Linear Regression

- Mohamed Imran

Relationship

Y = ?????????

x	Y
2	8
6	20
4	14
3	11
7	23
4	14
2	8
5	17

Relationship

$$Y = 2 + 3(X)$$

X	Y
2	8
6	20
4	14
3	11
7	23
4	14
2	8
5	17

What is 2 here?

$$Y = 2 + 3(X)$$

X	Y
2	8
6	20
4	14
3	11
7	23
4	14
2	8
5	17

Find the Y in ?

$$Y = 2 + 3(X)$$

X	Y
2	8
6	20
4	14
3	11
7	23
4	14
2	8
5	17
10	?
1	?

Value for Y with given X

$$Y = 2 + 3(X)$$

X	Y
2	8
6	20
4	14
3	11
7	23
4	14
2	8
5	17
10	32
1	5

Terminology

$$Y = 2 + 3(X)$$

Y = Model

X	Y
2	8
6	20
4	14
3	11
7	23
4	14
2	8
5	17
10	32
1	5

Terminology

$$Y = 2 + 3(X)$$

Y = Model

2 = Intercept

X	Y
2	8
6	20
4	14
3	11
7	23
4	14
2	8
5	17
10	32
1	5

Terminology

$$Y = 2 + 3(X)$$

Y = Model

2 = Intercept

3 = Slope

X	Y
2	8
6	20
4	14
3	11
7	23
4	14
2	8
5	17
10	32
1	5

Terminology

$$Y = 2 + 3(X)$$

Y = Model

2 = Intercept

3 = Slope

X = input

X	Y
2	8
6	20
4	14
3	11
7	23
4	14
2	8
5	17
10	32
1	5

Formula for a line

Linear regression equation

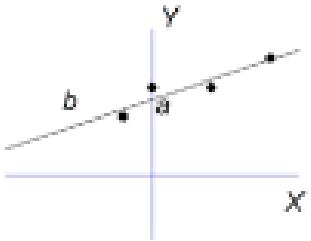
(without error)

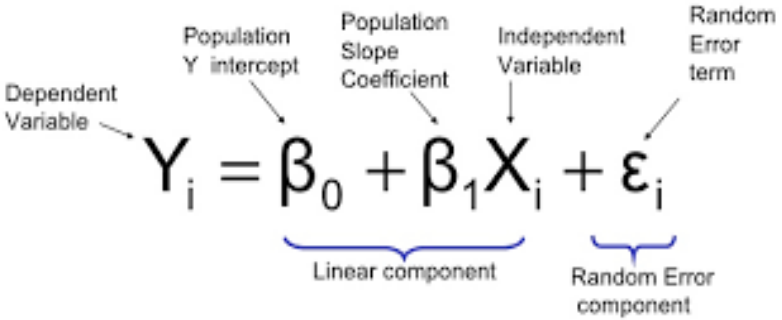
$$\hat{Y} = bX + a$$

predicted values of Y

b = slope = rate of predicted ↑/↓ for Y scores for each unit increase in *X*

Y-intercept = level of Y when X is 0





Linear Regression

Welcome to the world of data science

What is linear?

What is linear?

A Straight line

What is Regression?

What is Regression?

Relationship between two points

What is Linear Regression?

What is Linear Regression?

A Straight line that attempts to predict the relationship between two points

What is Simple Linear Regression?

- Simple Linear Regression is a method used to fit the **best straight line** between a set of data points.
- After a graph is properly scaled, the data points must "look" like they would fit a straight line, not a parabola, or any other shape.
- The line is used as a model in order to predict a variable y from another variable x.
- A regression line must involve 2 variables, the dependent and the independent variable.
- Finding the "best-fit" line is the **goal** of simple linear regression.

Definitions:

Input, Predictive, Or Independent Variable X – 3 names mean the same thing. This is the variable whose value that is believed to influence the value of another variable. This variable should not be dependent on another variable (by definition)

Output, Response, Or Dependent Variable Y – 3 names mean the same thing. This is the variable whose value that is believed to be influenced by the value of another variable. It is by definition, dependent on another variable.

Best-Fit Line – Represents our model. It is the line that "best fits" our data points. The line represents the best estimate of the y value for every given input of x.

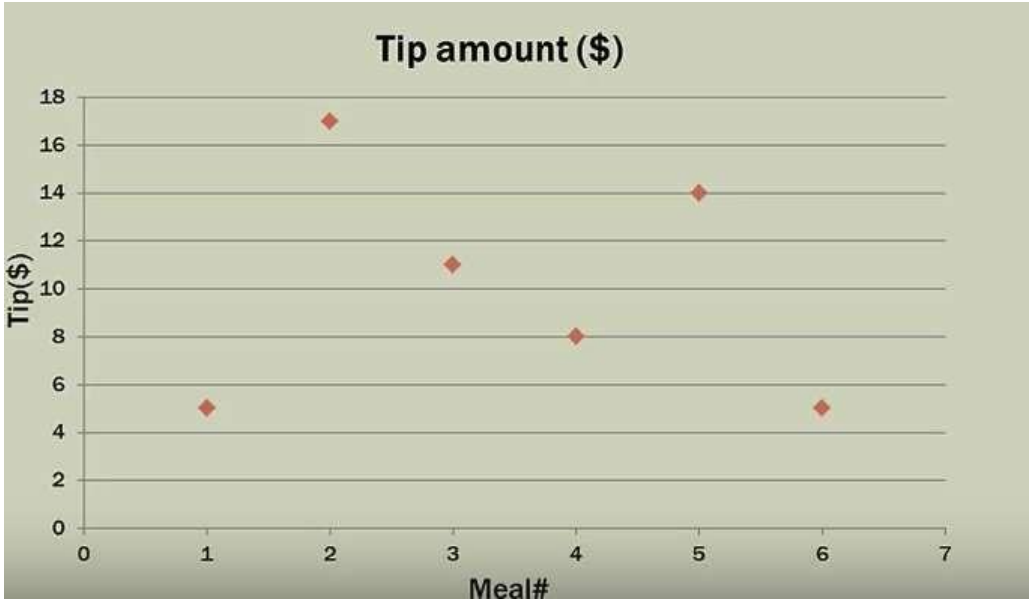
Sum Of Squares – An important calculation we will use to find the best-fit line.

One Variable

- **Problem: A waiter wants to predict his next tip, but he forgot to record the bill amounts for previous tips.**
- **Here is a graph of his tips. The tips is the only variable. Let's call it the y variable.**
- **Meal# is not a variable. It is simply used to identify a tip.**

Meal#	Tip amount ($)
1	5.00
2	17.00
3	11.00
4	8.00
5	14.00
6	5.00

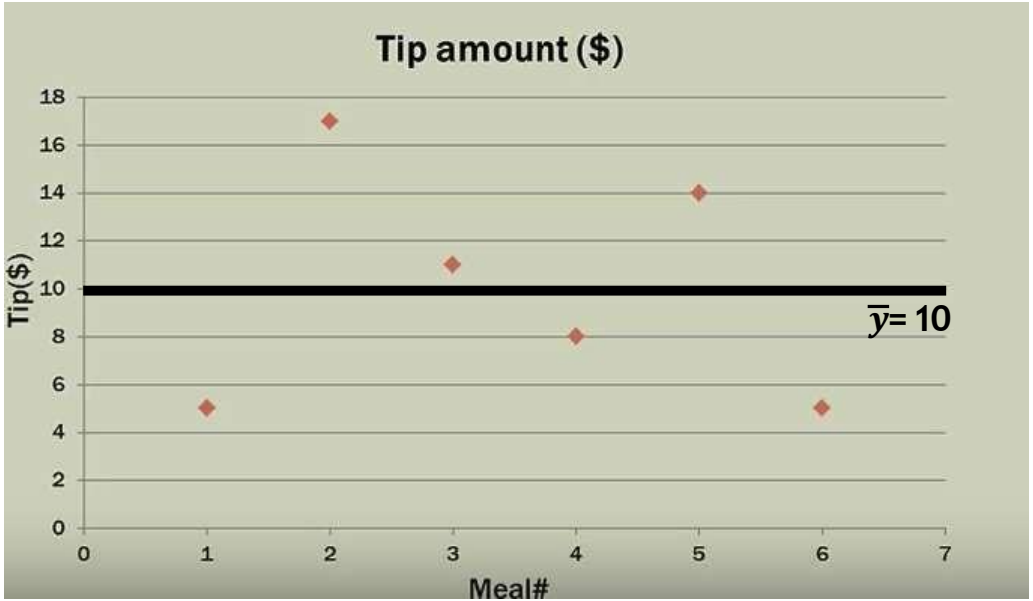
y variable



Can we come up with a model for this problem with only 1 variable?

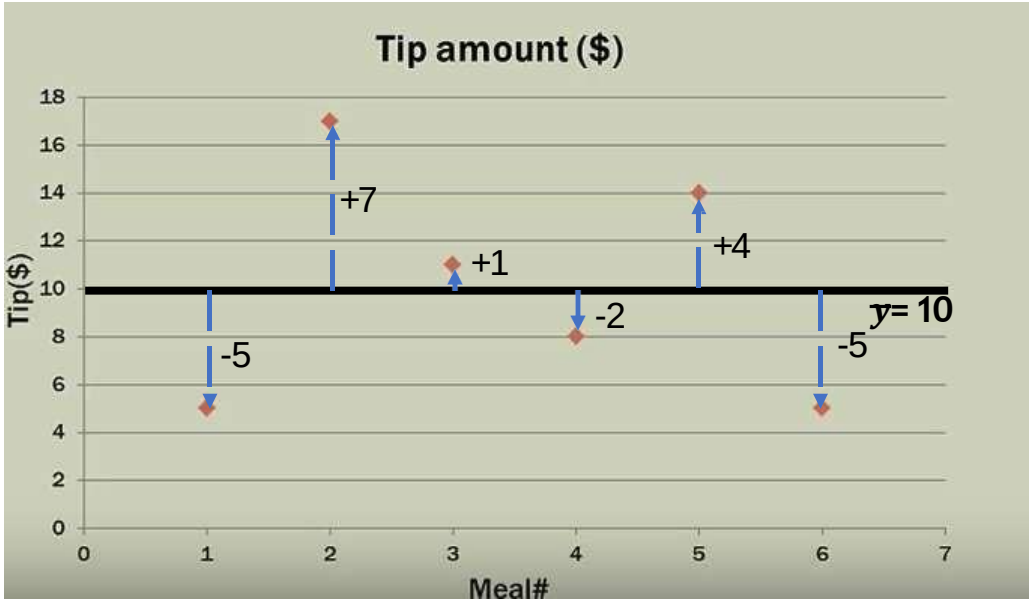
- The only option for our model is to use the mean of the Tips($)
- Tips are on the y access. We would call the mean $\bar{y}$ (y bar).
- The mean for the tip amounts is 10.
- The model for our problem is simply $\bar{y} = 10$.
- $\bar{y} = 10$ is our *best fit line* (represented by bold blackline).

