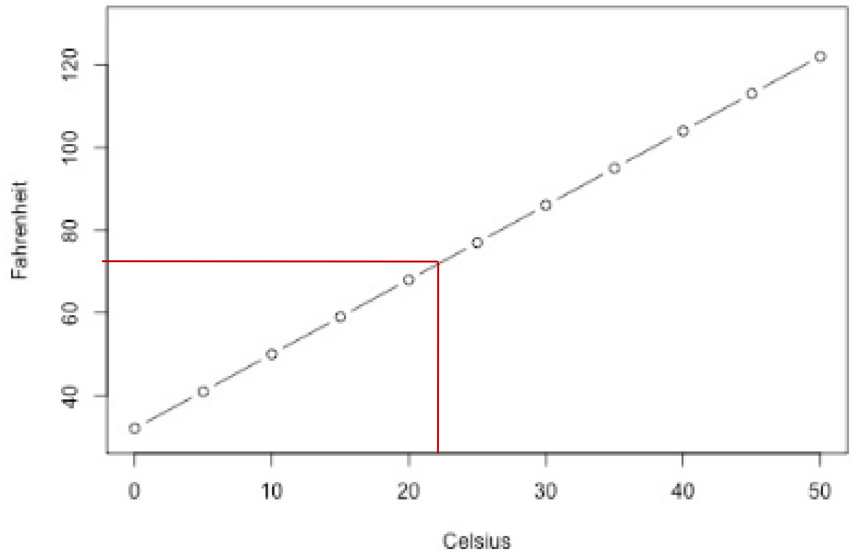
Unit 2

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

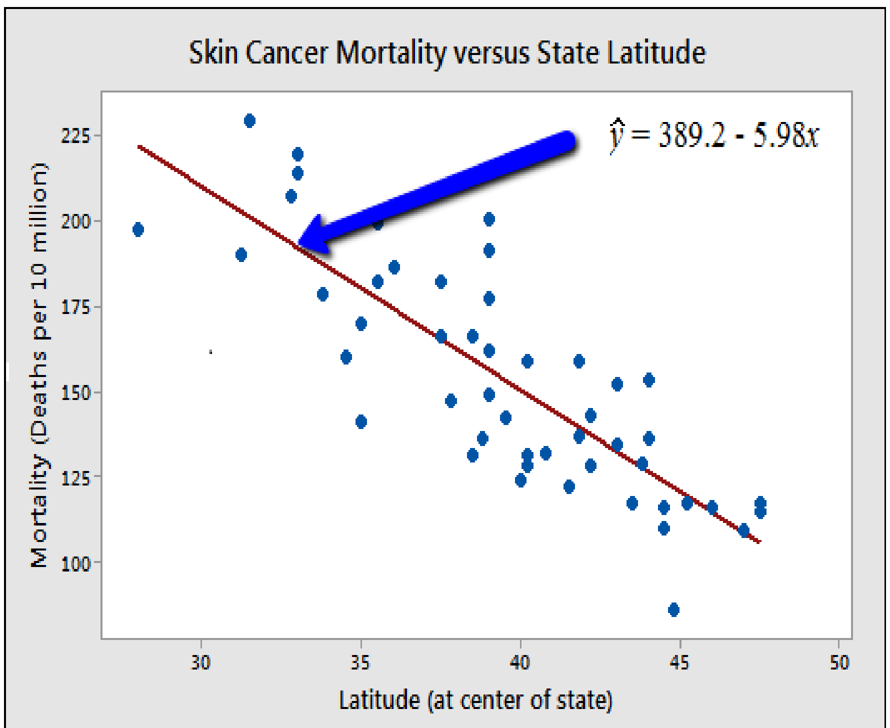
- Circumference = $\pi \times$ diameter
- Ohm's Law: $I = V/r$

  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

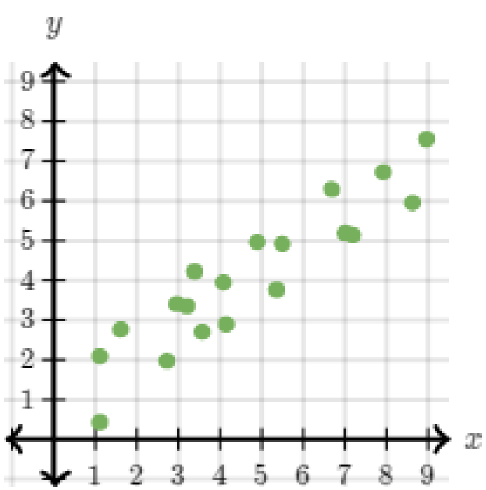
Examples: Statistical Relationship

- 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 decease perfectly

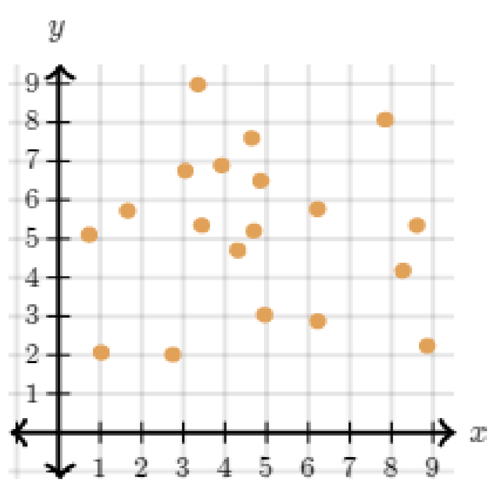
Types of Association

Scatter plots are used to see relationships between variables

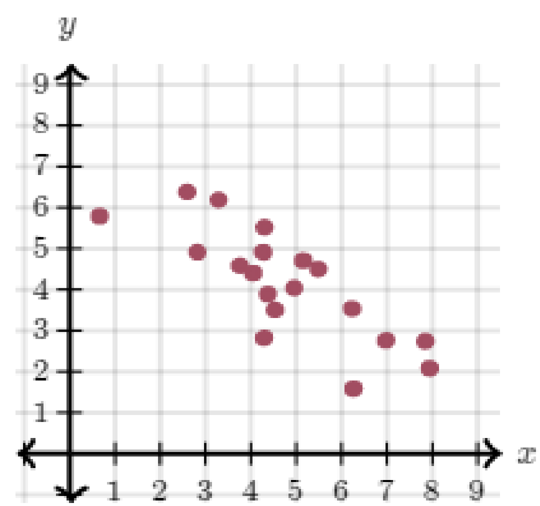
Positive association



No association

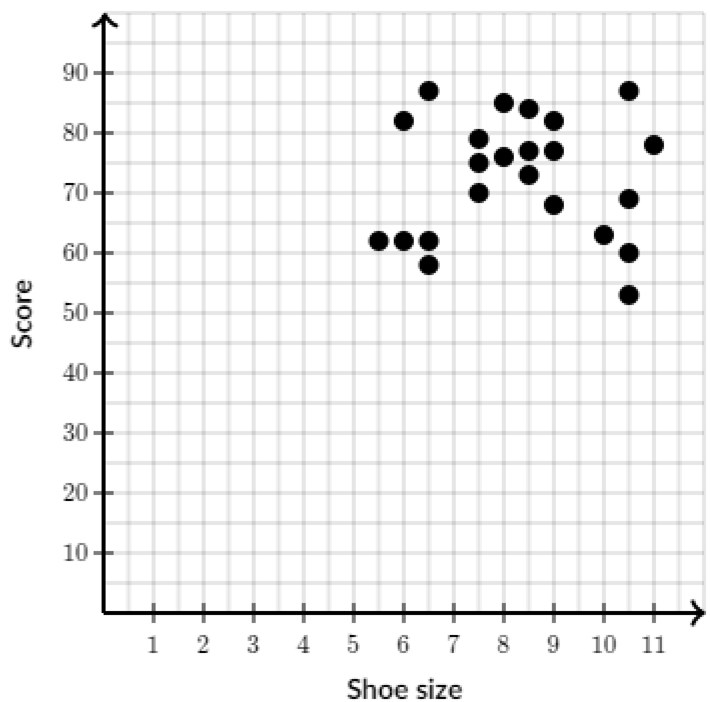


Negative association



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?

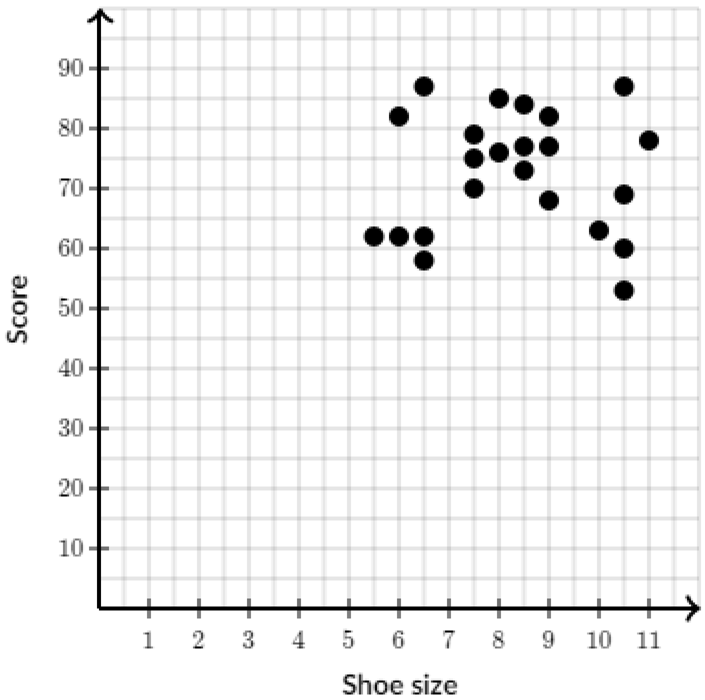
Positive association

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

  $\mathbf{x} = (x_1, \ldots, x_r)$, where r is the number of inputs

When Do You Need Regression?

