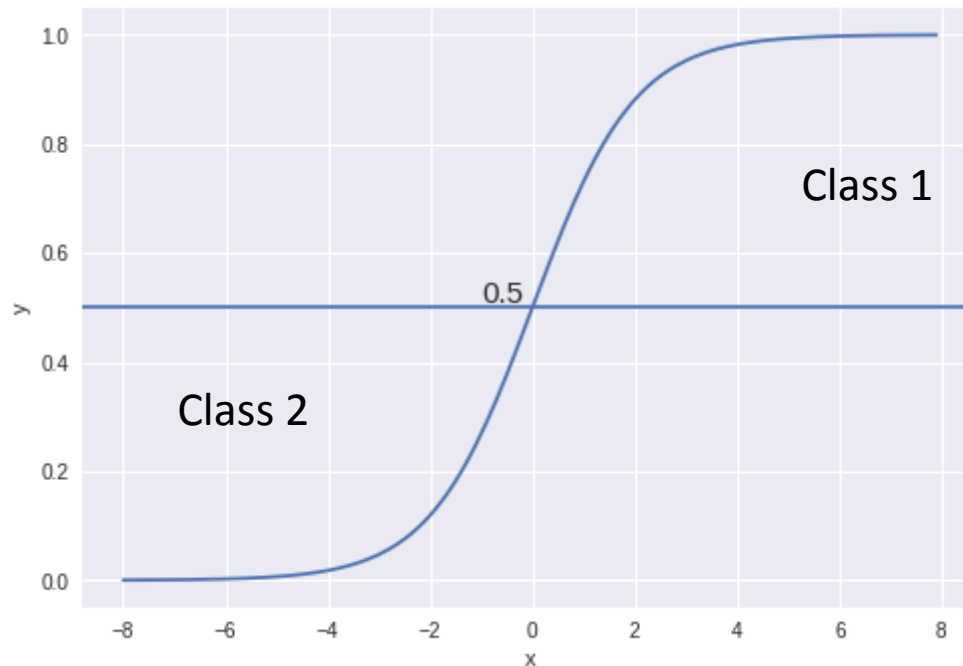
# Hypothesis Representation

- For linear regression hypothesis is
$$h(x) = \beta_0 + \beta_1 X$$
- For logistic regression hypothesis is
$$\sigma(Z) = \sigma(\beta_0 + \beta_1 X), \text{ where } Z = \beta_0 + \beta_1 X$$
- Where  $\sigma(Z)$  is sigmoid function
- Hypothesis for logistic regression should give values between 0 and 1
- $$h(x) = \text{sigmoid}(Z)$$
$$= 1/(1 + e^{-(\beta_0 + \beta_1 X)})$$



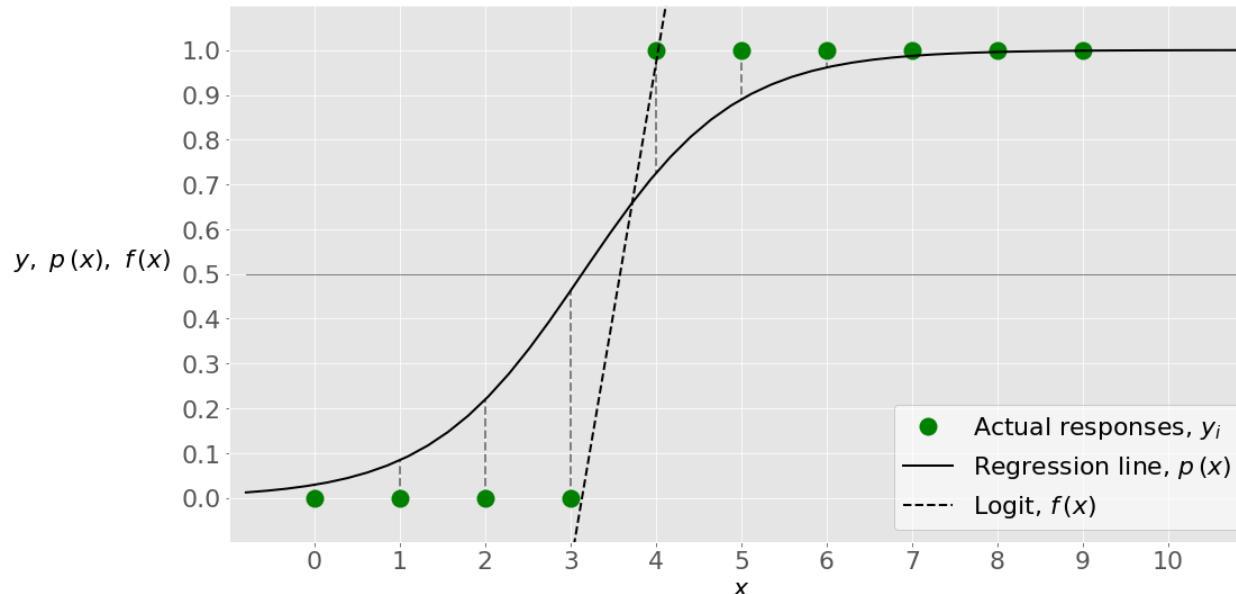
# Example

- Threshold is 0.5
- If prediction function returns 0.7
- Then this observation is in Class 1
- If prediction return 0.2
- then classify the observation as Class 2



# Example: Logistic Regression

- Logistic regression finds the weights  $b_0$  and  $b_1$  that correspond to the maximum Likelihood Function
- Therefore  $f(x) = b_0 + b_1x$  (dashed black line)
- Predicted probability  $p(x) = 1 / (1 + \exp(-f(x)))$  (full black line)
- In this case, the threshold  $p(x) = 0.5$



# Making Predictions with Logistic Regression

- Have a linear regression model that can predict whether a person is male or female based on their height (completely fictitious)
- Model is,  $y = -100 + 0.6x$
- Given a height of 150 cm, is the person male or female?
- Coefficients are  $b_0 = -100$  and  $b_1 = 0.6$
- $y = 1 / (1 + e^{(b_0 + b_1 * X)})$
- $y = 1 / (1 + e^{(-100 + 0.6 * X)})$
- $y = 1 / (1 + 4.5) = 0.18$
- If  $y < 0.5$  then male else female
  - 0 if  $p(\text{male}) < 0.5$
  - 1 if  $p(\text{male}) \geq 0.5$
- Therefore the person with height 150 cm is a male

# Example: Logistic Regression

- Given set of input-output (or  $x$ - $y$ ) pairs, represented by green circles
- Output  $y$  can only be 0 or 1
- For example, the leftmost green circle has the input  $x = 0$  and the actual output  $y = 0$
- Rightmost observation has  $x = 9$  and  $y = 1$ .