Meal#	Tip amount ($)
1	5.00
2	17.00
3	11.00
4	8.00
5	14.00
6	5.00



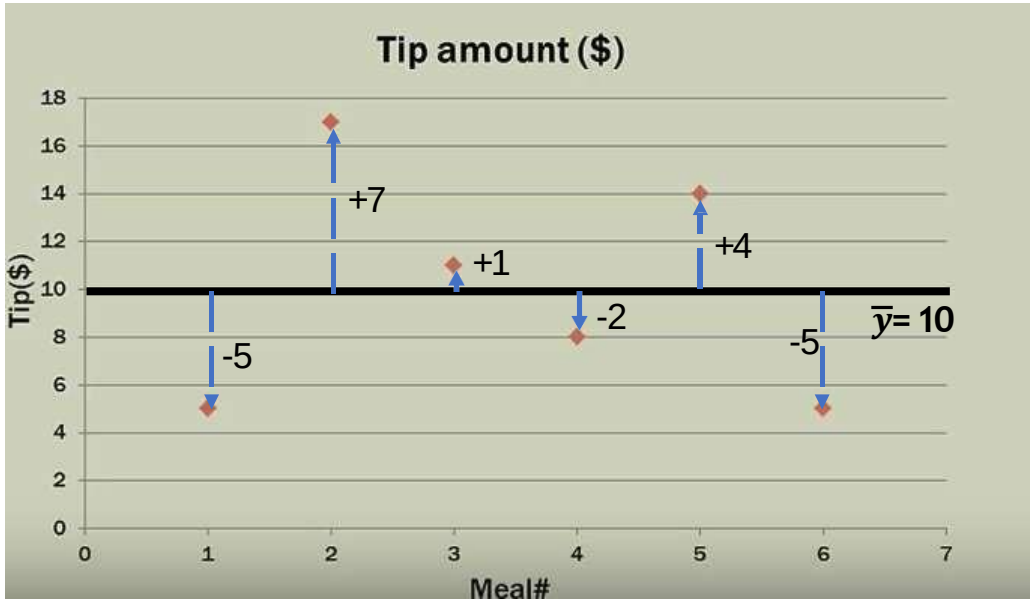
- Now, let's talk about goodness of fit. This will tell us how good our data points fit the line.
- We need to calculate the residuals (errors) for each point.

Meal#	Tip amount ($)
1	5.00
2	17.00
3	11.00
4	8.00
5	14.00
6	5.00



- The best fit line is the one that minimizes the sum of the squares of the residuals (errors).
- The error is the difference between the actual data point and the point on the line.
- SSE (Sum Of Squared Errors) = $(-5)^2 + 7^2 + 1^2 + (-2)^2 + 4^2 + (-5)^2 = 120$

Meal#	Tip amount ($)
1	5.00
2	17.00
3	11.00
4	8.00
5	14.00
6	5.00



- SST (Sum Of Squared Total) = SSR (Sum Of Squared Regression) + SSE is the Sum Of Squares Equation.
- Since there is no regression line (as we only have 1 variable), we can not make the SSE any smaller than 120, because SSR = 0.

Two Variables

Goal Of Simple Linear Regression

- The goal of simple linear regression is to create a **linear model that minimizes the sum of squares of the errors (SSE).**

- From a previous slide: **SST (Sum Of Squared Total) = SSR (Sum Of Squared Regression) + SSE** When we have 2 variables, we can create a regression line; and therefore, we can calculate an SSR > 0. If SSR > 0, then we can reduce SSE. Minimizing the errors means that the line will fit the data better.

- 2 variables: One is the dependent variable: y. The other is the independent variable x.

So now, we need to introduce a little math…..

Remember from the previous slide that we want to **minimize the SSE**.

We write this mathematically this way.:

$$\min \sum (y_i - \hat{y}_i)^2$$

y_i = observed value of dependent variable (tip amount)

$\hat{y}_i$ = estimated(predicted)value of the dependent variable (predicted tip amount)

y is often referred to as y-hat.

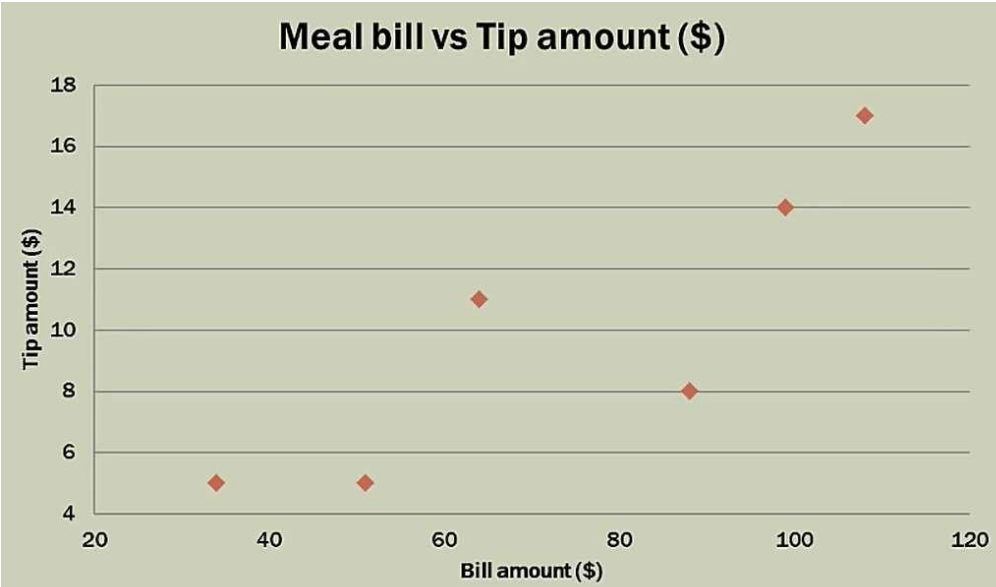
- **Repeating the Problem: As a waiter, how do we predict the tips we will receive for service rendered?**
- **Let's say, we didn't forget to record the bill amount.**

Independent Variable (x)

Dependent Variable (y)

Total bill ($)	Tip amount ($)
34.00	5.00
108.00	17.00
64.00	11.00
88.00	8.00
99.00	14.00
51.00	5.00

If we scale the graph according to the data points available, we can then plot the points.

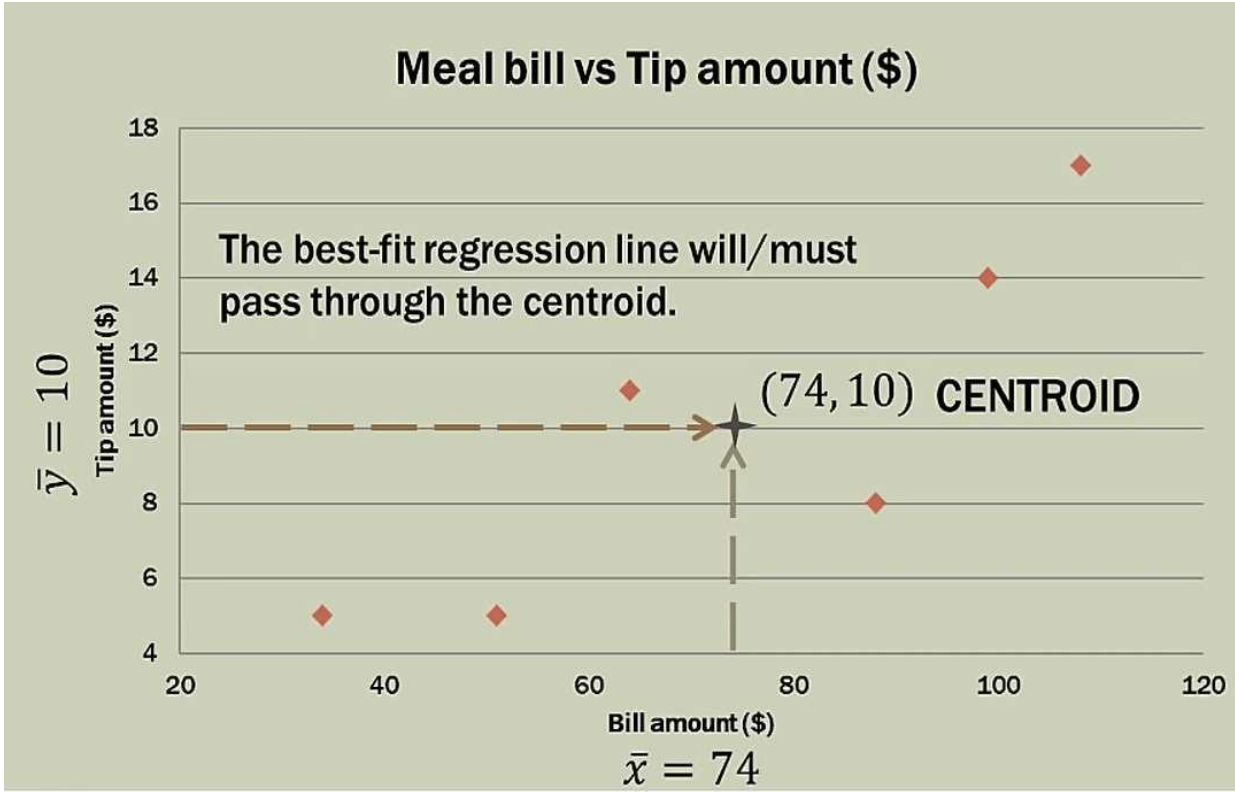


Bill ($)	Tip ($)
34.00	5.00
108.00	17.00
64.00	11.00
88.00	8.00
99.00	14.00
51.00	5.00

Does the data seem to fall along a line?

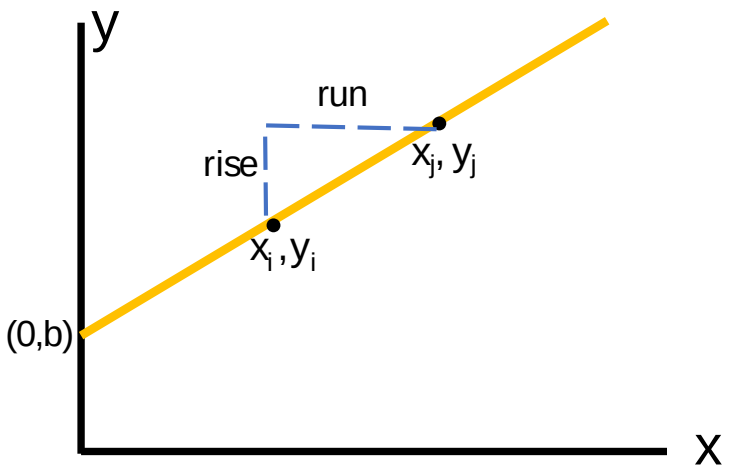
In this case, YES! Proceed.

If not…if it's a BLOB with no linear pattern, then stop.



Bill ($)	Tip ($)
34.00	5.00
108.00	17.00
64.00	11.00
88.00	8.00
99.00	14.00
51.00	5.00
$\bar{x} = 74$	$\bar{y} = 10$

Algebra Of Lines - Slope



Equation Of A Straight Line

$$Y = mX + b$$

slope, y-intercept

$m = (\text{rise/run}) = (y_j - y_i) / (x_j - x_i)$

b is the point where $x = 0$ and y intersects the y-axis

Regression Line - Slope

$$\hat{y}_i = b_0 + b_1 x_i$$

Slope

The formula for the slope, m, of the best-fitting line is

$$m = r\left(\frac{s_y}{s_x}\right)$$

where r is the correlation between X and Y, and s_x and s_y are the standard deviations of the x-values and the y-values, respectively. You simply divide s_y by s_x and multiply the result by r.