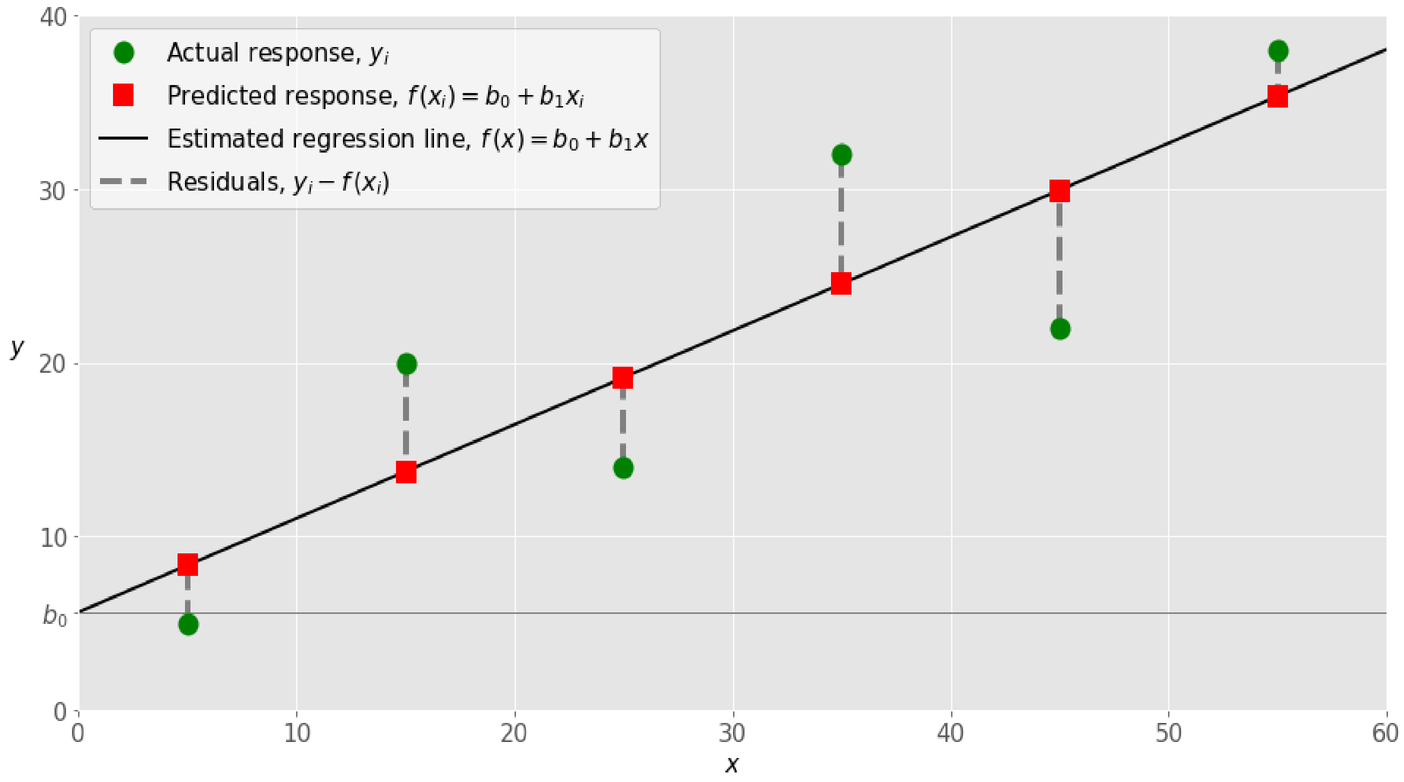
- How one or more variables influence the other
- Ex 1: Use regression to determine *if* and *to what extent* the experience or gender impact salaries
- Also useful when you want to forecast a response using a new set of predictors
- Ex 2: Try to predict electricity consumption of a household for the 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

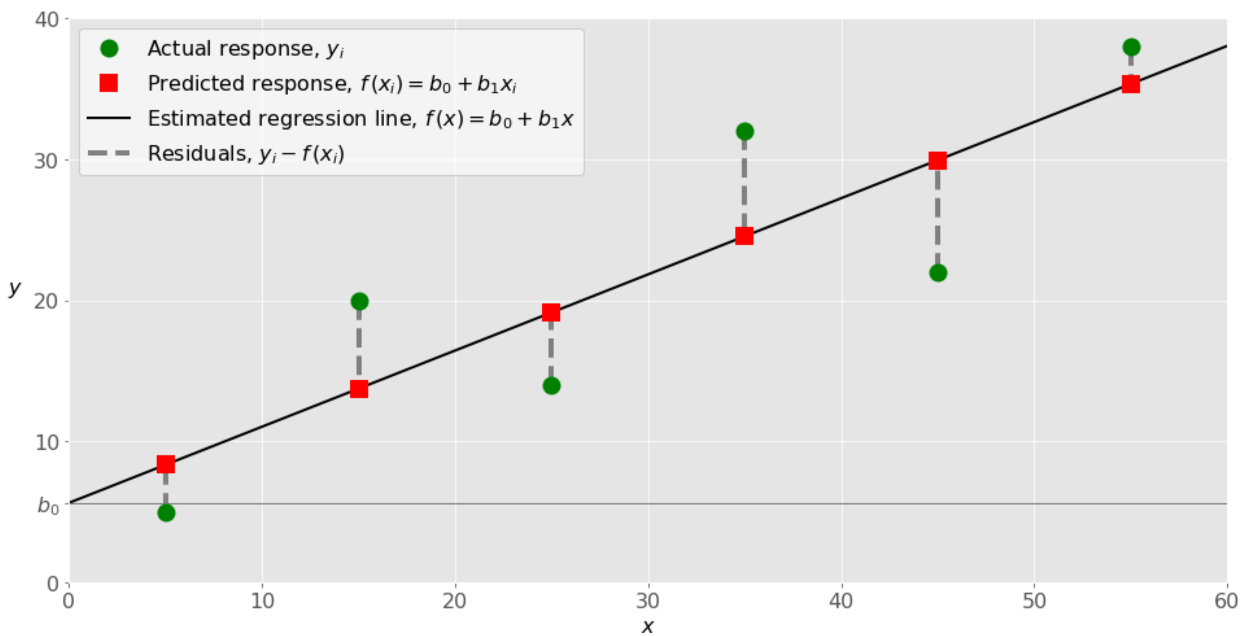
Simple Linear Regression

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



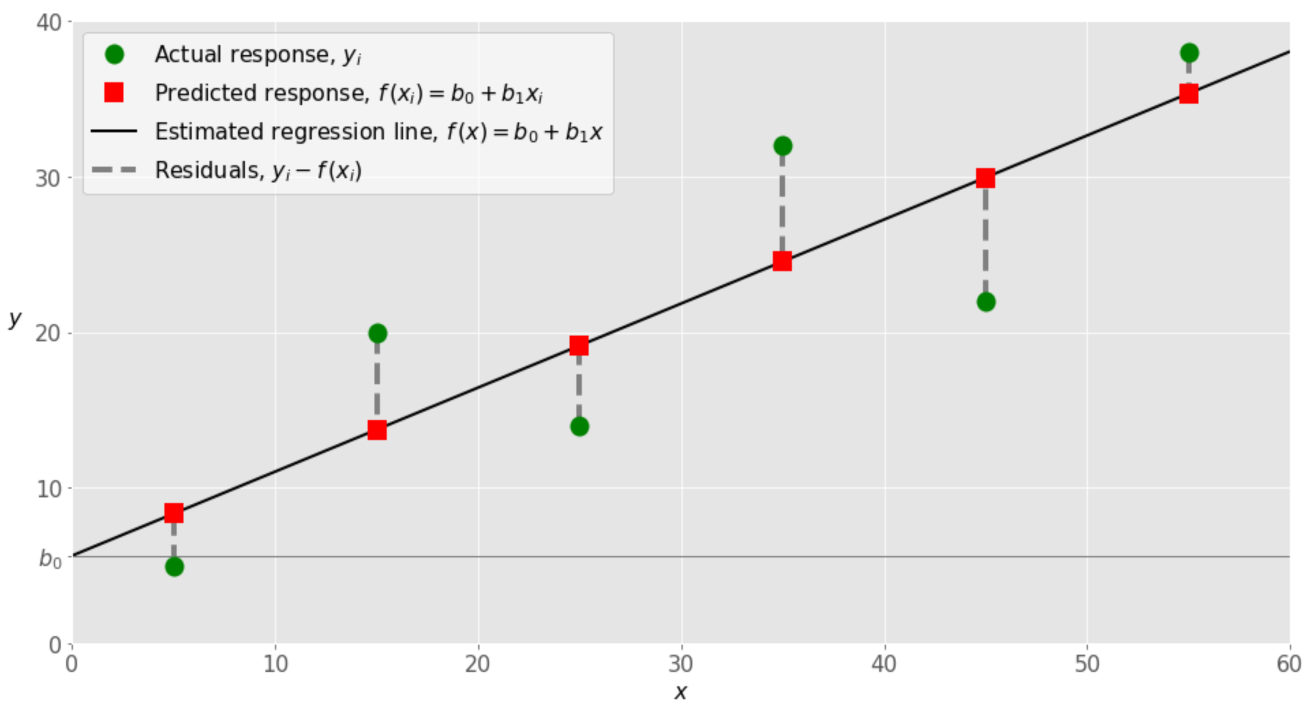
Simple Linear Regression

- The estimated regression function (black line) is represented by
  $f(x) = b_0 + b_1 x$
- 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.



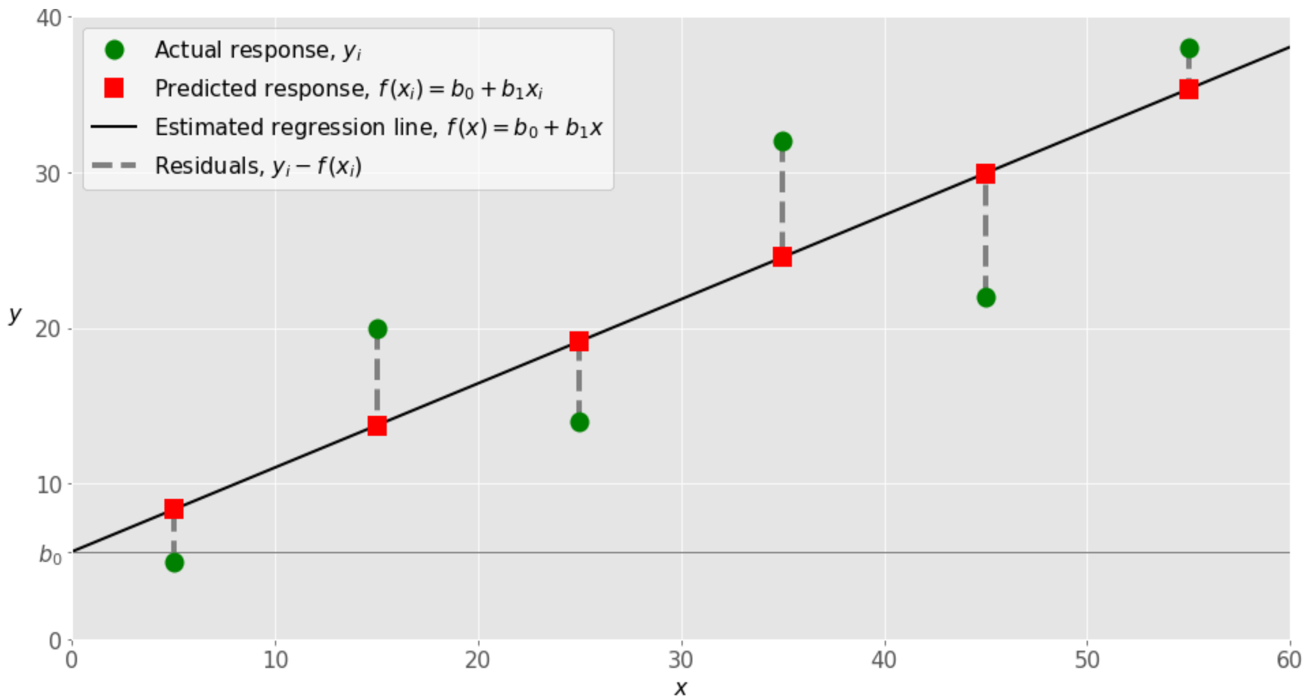
Simple Linear Regression

- 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

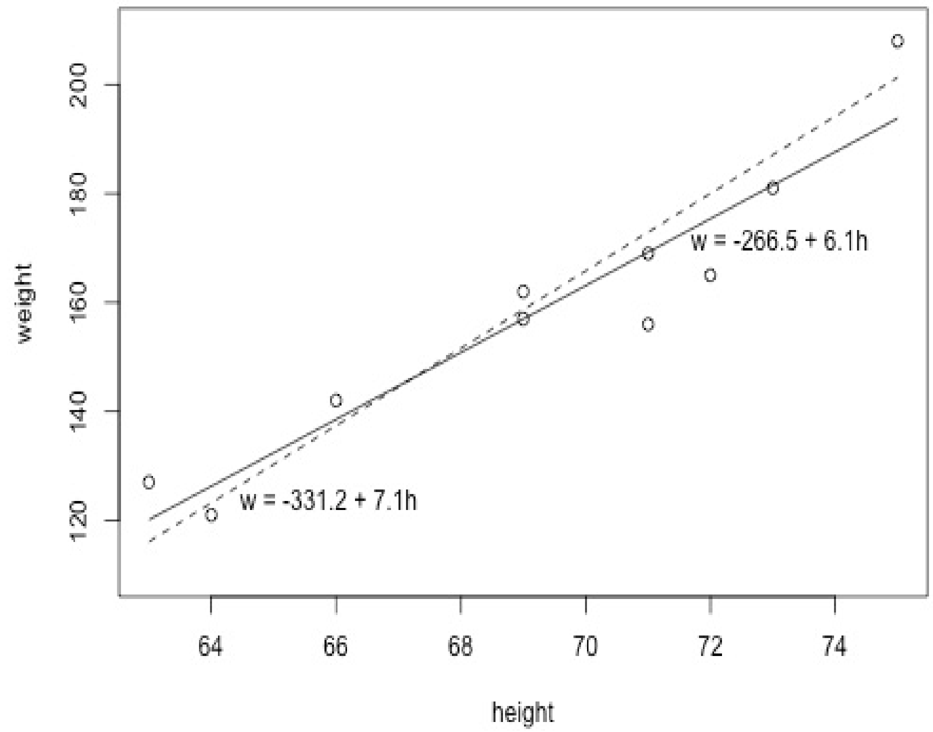
- Residuals (vertical dashed gray lines) can be calculated as
  $y_i - f(\mathbf{x}_i) = y_i - (b_0 - b_1 x_i)$ for $i = 1, \ldots, n$
- Residuals are the distances between the green circles and red squares
- Linear regression minimizes these distances and make the red squares as close to the predefined green circles as possible



What is the "Best Fitting Line"?