Think of s_y divided by s_x as the variation (resembling change) in Y over the variation in X, in units of X and Y. Rise Over Run!

Derivation to this

$$b_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$$

$\bar{x}$ = mean of the independent variable

x_i = value of independent variable

$\bar{y}$ = mean of the dependent variable

y_i = value of dependent variable

Let's create a table of calculations that we can use to calculate the slope of the regression (best-fit) line.

Meal	Total bill ($)	Tip amount ($)	Bill deviation	Tip Deviations	Deviation Products	Bill Deviations Squared
	x	y	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$
1	34	5	-40	-5	200	1600
2	108	17	34	7	238	1156
3	64	11	-10	1	-10	100
4	88	8	14	-2	-28	196
5	99	14	25	4	100	625
6	51	5	-23	-5	115	529
	$\bar{x} = 74$	$\bar{y} = 10$			$\sum = 615$	$\sum = 4206$

$$b_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$$

$$b_1 = \frac{615}{4206}$$

$$b_1 = 0.1462$$

Deviation Products	Bill Deviations Squared
$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$
200	1600
238	1156
-10	100
-28	196
100	625
115	529
$\sum = 615$	$\sum = 4206$

- Now that we have the slope, we can calculate the y-intercept because we know 1 point on the line already (74,10).
- What do we call 74,10?

$$b_0 = \bar{y} - b_1\bar{x} \qquad b_1 = 0.1462$$

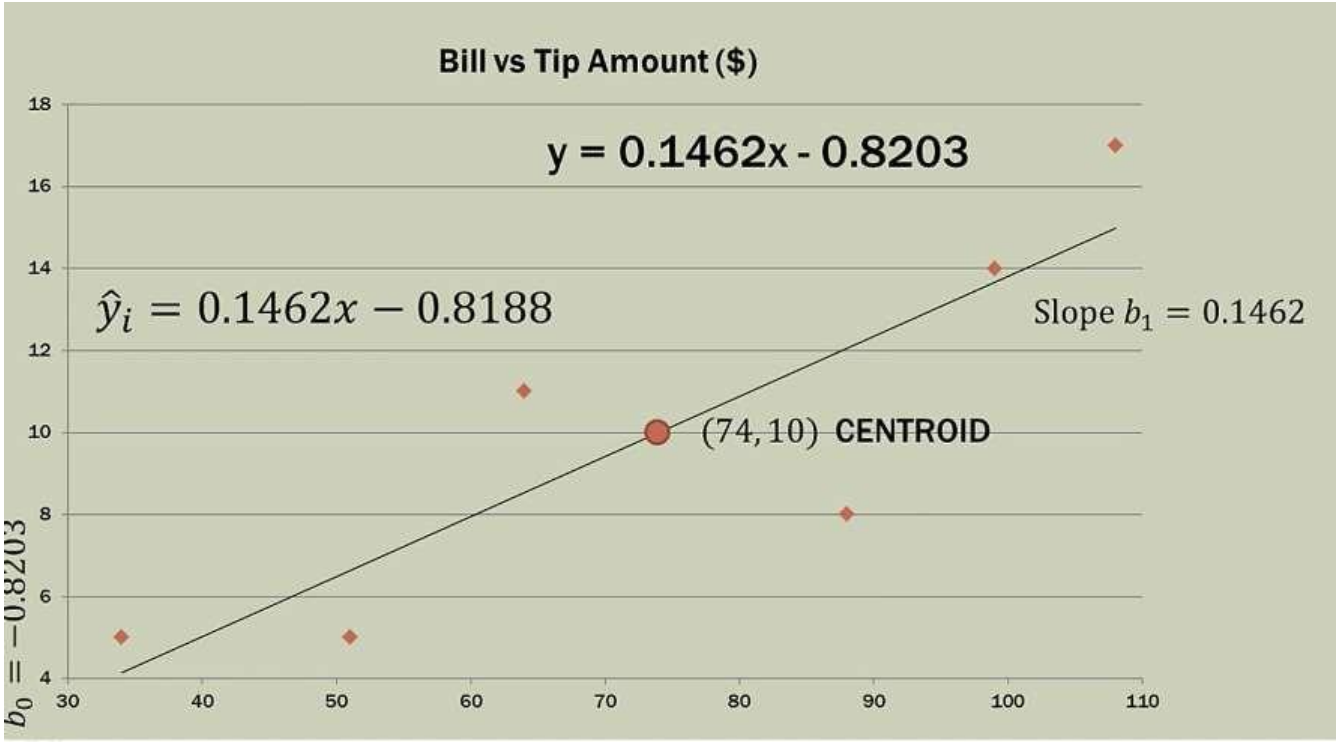
$$b_0 = 10 - 0.1462(74)$$

$$b_0 = 10 - 10.8188$$

$$b_0 = -0.8188$$

Total bill ($)	Tip amount ($)
x	y
34	5
108	17
64	11
88	8
99	14
51	5
$\bar{x} = 74$	$\bar{y} = 10$

- **(74,10) is the Centroid.**
- **For comparison, Excel has calculated the regression equation very close to our manual calculation.**



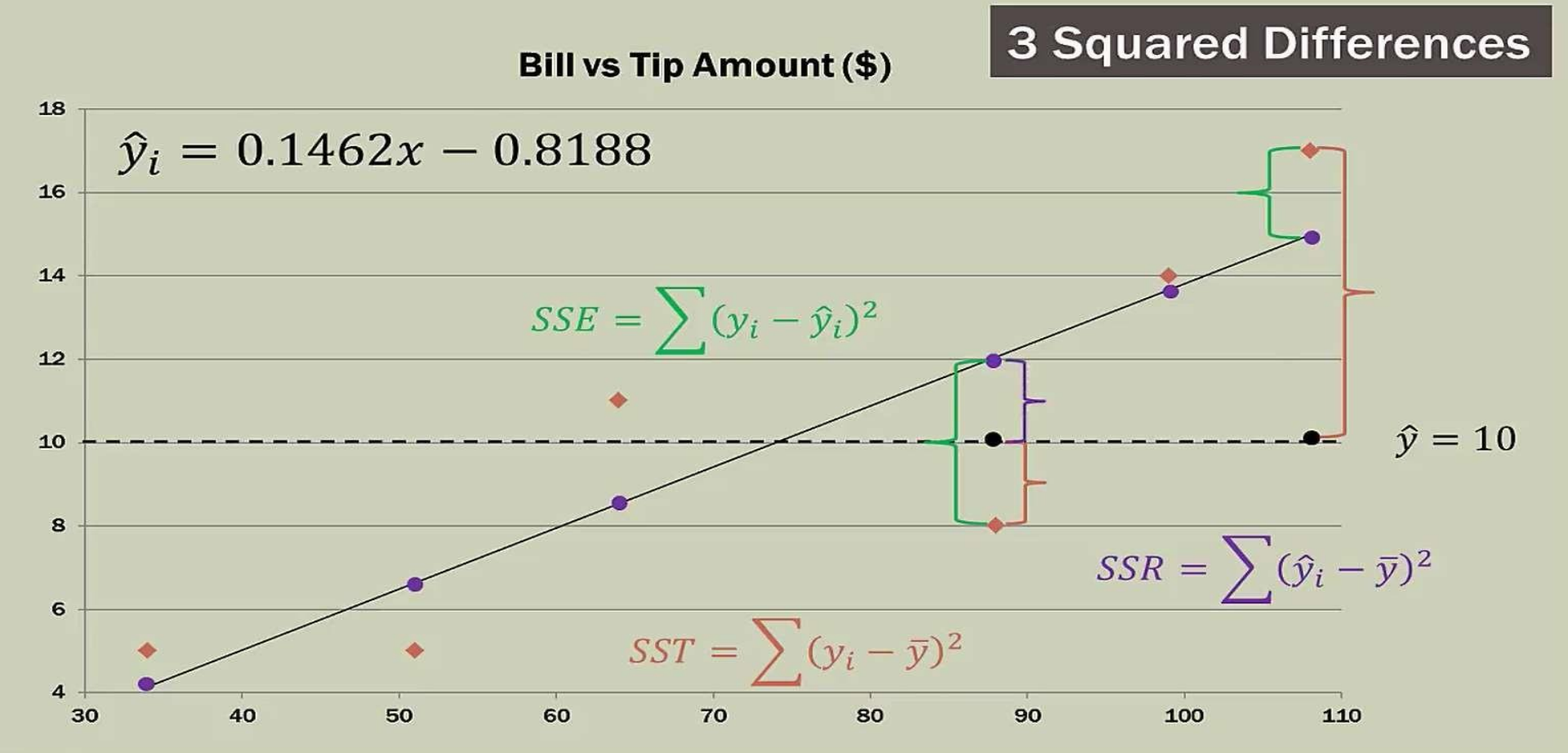
Meaning of our equation….

$$\hat{y}_i = 0.1462x - 0.8188$$

For every \$1 the bill amount (x) increases, we would expect the tip amount to increase by \$0.1462 or about 15-cents.

If the bill amount (x) is zero, then the expected/predicted tip amount is \$-0.8188 or negative 82-cents! Does this make sense? NO. The intercept may or may not make sense in the "real world."

SST = SSR + SSE


3 Squared Differences
Bill vs Tip Amount ($)
$\hat{y}_i = 0.1462x - 0.8188$
$SSE = \sum (y_i - \hat{y}_i)^2$
$\hat{y} = 10$
$SSR = \sum (\hat{y}_i - \bar{y})^2$
$SST = \sum (y_i - \bar{y})^2$
18
16
14
12
10
8
6
4
30
40
50
60
70
80
90
100
110

Meal	Total bill ($)	Observed tip amount ($)	$\hat{y}_i$ (predicted tip amount)	Error $(y - \hat{y}_i)$	Squared Error $(y - \hat{y}_i)^2$
	x	y			
1	34	5	4.1505	0.8495	0.7217
2	108	17	14.9693	2.0307	4.1237
3	64	11	8.5365	2.4635	6.0688
4	88	8	12.0453	−4.0453	16.3645
5	99	14	13.6535	0.3465	0.1201
6	51	5	6.6359	−1.6359	2.6762
	$\bar{x} = 74$	$\bar{y} = 10$		SSE=	$\sum = 30.075$

General Regression using Linear Algebra

- Find the best-fitting line for the data points

x	y
1	2
2	3
4	7
5	5
7	11

Solve the following system of linear equations for x, y, and z:

$$\begin{aligned} x &+ 2y &+ z &= 12 \\ 2x &- y &+ z &= 1 \\ x &+ y &- 3z &= -4 \end{aligned} .$$

While there are many ways to solve these types of systems, one of special interest is by treating the three lines as a linear transformation and looking at its corresponding matrix:

$$\begin{aligned} x &+ 2y &+ z &= 12 \\ 2x &- y &+ z &= 1 \\ x &+ y &- 3z &= -4 \end{aligned} \implies \begin{pmatrix} 1 & 2 & 1 \\ 2 & -1 & 1 \\ 1 & 1 & -3 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 12 \\ 1 \\ -4 \end{pmatrix}.$$

Then, $\begin{pmatrix} x \\ y \\ z \end{pmatrix}$ is an element of the vector space $\mathbb{R}^3$, and the matrix describes a linear transformation from $\mathbb{R}^3$ to itself. Finding the matrix's inverse then yields the answer:

$$\begin{pmatrix} \frac{2}{19} & \frac{7}{19} & \frac{3}{19} \\ \frac{7}{19} & -\frac{4}{19} & \frac{1}{19} \\ \frac{3}{19} & \frac{1}{19} & -\frac{5}{19} \end{pmatrix} \cdot \begin{pmatrix} 1 & 2 & 1 \\ 2 & -1 & 1 \\ 1 & 1 & -3 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} \frac{2}{19} & \frac{7}{19} & \frac{3}{19} \\ \frac{7}{19} & -\frac{4}{19} & \frac{1}{19} \\ \frac{3}{19} & \frac{1}{19} & -\frac{5}{19} \end{pmatrix} \cdot \begin{pmatrix} 12 \\ 1 \\ -4 \end{pmatrix}$$

$$\Rightarrow \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 1 \\ 4 \\ 3 \end{pmatrix}. \square$$

https://www.mathsisfun.com/algebra/matrix-inverse.html

Solving with Linear algebra

The previous example can be rewritten in matrix language: we seek a least-squares approximation to the equation

$$\begin{pmatrix} 1 & 1 \\ 2 & 1 \\ 4 & 1 \\ 5 & 1 \\ 7 & 1 \end{pmatrix} \begin{pmatrix} m \\ b \end{pmatrix} = \begin{pmatrix} 2 \\ 3 \\ 7 \\ 5 \\ 11 \end{pmatrix}.$$

This equation has no solutions (since no line goes through all five points), but the least squares solution is given by multiplying both sides by A^T and solving

$$\begin{pmatrix} 1 & 2 & 4 & 5 & 7 \\ 1 & 1 & 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 2 & 1 \\ 4 & 1 \\ 5 & 1 \\ 7 & 1 \end{pmatrix} \begin{pmatrix} m \\ b \end{pmatrix} = \begin{pmatrix} 1 & 2 & 4 & 5 & 7 \\ 1 & 1 & 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} 2 \\ 3 \\ 7 \\ 5 \\ 11 \end{pmatrix}$$

$$\begin{pmatrix} 95 & 19 \\ 19 & 5 \end{pmatrix} \begin{pmatrix} m \\ b \end{pmatrix} = \begin{pmatrix} 138 \\ 28 \end{pmatrix},$$

which is the same system of equations we got by taking partial derivatives, and leads again to the unique solution $m = \frac{79}{57}$ and $b = \frac{1}{3}$. $\square$

https://www.mathsisfun.com/algebra/matrix-inverse-row-operations-gauss-jordan.html

Gradient Descent

Linear Regression with Gradient Descent

- **Gradient descent** is a first-order iterative optimization algorithm for finding the minimum of a function.

Linear Regression with Gradient Descent

- **Gradient descent** is a first-order **iterative** optimization algorithm for finding the minimum of a function.

Linear Regression with Gradient Descent

- **Gradient descent** is a first-order **iterative optimization** algorithm for finding the minimum of a function.

Linear Regression with Gradient Descent

- **Gradient descent** is a first-order **iterative optimization** algorithm for finding the **minimum** of a **function**.

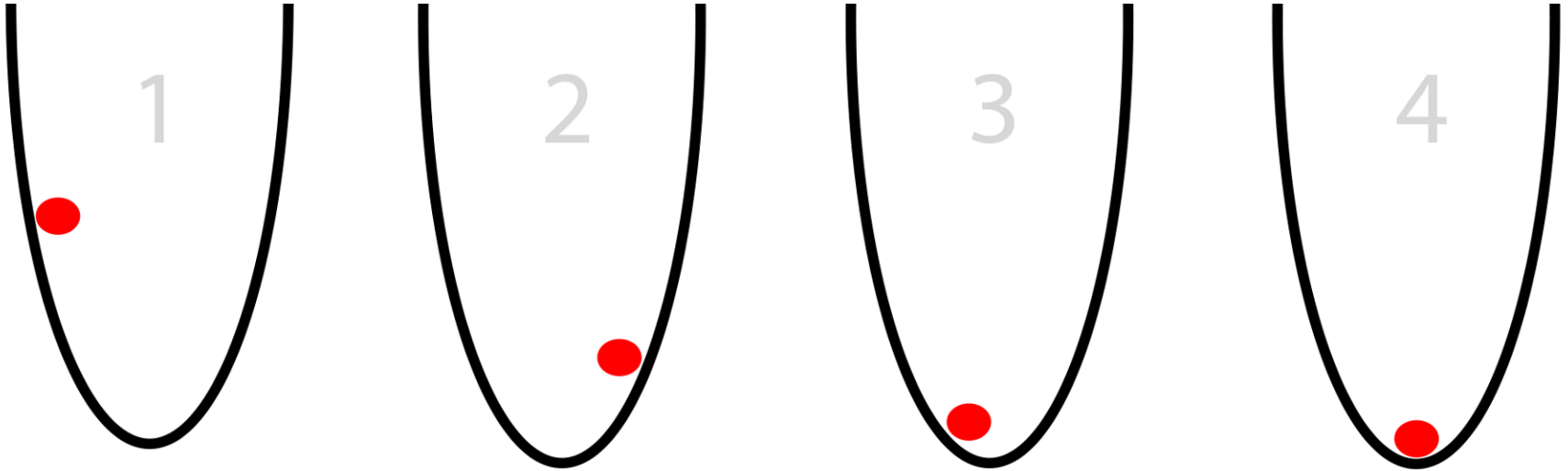
Linear Regression with Gradient Descent

- By **tweaking *m* and *b***, we can create a line that will best describe the relationship. How do we know we're close? By using a thing called a **cost function**. It literally tells us the cost.
- A high cost value means it's expensive—our approximation is far from describing the real relationship. On the other hand, a **low cost value** means it's **cheap**—our approximation is close to describing the relationship.

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^{n} (y_i - \tilde{y}_i)^2$$

Linear Regression with Gradient Descent

- **Brute force isn't helpful**. A more efficient way is gradient descent. Imagine trying to find the lowest point blindfolded as can be seen below. What you would do is to check left and right and then feel which one brings you to a lower point. You do this every step of the way until checking left and right both brings you to a higher point.



Linear Regression with Gradient Descent

- What we're doing here is applying **partial derivatives** with respect to both *m* and *b* to the **cost function** to point us to the lowest point.
- derivative of zero means you are at either a local minima or maxima
- Which means that the closer we get to zero, **the better**.

$$\frac{\partial}{\partial \mathrm{m}} = \frac{2}{N}\sum_{i=1}^{N} - x_i(y_i - (mx_i + b))$$

$$\frac{\partial}{\partial \mathrm{b}} = \frac{2}{N}\sum_{i=1}^{N} - (y_i - (mx_i + b))$$

What is partial derivative?

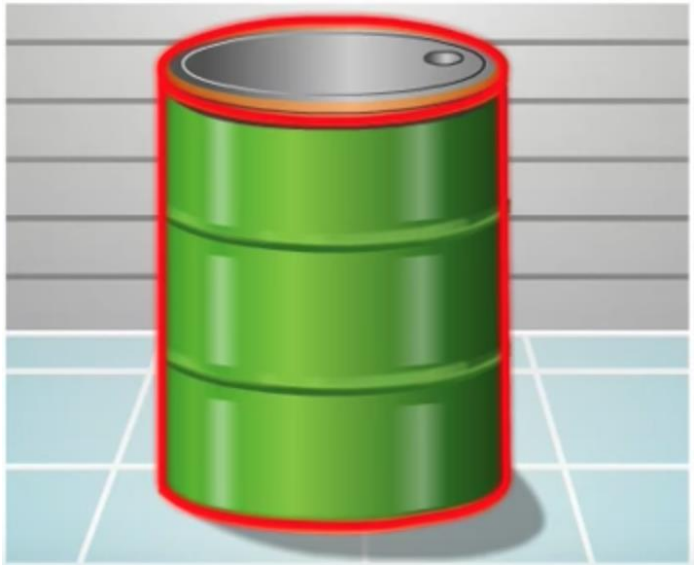
- Let us understand the function of several variables
- Volume of cylinder

$$V = \pi r^2 h$$

Where,

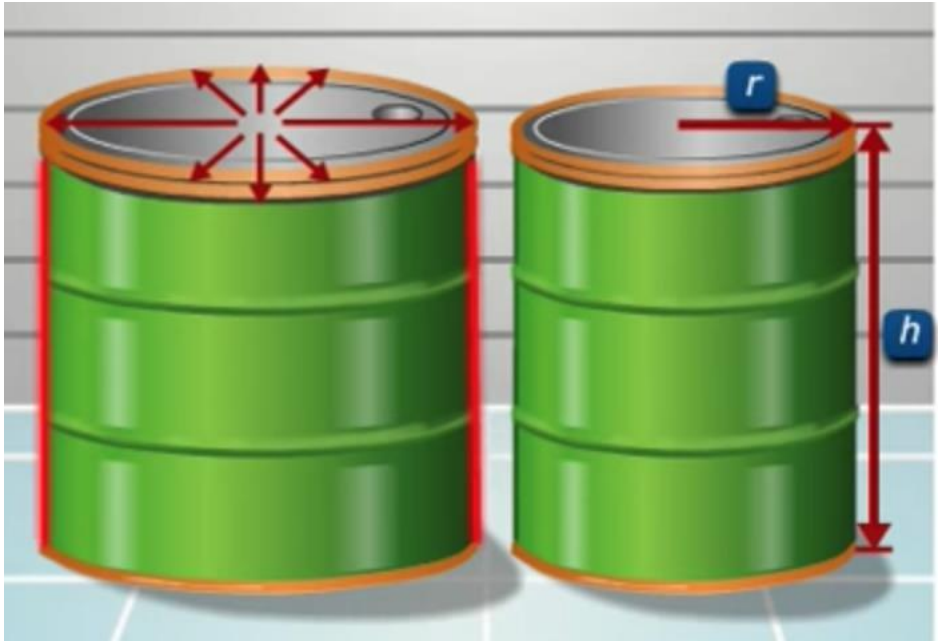
r is the radius of the cylinder

h is the height of the cylinder



What is partial derivative?

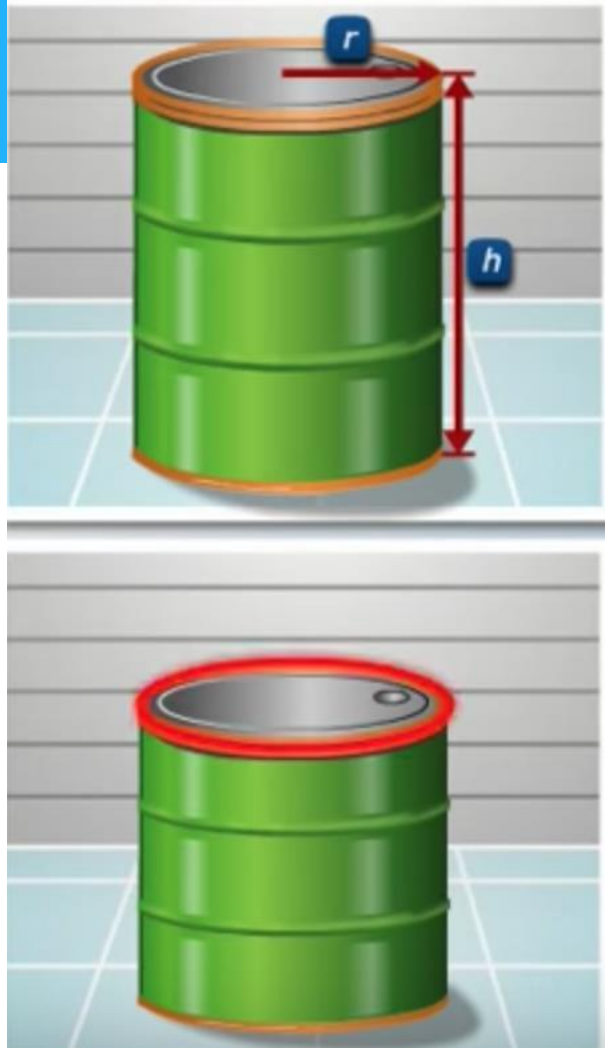
- When we change value of r no Change in h



What is partial derivative?