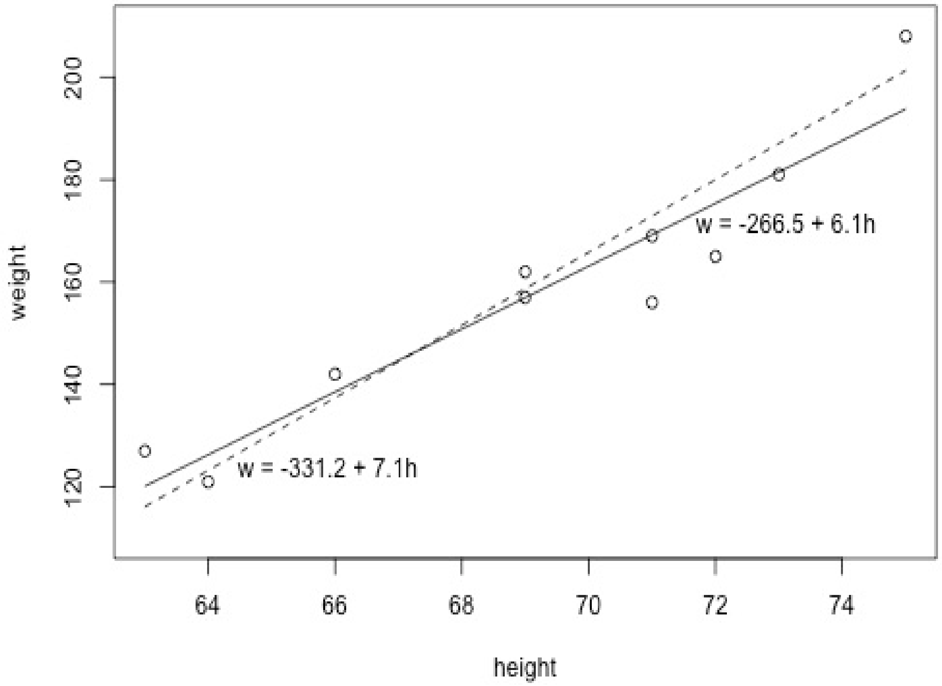
- Line which summarizes the trend between height and weight in the best way

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 $w = -266.53 + 6.1376\,h$

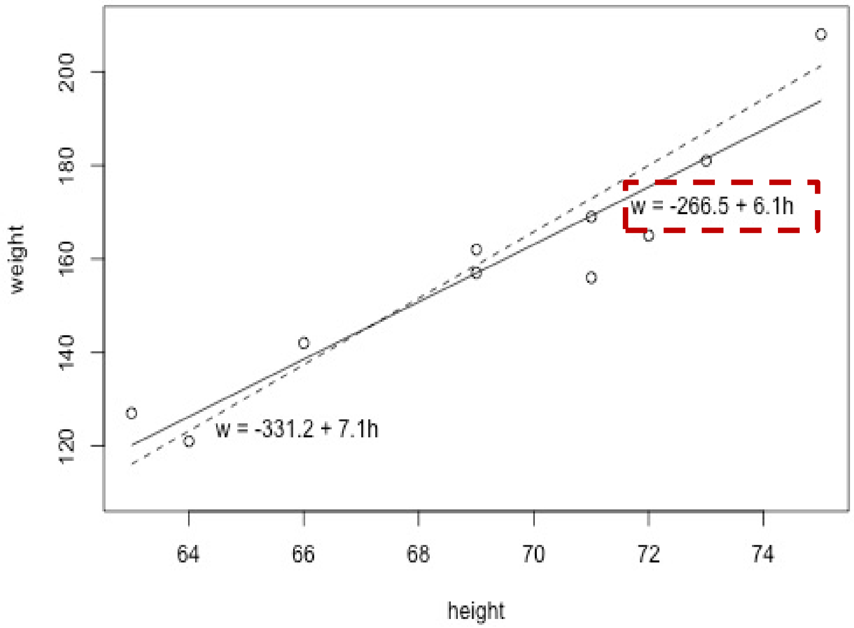


What is the "Best Fitting Line"?

- 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,

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

  $e_i = y_i - y\hat{}_i$

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

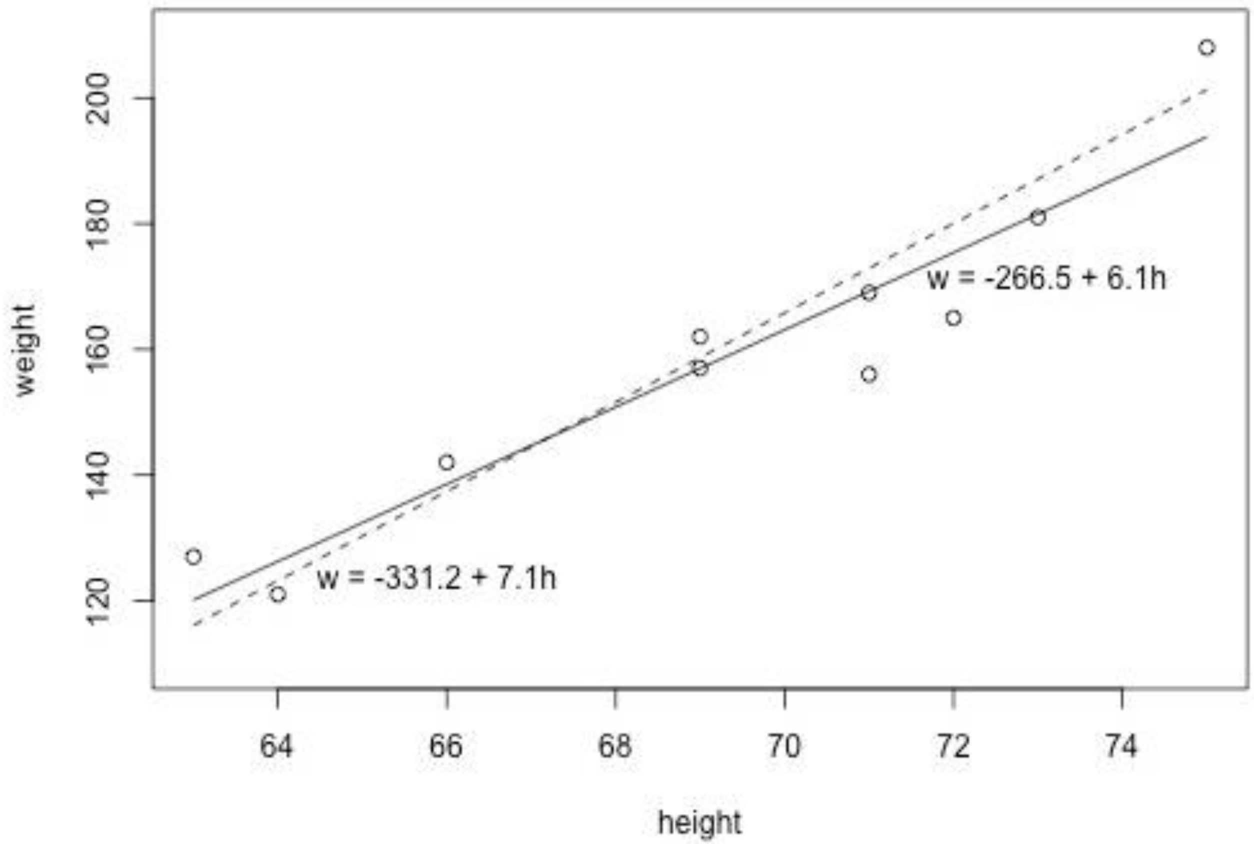
$$Q = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

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
					766.5

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
					597.4

- 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

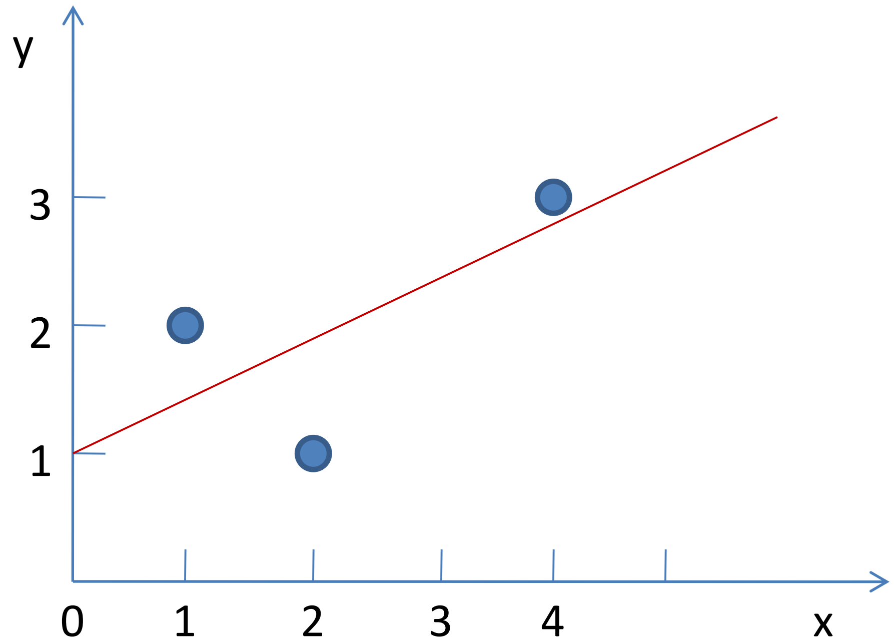
$$b_1 = (\bar{x}\bar{y} - \overline{xy})/(\bar{x}^2 - \overline{x^2})$$

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

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

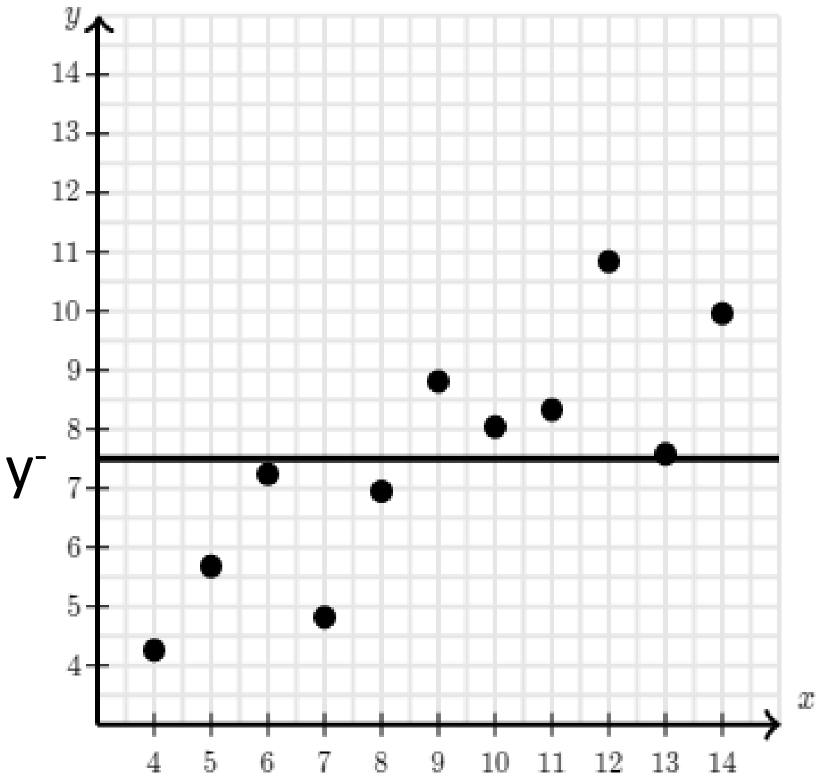
- Given relation between height and weight is
  w = -266.53 + 6.1376h
- If a person is 0 inches tall then his predicted weight is 266.53 pounds!
- Scope of the model does not include $x = 0$
- It is "extrapolated" beyond the "scope of the model“
- If the "scope of the model" includes $x = 0$, then b_0 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

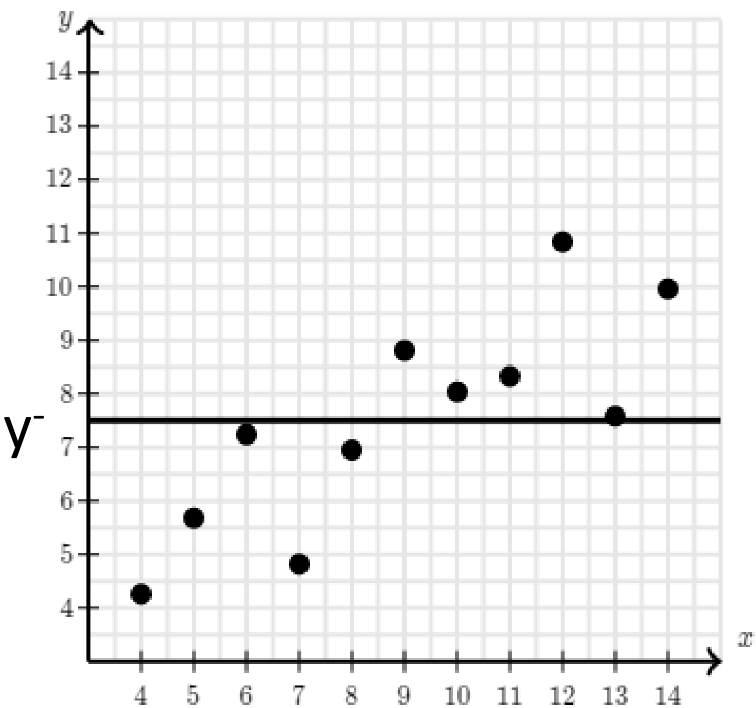
Predicting without Regression

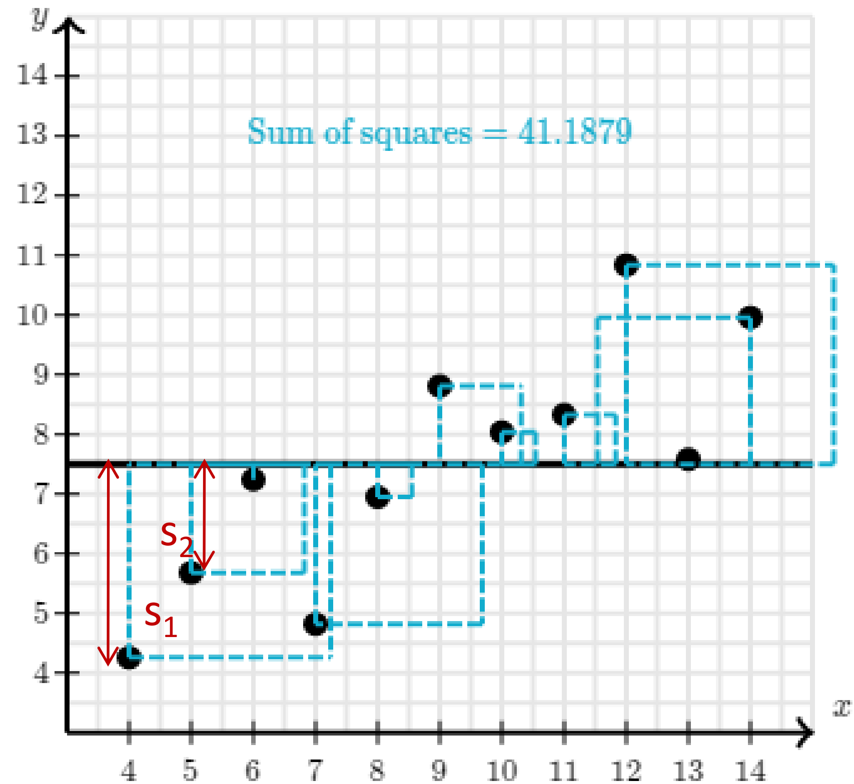
- Without using regression on the x variable, most reasonable estimate is to predict the average of the y values.
- Line shows average value of output



- 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





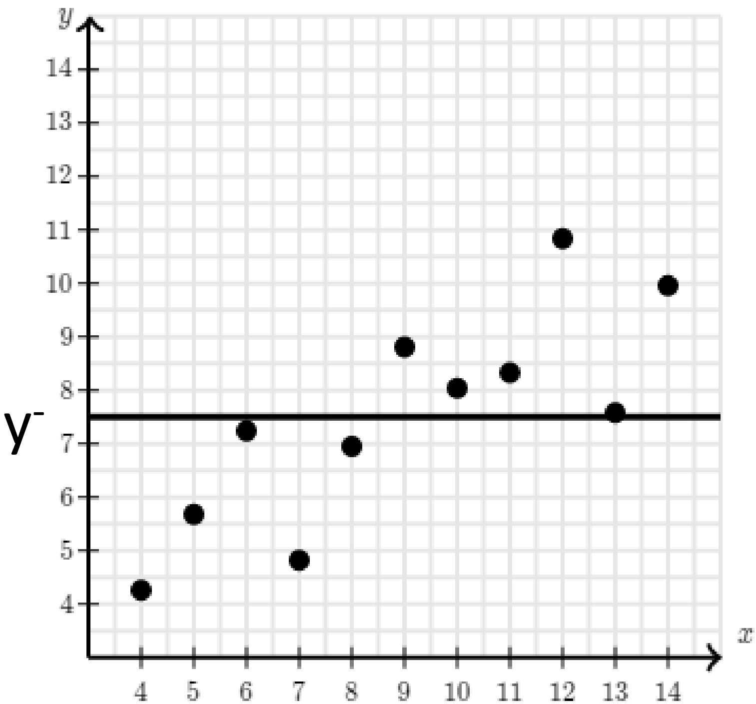
- Prediction Error is Sum of Square, SSR