- When we change value of h no

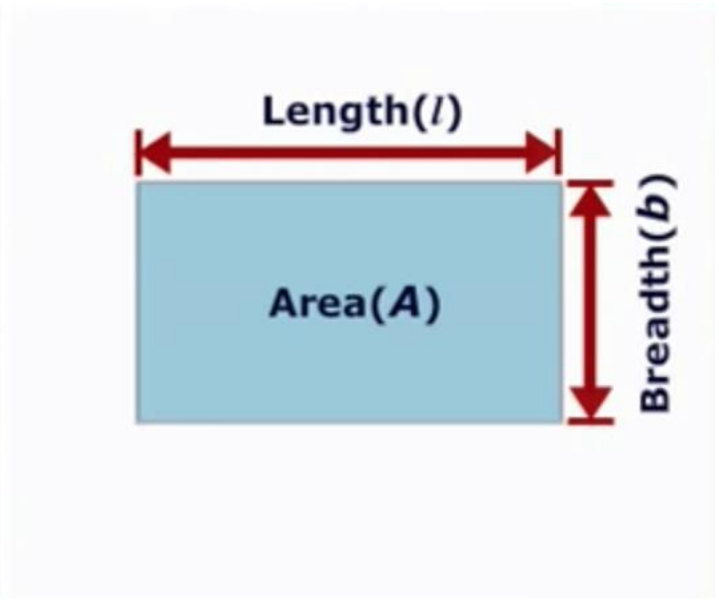
Change in r



What is partial derivative?

- Therefore we see that
- If r changes, is no change in h and vice versa.
- Therefore r and h are independent variables.
- Therefore, V is an example of **Function of several variables**

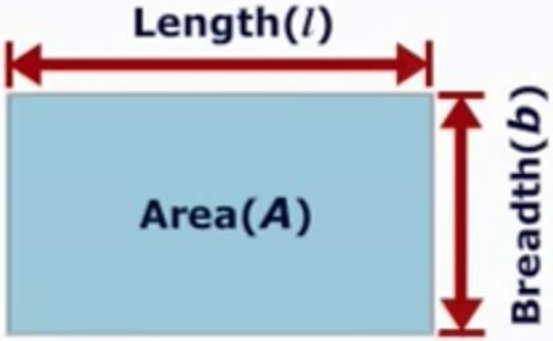
What is partial derivative?



What is partial derivative?

- Area, A = l x b
- Rate of change of Area wrt Length
- Area, A = l x b (we will make b as constant)

$$\text{Partial Derivative} = \frac{\partial A}{\partial l}$$



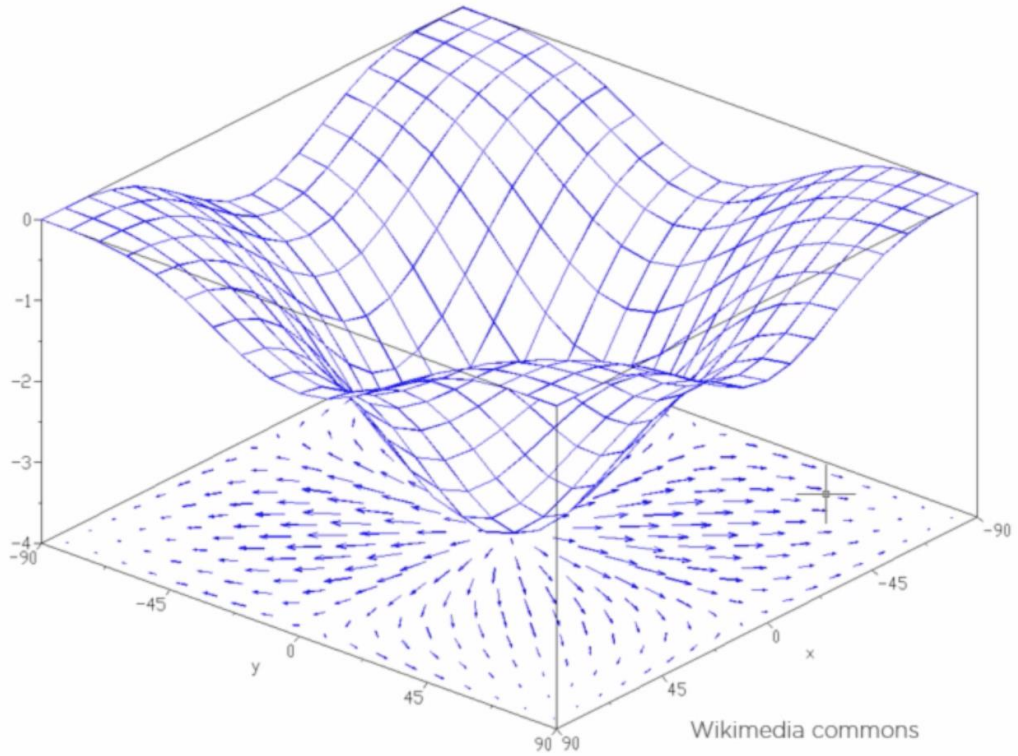
PARTIAL DIFFERENTIATION is a method to *Differentiate a Function* with respect to *One Independent Variable* while Treating the *Other Variables as Constant* It is represented as

$$\text{Partial Derivative} = \frac{\partial(\text{function})}{\partial(\text{Independant Variable})}$$

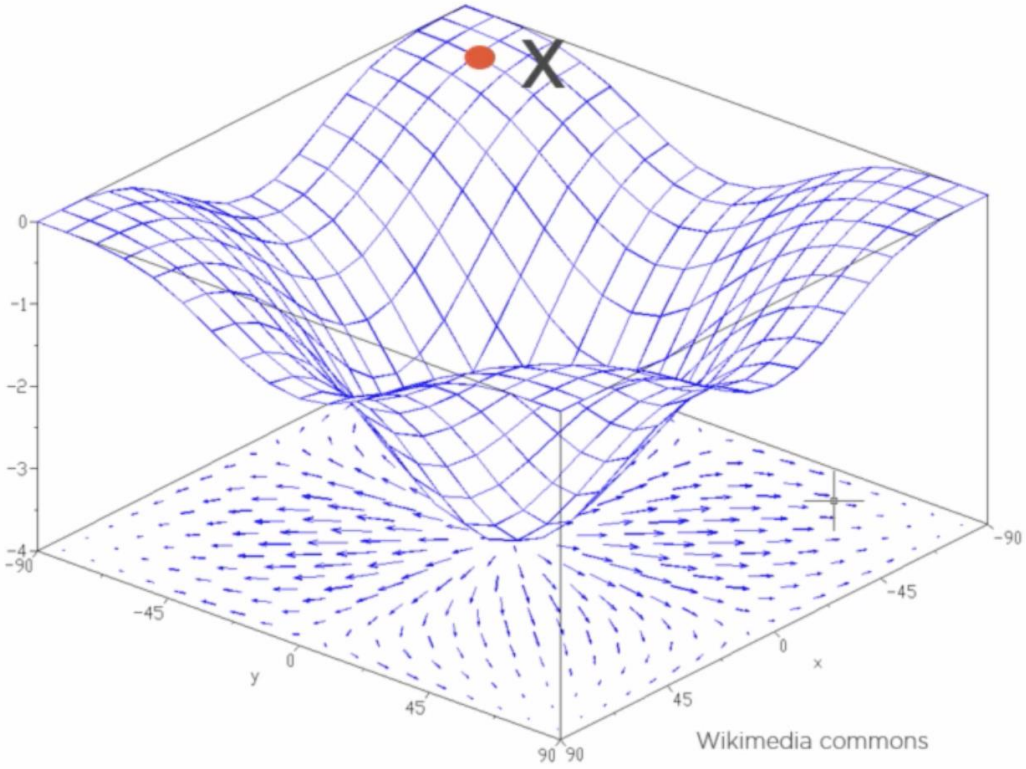
Linear Regression with Gradient Descent

- Step 1: Substitute m and b to zero
- Step 2: Find the prediction and cost function. ie. summation(y-(mx+b))^2
- Step 3: Find b_gradient and m_gradient by substituting in partial derivative function
- Step 4: Update b and m by with respective to gradient value found in step 3 with learning rate
- Step 5: Find the cost function and repeat step 2 to 4