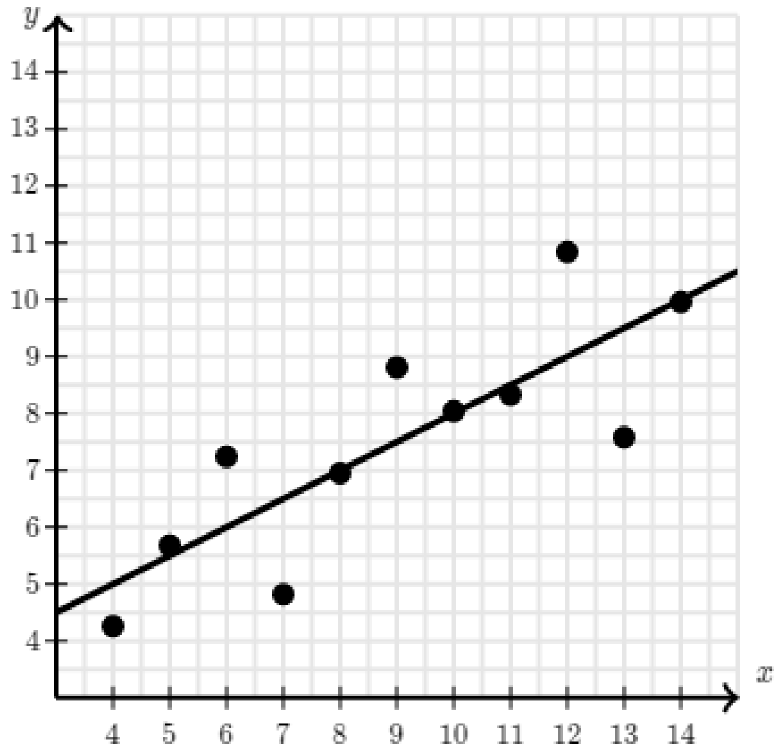
$S_1^2 + S_2^2 + \ldots = 41.1879$

Predicting with Regression

- Regression line is

  $\hat{y} = 0.5x + 1.5$

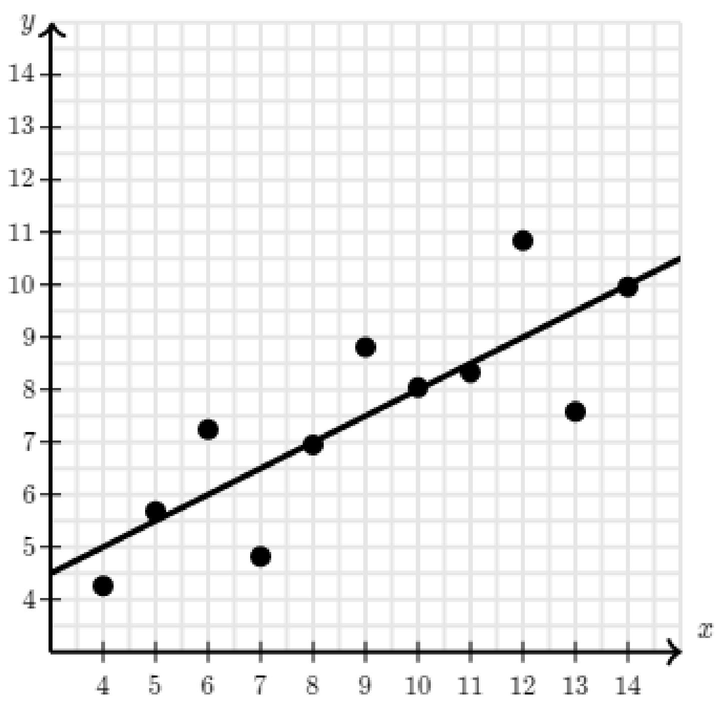


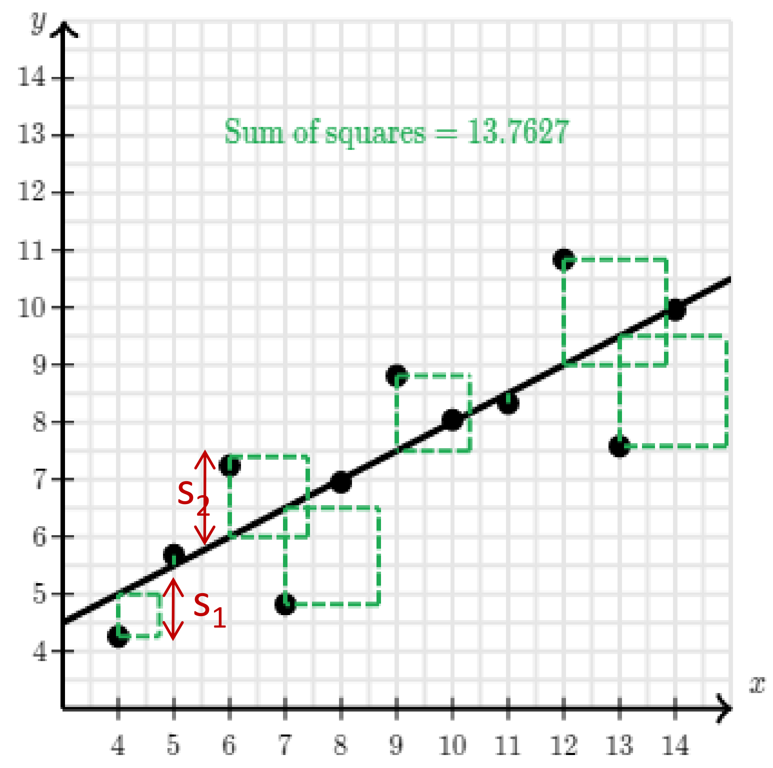


Predicting with Regression

- Prediction Error is Sum of Square, SSR

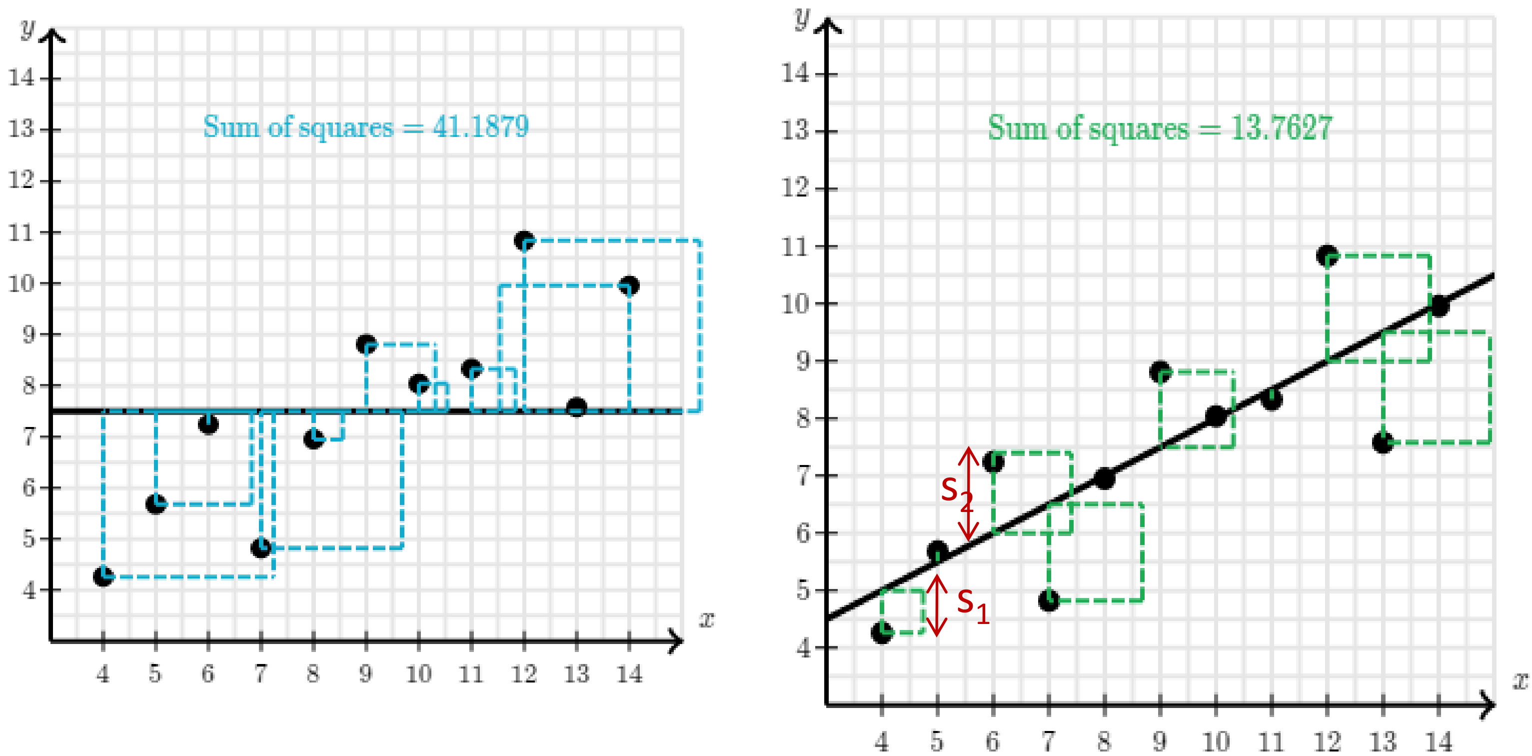
  $S_1^2 + S_2^2 + \ldots = 13.7627$
- Least-squares regression reduces the amount of prediction error





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 1: r-squared

x	y	$\hat{y} = (41/42)x - (5/21)$	Squared error from line $(\hat{Y}-y)^2$	Squared error from mean $(\hat{y} - \bar{y})^2$
-2	-3			
-1	-1			
1	2			
4	3			
	$\bar{y}$ =			

Example 1: r-squared

x	y	y^ = (41/42) x – (5/21)	Squared error from line $(Y\hat{}-y)^2$	Squared error from mean $(y\hat{} - y^{-})^2$
-2	-3	-2.1905		
-1	-1	-1.2143		
1	2	0.7381		
4	3	3.66667		
	y^{-} = 0.25	Total		

Example 1: r-squared

x	y	y^ = (41/42) x – (5/21)	Squared error from line $(Y^\wedge\text{-}y)^2$	Squared error from mean $(y^\wedge - \bar{y})^2$
-2	-3	-2.1905	0.655328798	
-1	-1	-1.2143	0.045918367	
1	2	0.7381	1.592403628	
4	3	3.66667	0.444444444	
	$\bar{y}$ = 0.25	Total	SSE = 2.738095238	

Example 1: r-squared

x	y	y^ = (41/42) x – (5/21)	Squared error from line $(Y\hat{}-y)^2$	Squared error from mean $(y\hat{} - \bar{y})^2$
-2	-3	-2.1905	0.655328798	10.5625
-1	-1	-1.2143	0.045918367	1.5625
1	2	0.7381	1.592403628	3.0625
4	3	3.66667	0.444444444	7.5625
	$\bar{y}$ = 0.25	Total	SSE = 2.738095238	SSR = 22.75

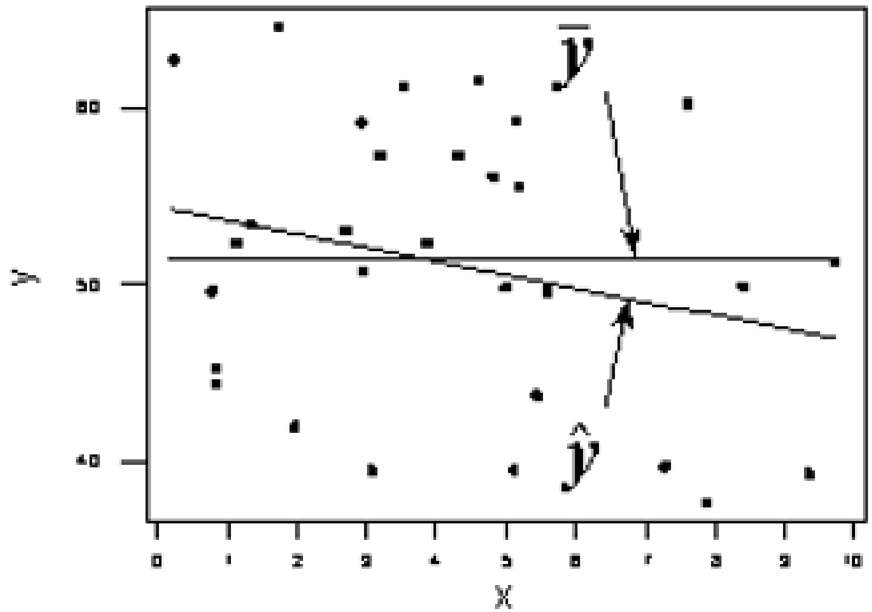
Example 1: r-squared

x	y	y^ = (41/42) x – (5/21)	Squared error from line $(y-y^\wedge)^2$	Squared error from mean $(y^\wedge - \bar{y})^2$
-2	-3	-2.1905	0.655328798	3.76
-1	-1	-1.2143	0.045918367	2.01
1	2	0.7381	1.592403628	0.23
4	3	3.66667	0.444444444	11.68
	$\bar{y}$ = 0.25	Total	SSE = 2.738095238	SSR = 11.68

- SSE = 2.74, SSR = 11.68
- % of total variation not explained by the variation in x, SSE / SSR = 2.74/11.68 = 23.45 = 23.45%
- % of total variation is explained by the variation in x,
- r^2= 1- (SSE/ SSR)= 1- (2.74/22.75) = 0.7655 = 76.55%
- 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, y¯, and estimated regression line, y^
- Slope of the estimated regression line, 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

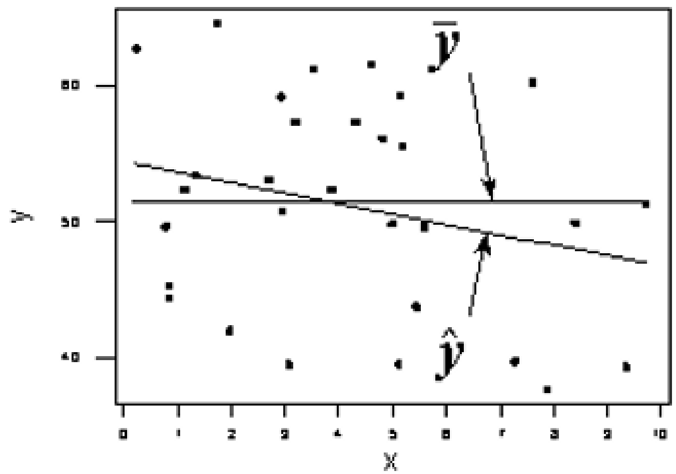


Ex 2: coefficient of determination, r^2

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

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

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



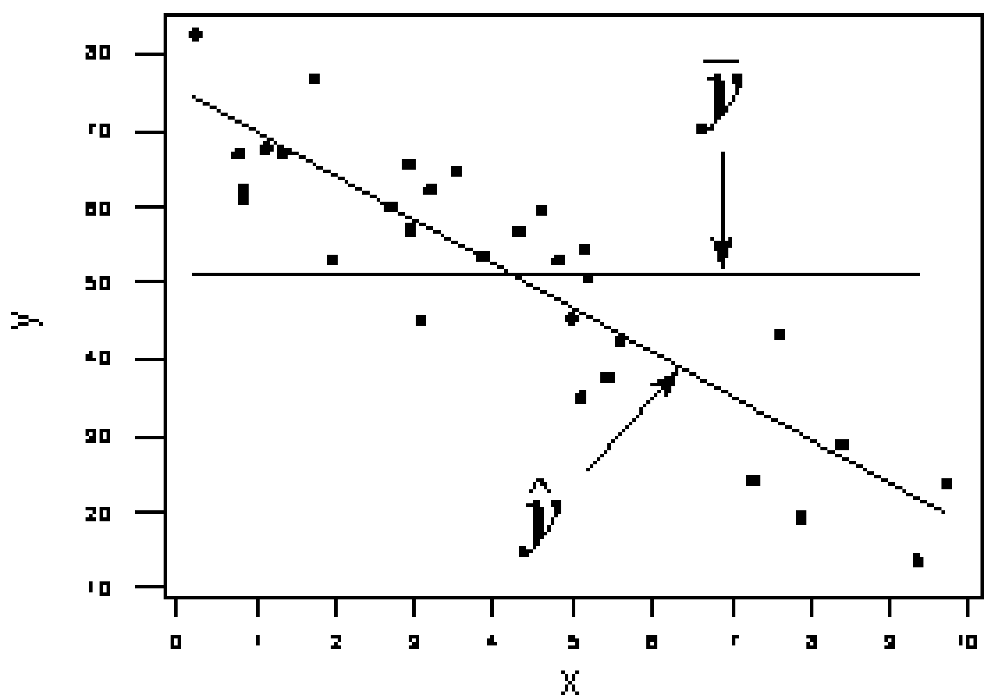
- The sums of squares convey most of the information
- Represent most of the variation in the response *y* (*SSTO* = 1827.6) is

  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

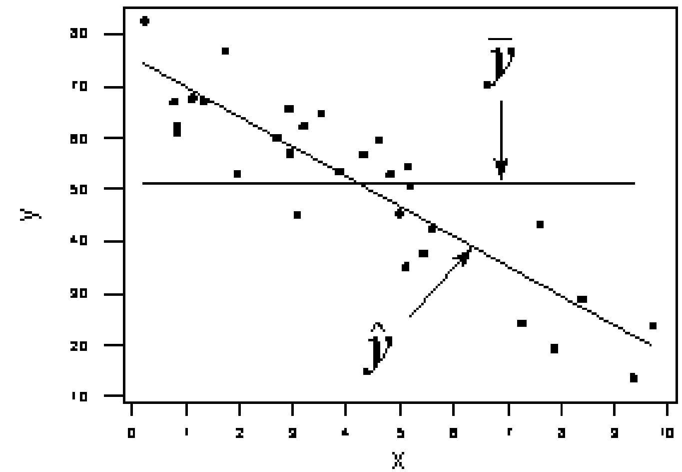


Ex2: coefficient of determination, r^2

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

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

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



- 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<= r^2 <=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

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

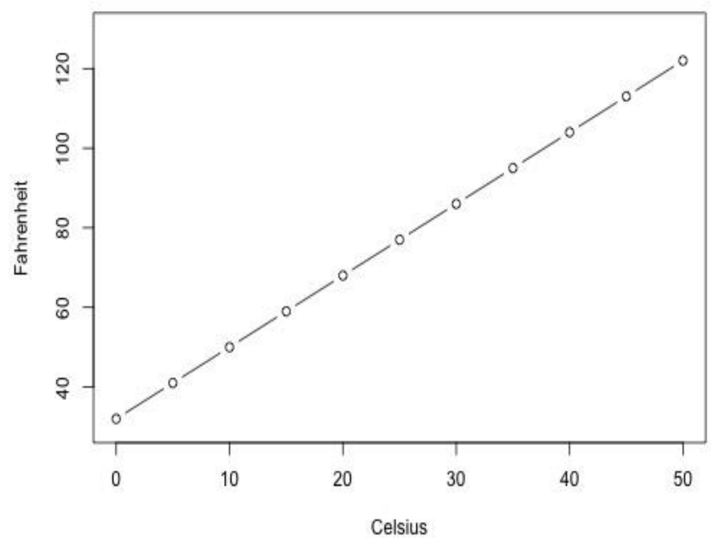
- The estimated slope b_1 of the regression line and the correlation coefficient *r* always share the same sign
- If the estimated slope b_1 of the regression line is 0, then the correlation coefficient *r* must also be 0
- If *r* = -1, then there is a perfect negative linear relationship between *x* 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)