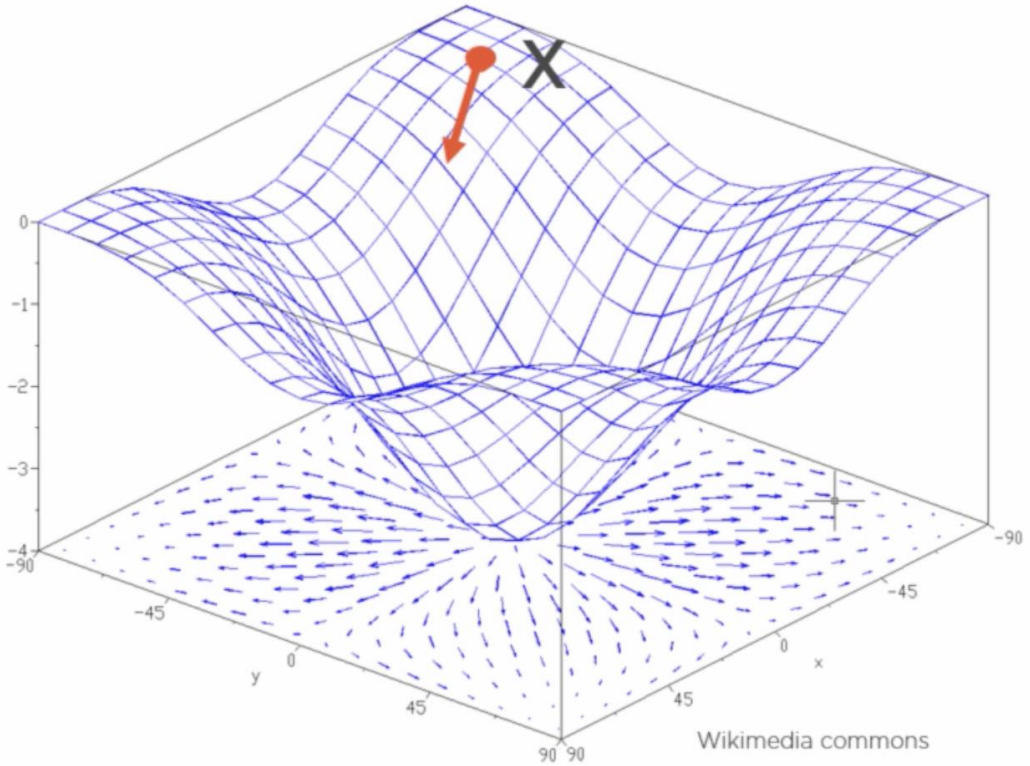
Gradient Descent



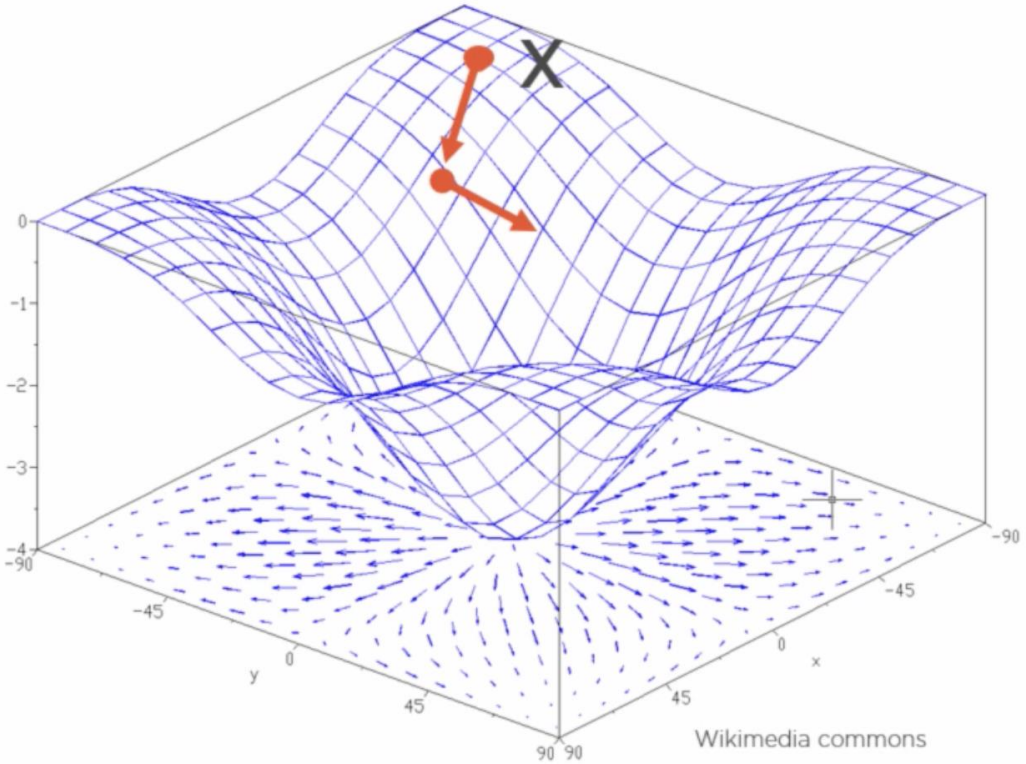
Gradient Descent



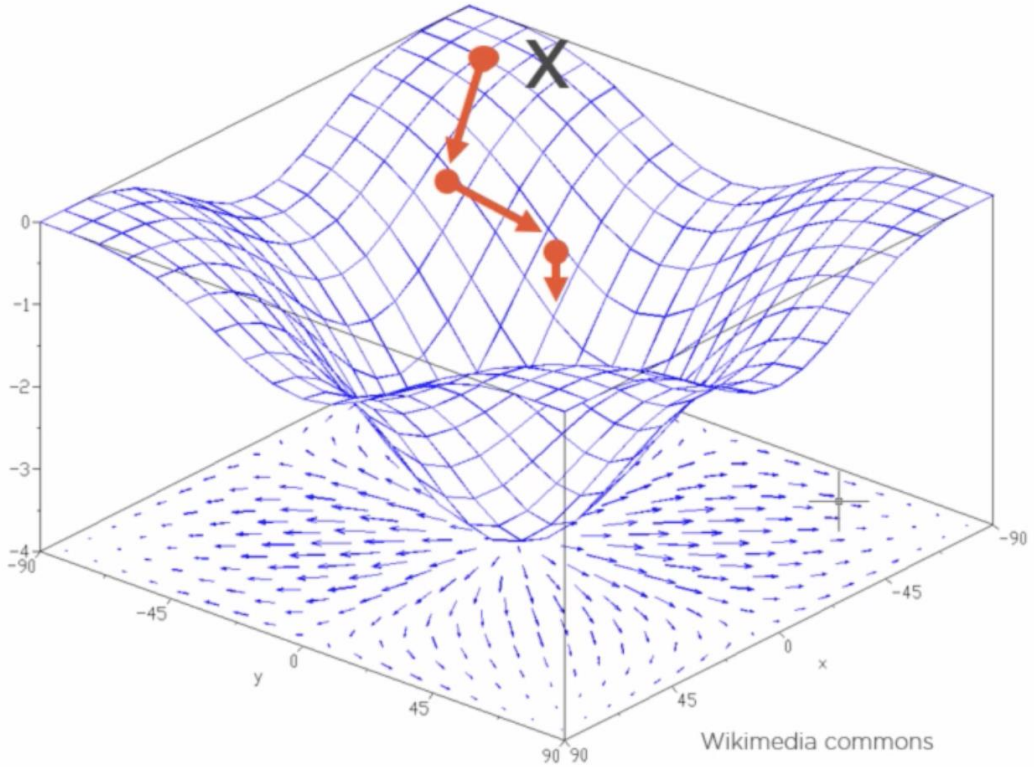
Gradient Descent



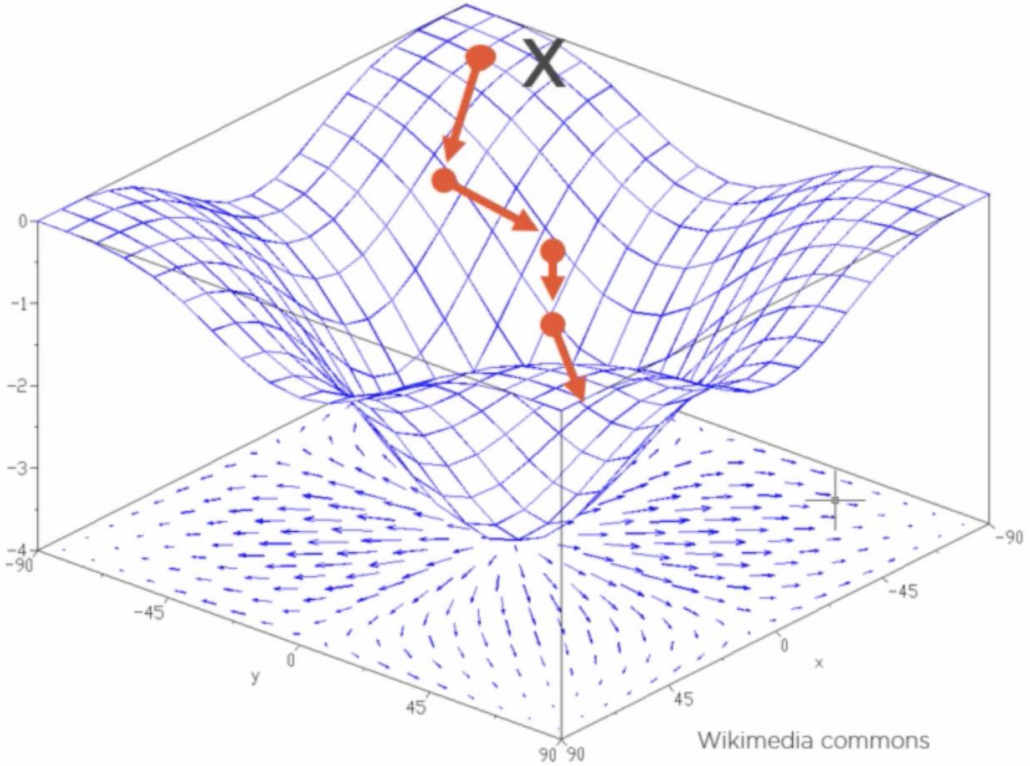
Gradient Descent



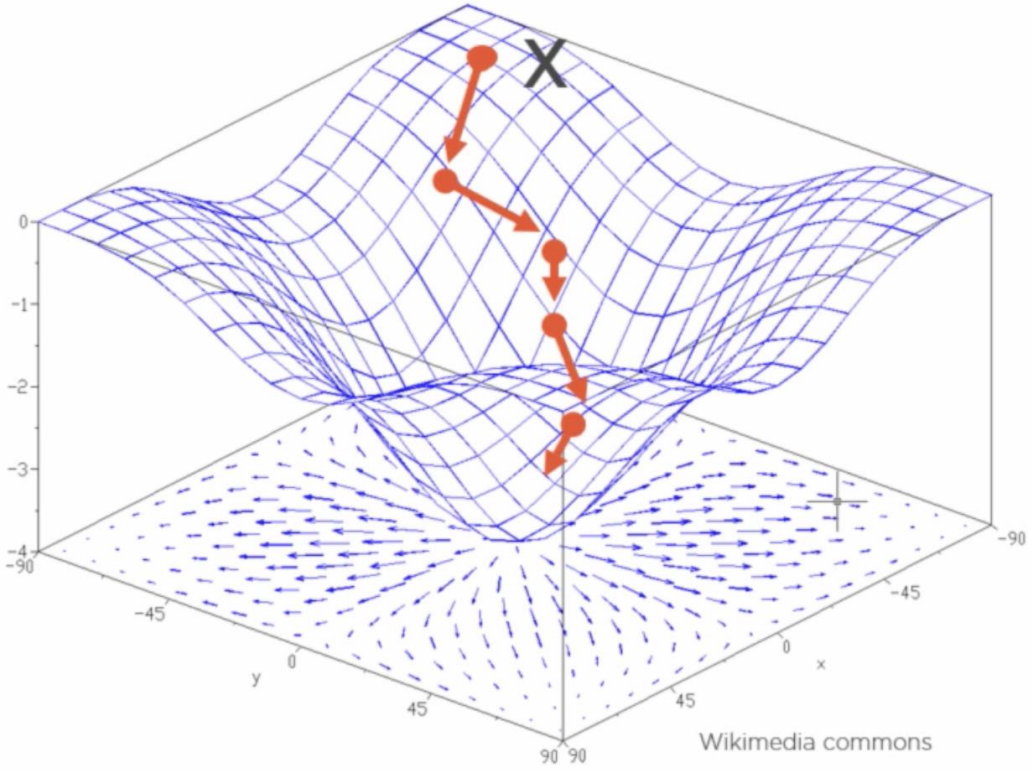
Gradient Descent



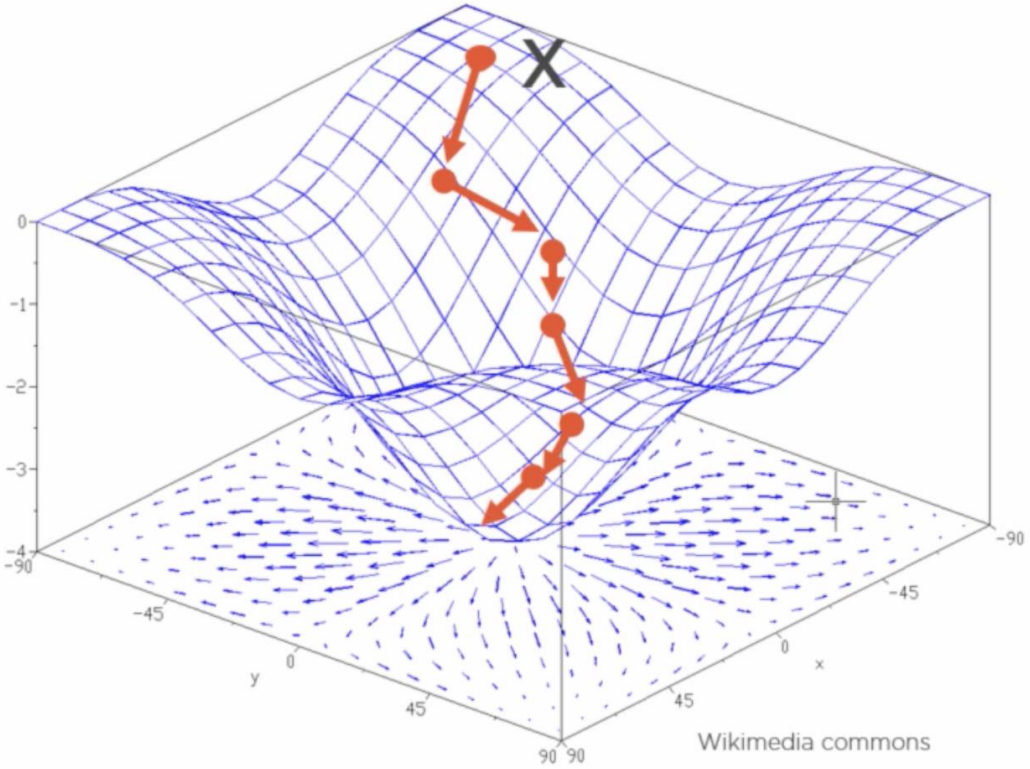
Gradient Descent



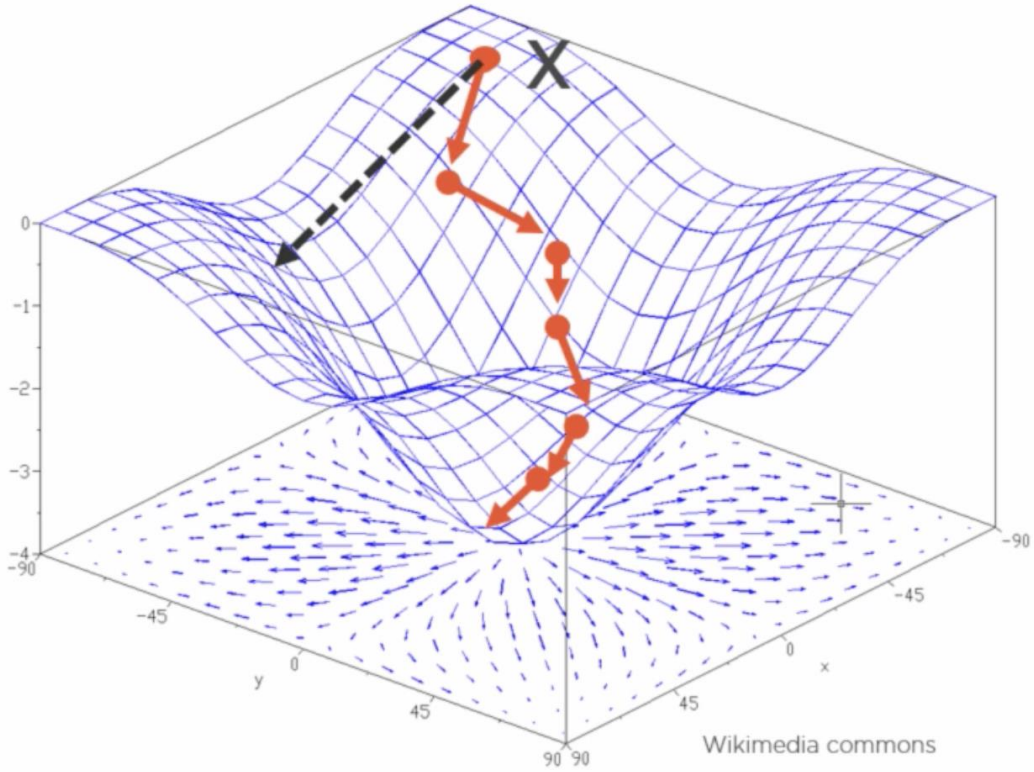
Gradient Descent



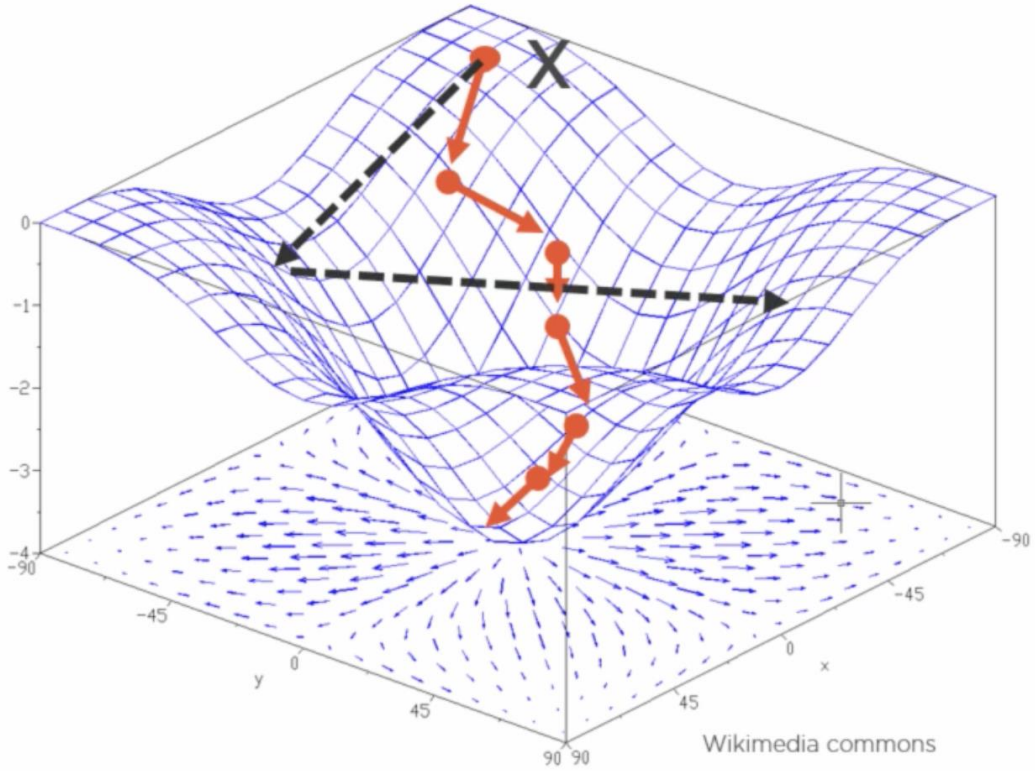
Gradient Descent



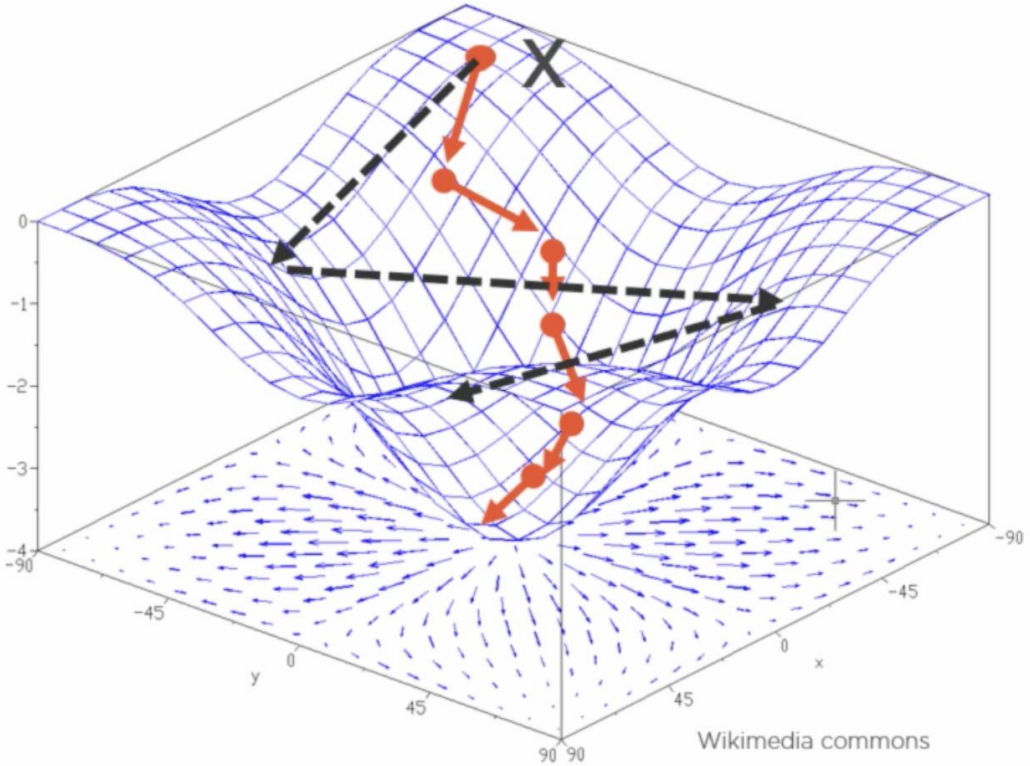
Gradient Descent



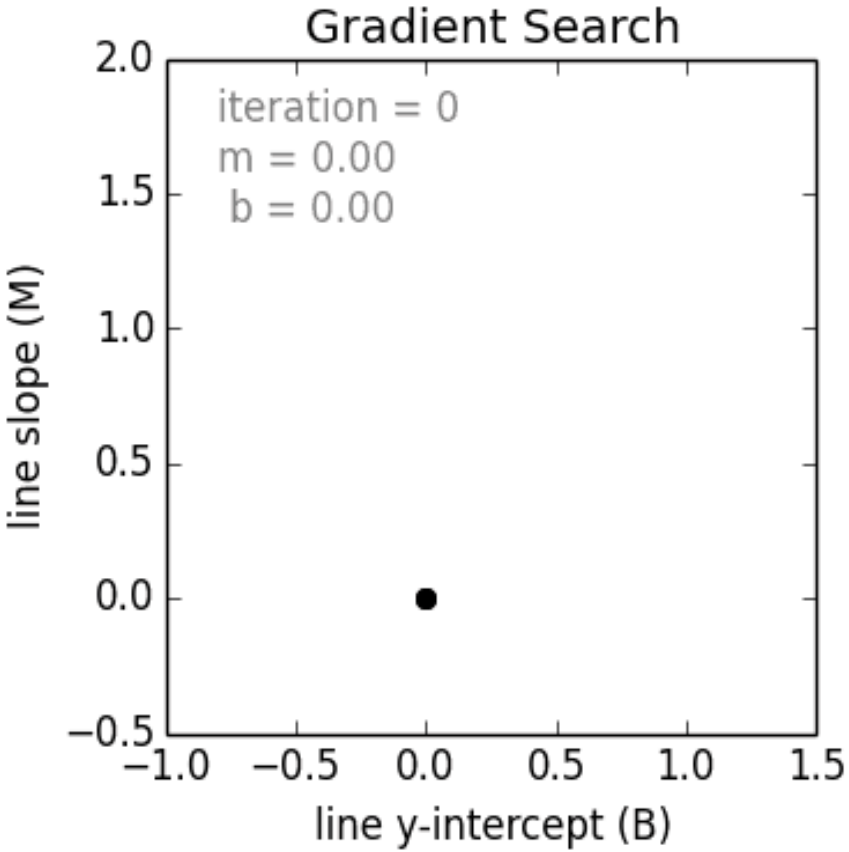
Gradient Descent

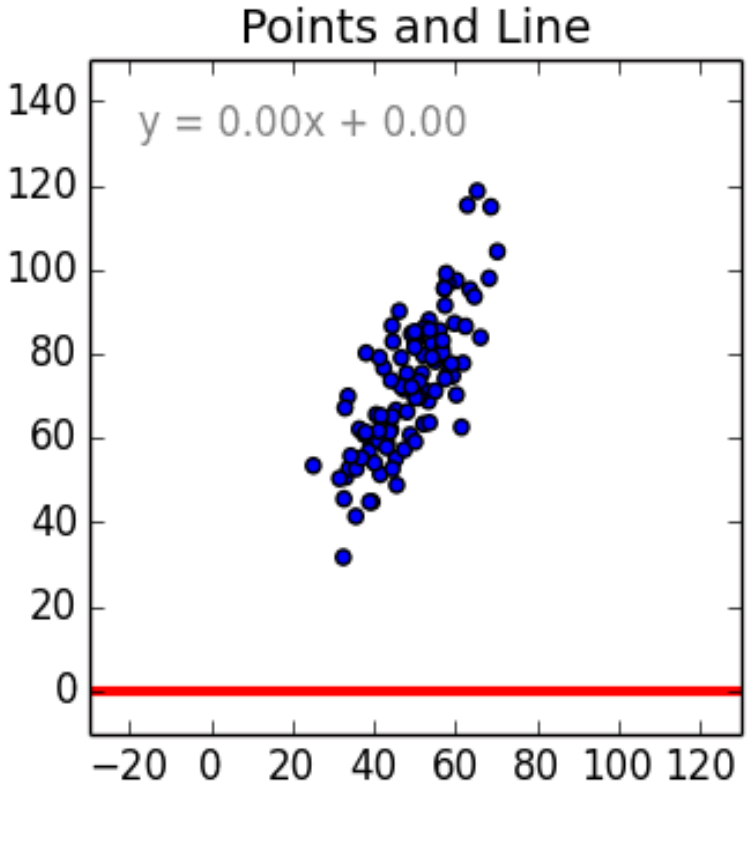


Gradient Descent



Gradient Descent





Coefficient of Determination, R^2

- The coefficient of determination is the portion of the total variation in the dependent variable that is explained by variation in the independent variable

- The coefficient of determination is also called R-squared and is denoted as R^2

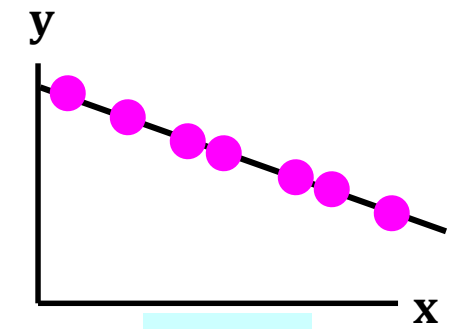
$$R^2 = \frac{SSR}{SST} \quad \text{where} \quad 0 \le R^2 \le 1$$

Coefficient of Determination, R^2

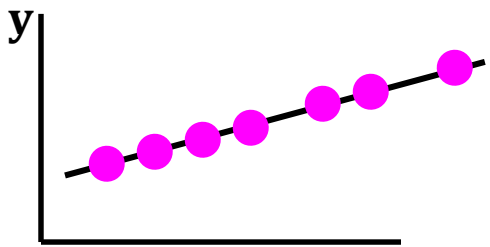