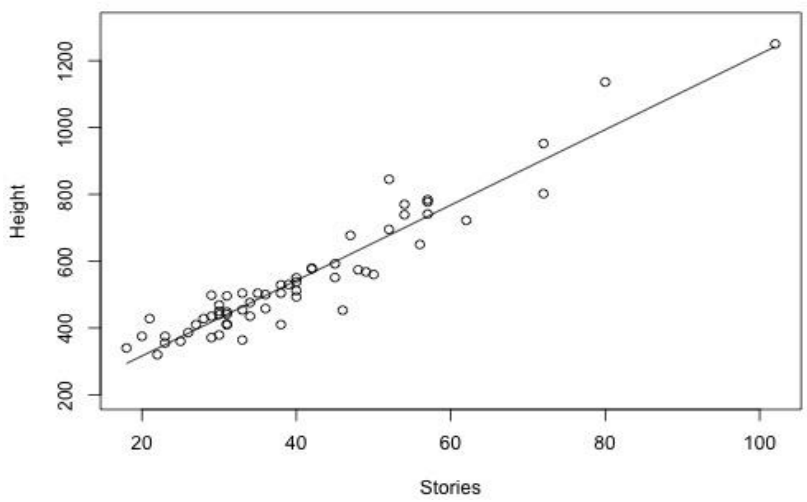
Example 1: r-square and r

- 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



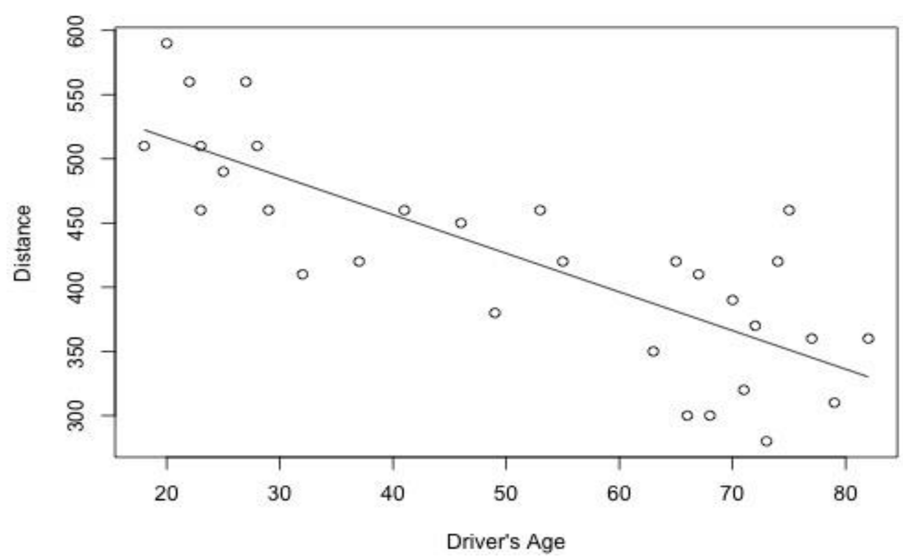
Example 2 r-square and r

- 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



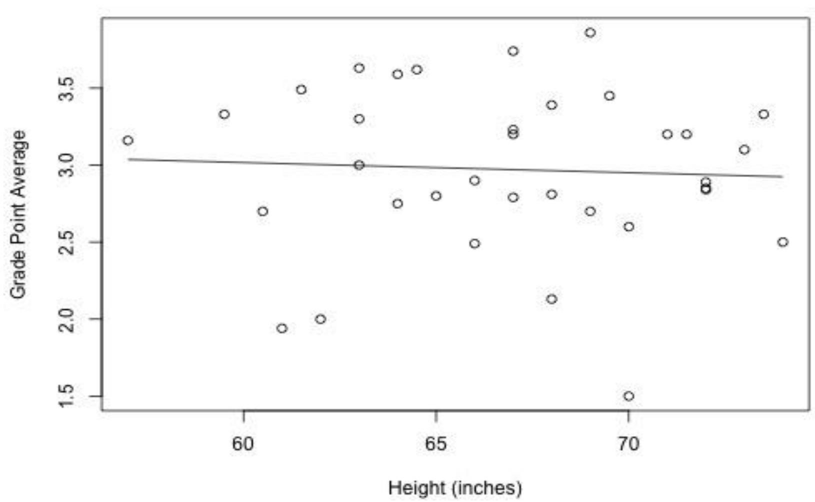
Example 3: r-square and r

- 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



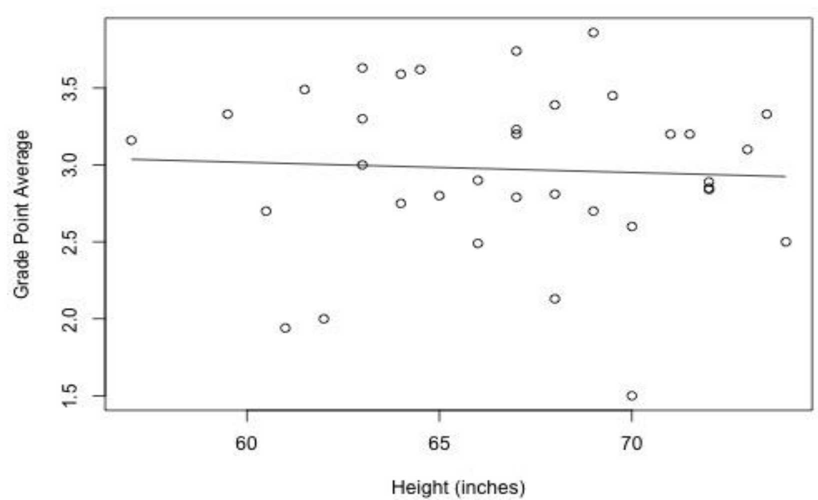
Example 4

- 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



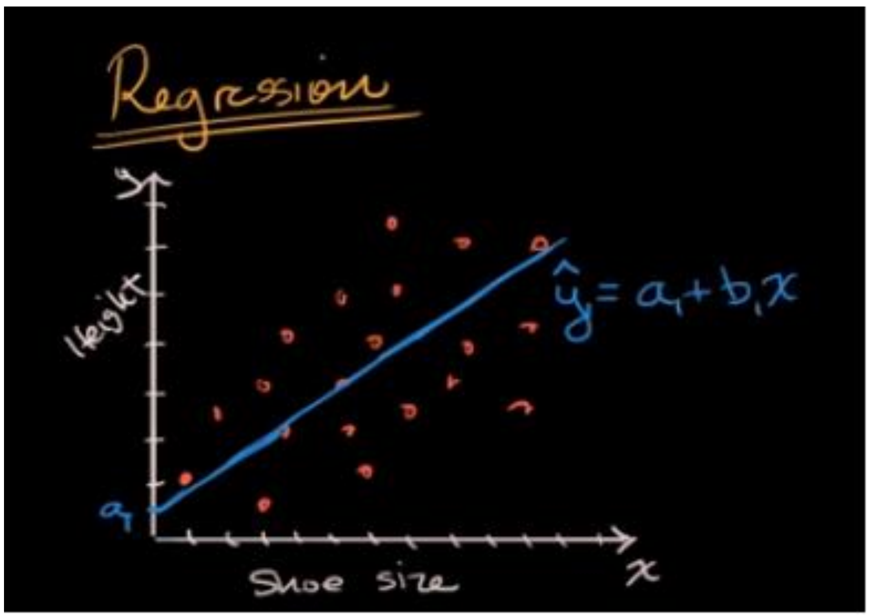
Example 4

- 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

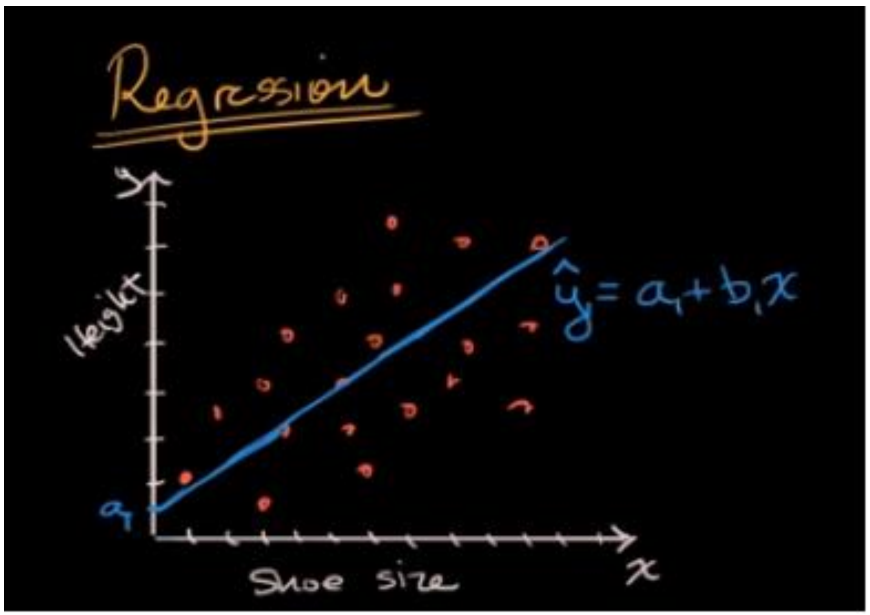
- Regression line for 20 samples



20 samples

Inference about slope

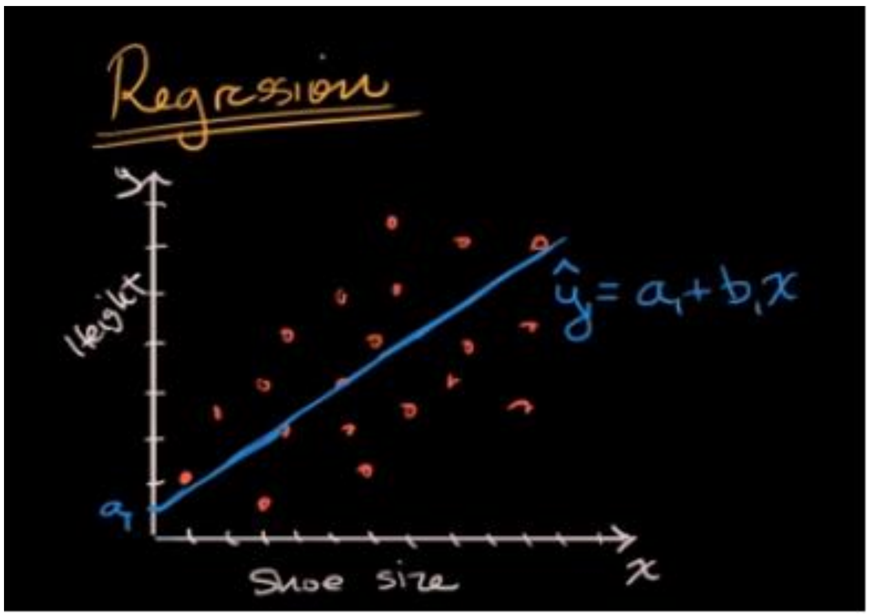
- Regression line for 20 samples
- After adding 20 more samples regression line changes



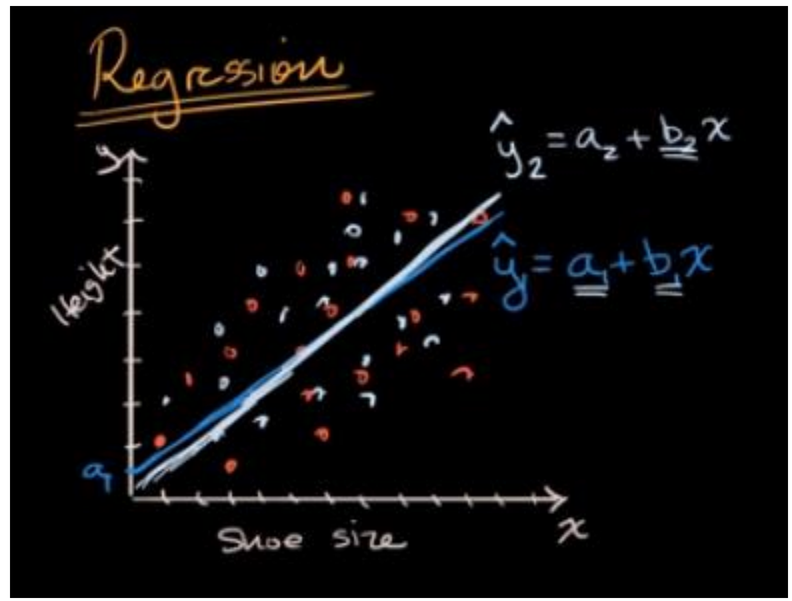
20 samples

Inference about slope

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



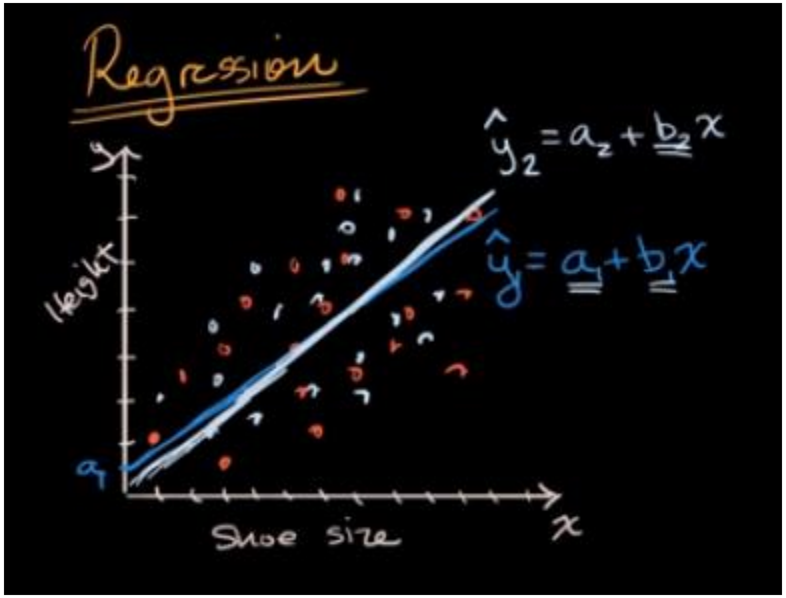
20 samples



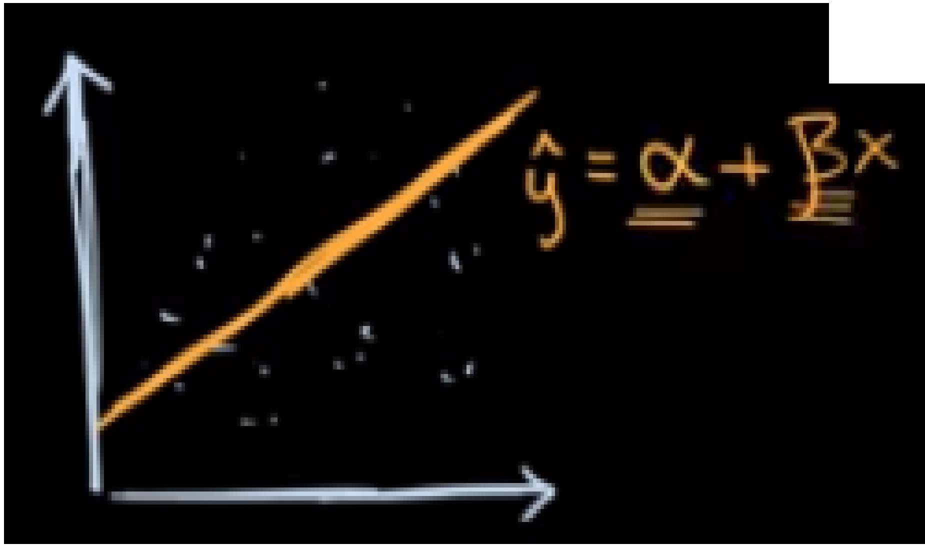
40 samples

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, $C = b_1 \pm t \, SSE_b$
- b_1 is slope and 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, x^{-} and sample standard deviation, s
- Confidence interval will be

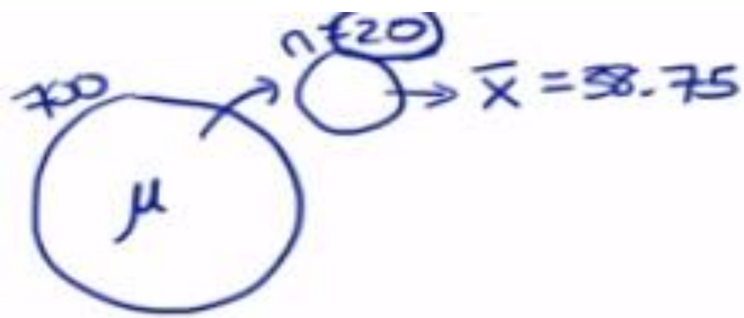
$$\bar{x} \pm t^{*} \frac{s}{\sqrt{n}}$$

- 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

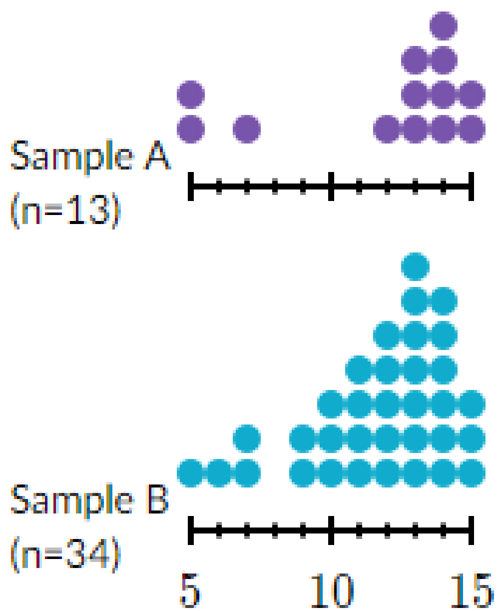
- The sampling distribution of $\bar{x}$, (a sample mean) should be approximately normal
- The shape of the sampling distribution of $\bar{x}$, mostly depends on the shape of the parent population and the sample size, n
- If parent population has normal distribution and sample size, n> 30
- then $\bar{x}$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

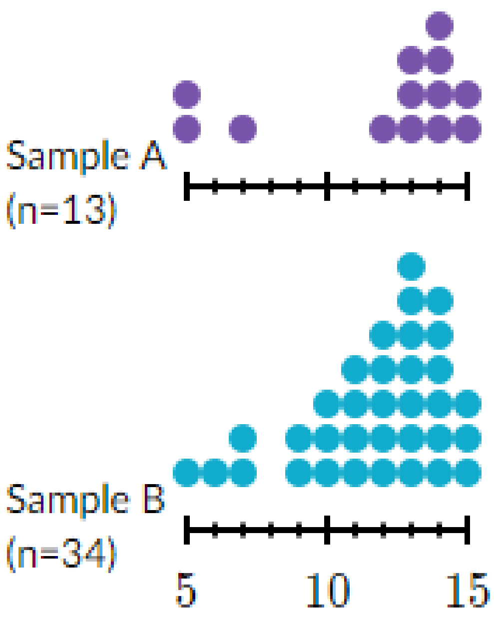
- Two different samples are drawn from two different populations



- 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)

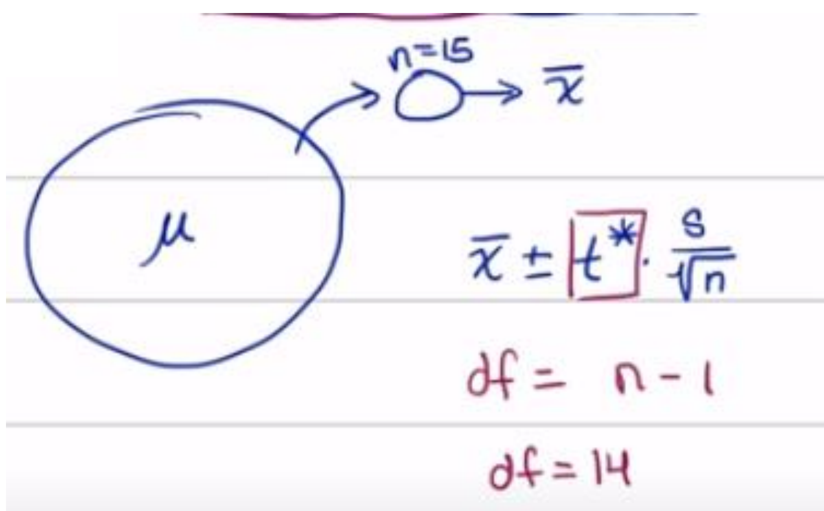
3. The Independence condition

- 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

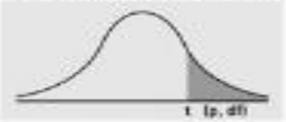
Critical value, t*

- 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



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	43178
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	1.644854	1.95996	2.32635	2.57583	3.2905
CI	———	———	80%	90%	95%	98%	99%	99.9%

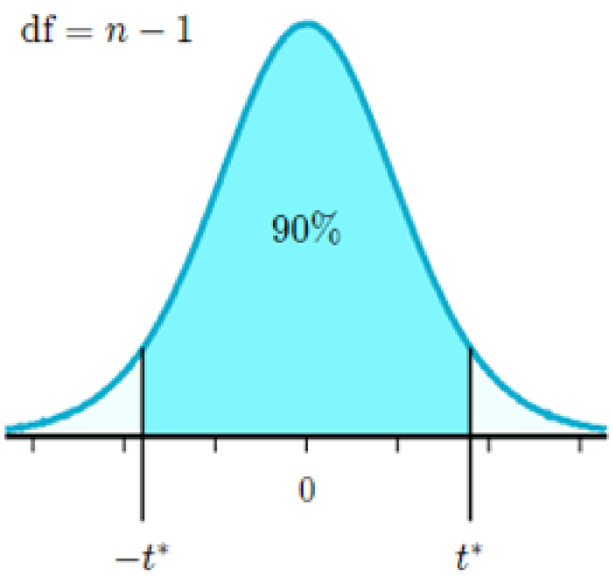
Critical value, t*

- 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	$\boxed{2.624}$	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
∞	.674	.841	1.036	1.282	1.645	1.960	2.054	2.326	2.576	2.807	3.091	3.291
C =	50%	60%	70%	80%	90%	95%	96%	98%	99%	99.5%	99.8%	99.9%

Critical Value of t

- There is a different t-distribution for each sample size, n
- Use t-distribution with degrees of freedom, df <=n
- 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%

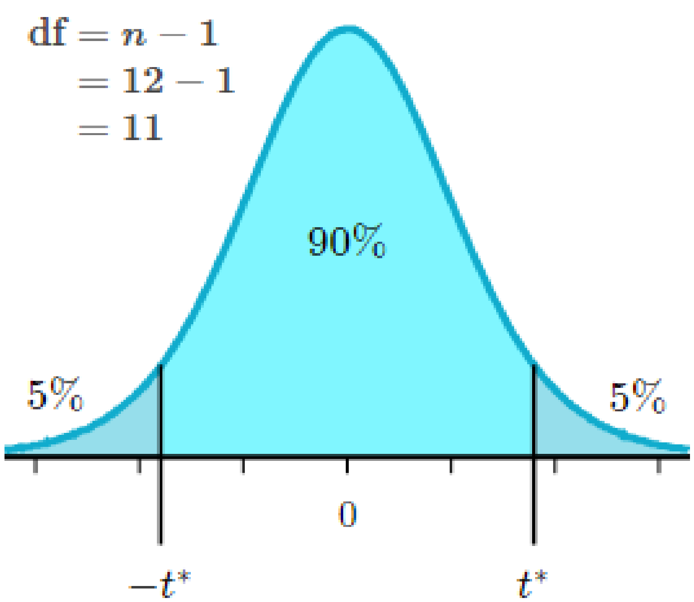


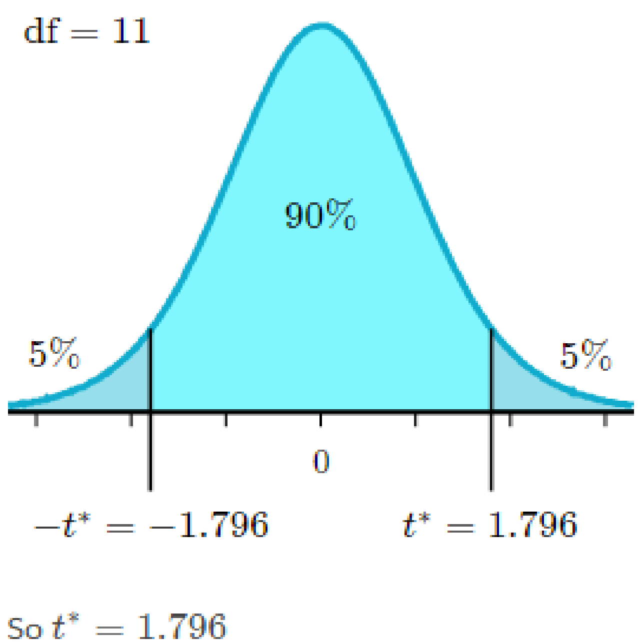
Ex: Critical Value of t (=t*)

- 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 σ = 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*

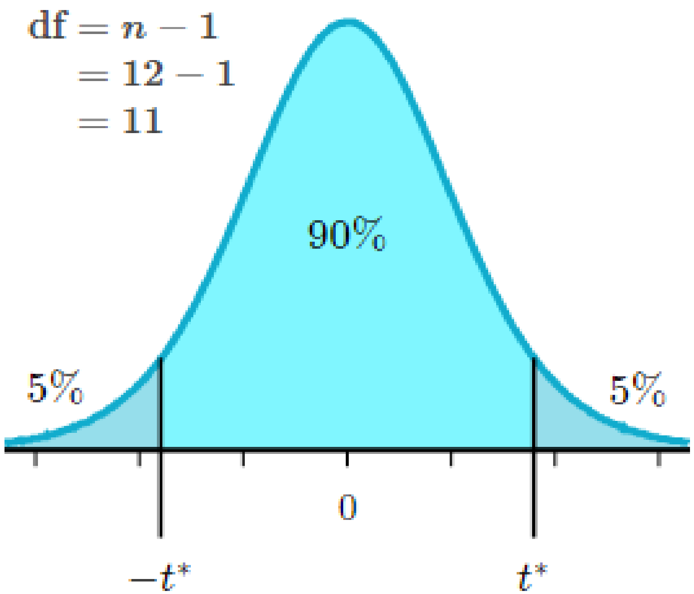
- Determine area remaining in the tails in a t-distribution with df = 12 - 1
- Remaining area= 100% - 90% = 10%
- 10%/2 = 5% per tail





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



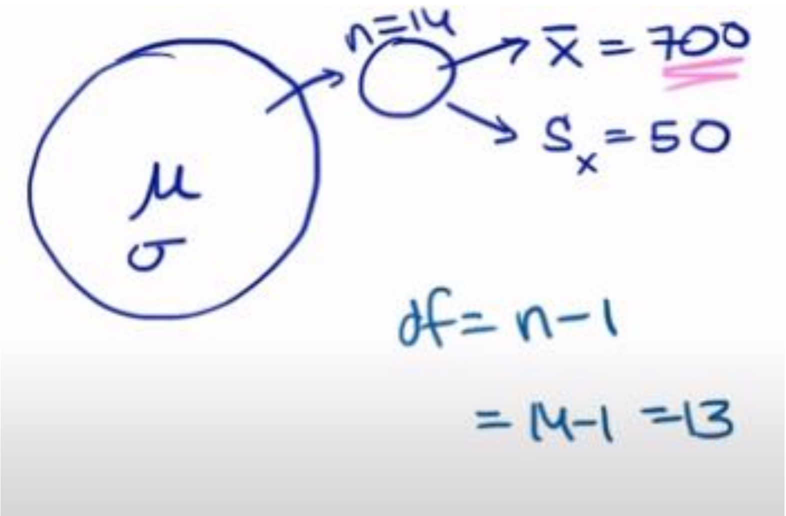
t-distribution

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

Example: t-interval for a mean

- A nutritionist wants to estimate the average caloric content of 14 pizzas and measure their caloric content
- 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 = x⁻ ± (t*s)/sqrt(n)

Example: t-interval for a mean

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

	Tail probability p							2.5%		2.5%		
df	.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.
3	.765	.978	1.250	1.638	2.353	3.182	3.482	4.541	5.841	7.453	10.21	12.
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	6.8
6	.718	.906	1.134	1.440	1.943	2.447	2.612	3.143	3.707	4.317	5.208	5.9
7	.711	.896	1.119	1.415	1.895	2.365	2.517	2.998	3.499	4.029	4.785	5.4
8	.706	.889	1.108	1.397	1.860	2.306	2.449	2.896	3.355	3.833	4.501	5.0
9	.703	.883	1.100	1.383	1.833	2.262	2.398	2.821	3.250	3.690	4.297	4.7
10	.700	.879	1.093	1.372	1.812	2.228	2.359	2.764	3.169	3.581	4.144	4.5
11	.697	.876	1.088	1.363	1.796	2.201	2.328	2.718	3.106	3.497	4.025	4.4
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	$\boxed{2.160}$	2.282	2.650	3.012	3.372	3.852	4.2
14	.692	.868	1.076	1.345	1.761	2.145	2.264	2.624	2.977	3.326	3.787	4.1
15	.691	.866	1.074	1.341	1.753	2.131	2.249	2.602	2.947	3.286	3.733	4.0
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	3.9
18	.688	.862	1.067	1.330	1.734	2.101	2.214	2.552	2.878	3.197	3.611	3.9
19	.688	.861	1.066	1.328	1.729	2.093	2.205	2.539	2.861	3.174	3.579	3.8
20	.687	.860	1.064	1.325	1.725	2.086	2.197	2.528	2.845	3.153	3.552	3.8

Example: t-interval for a mean

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

  $= 700 \pm (2.160^*50)/\text{sqrt}(14)$

  $= 700 \pm 28.9$

  $= 671.1$ to 728.9
- Confidence interval for the mean of a regression line is 671.1 to 728.9