- **Coefficient of determination**

$$R^2 = \frac{SSR}{SST} = \frac{\text{sum of squares explained by regression}}{\text{total sum of squares}}$$

Examples of Approximate R^2 Values



$R^2 = 1$



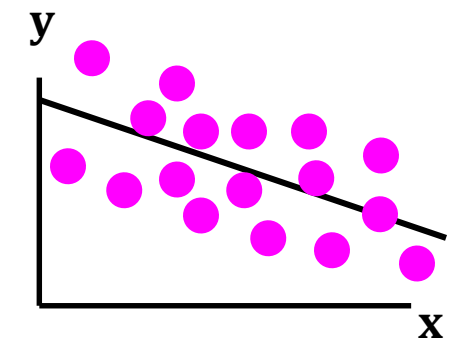
$R^2 = +1$

$R^2 = 1$

Perfect linear relationship between x and y:

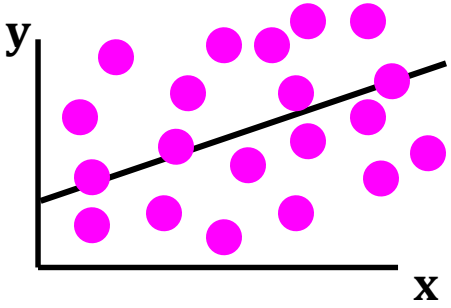
100% of the variation in y is explained by variation in x

Examples of Approximate R^2 Values



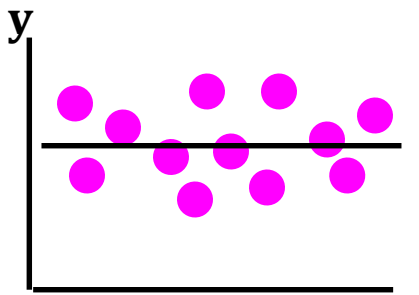
$0 < R^2 < 1$

Weaker linear relationship between x and y:



Some but not all of the variation in y is explained by variation in x

Examples of Approximate R^2 Values



$R^2 = 0$

$R^2 = 0$

No linear relationship between x and y:

The value of Y does not depend on x. (None of the variation in y is explained by variation in x)

Example

House Price in $1000s (y)	Square Feet (x)
245	1400
312	1600
279	1700
308	1875
199	1100
219	1550
405	2350
324	2450
319	1425
255	1700

Estimated Regression Equation:

$$\widehat{\text{house price}} = 98.25 + 0.1098\ (\text{sq.ft.})$$

Predict the price for a house with 2000 square feet

Example: House Prices

Predict the price for a house with
2000 square feet:

$$\widehat{\text{house price}} = 98.25 + 0.1098\ (\text{sq.ft.})$$

$$= 98.25 + 0.1098(200\ 0)$$

$$= 317.85$$

The predicted price for a house with 2000 square feet
is 317.85($1,000s) = $317,850

Linear Regression with Gradient Descent

- **Congratulations!** That’s the first step in your machine learning and artificial intelligence journey.
- Get an intuitive feel for how gradient descent works because this is actually used in more advanced models also.

How hard it is to code?

Step 1

`import` statement:

```
1 from sklearn import linear_model
```

Step 2

I have the height and weight data of some people. Let's use this data to do linear regression and try to predict the weight of other people.

```
1 height=[[4.0],[4.5],[5.0],[5.2],[5.4],[5.8],[6.1],[6.2],[6.4],[6.8]]
2 weight=[  42 ,  44 , 49, 55  , 53  , 58   , 60  , 64  ,  66 ,  69]
3
4 print("height weight")
5 for row in zip(height, weight):
6     print(row[0][0],"->",row[1])
```

Step 3

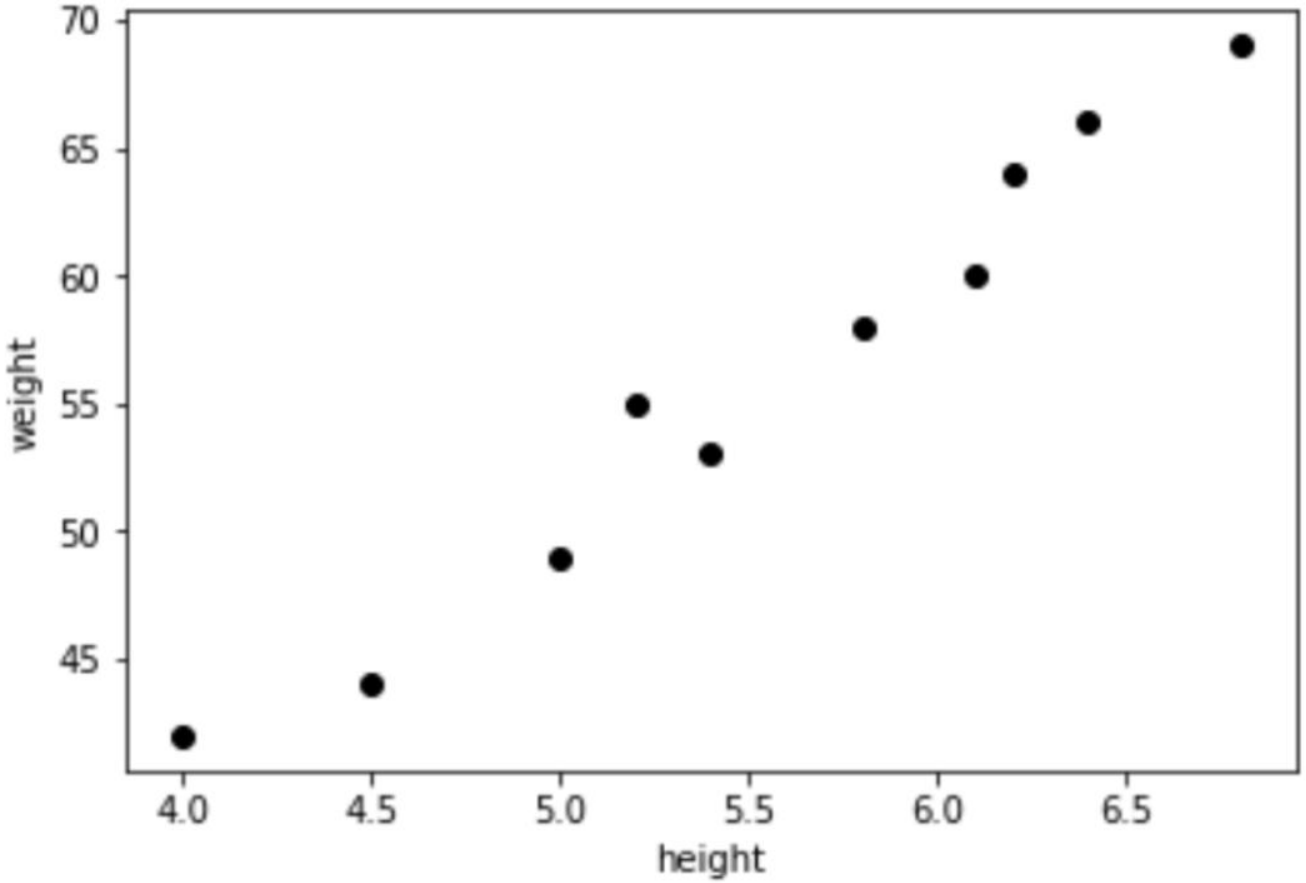
`import` statement to plot graph using `matplotlib` :

```
1 import matplotlib.pyplot as plt
```

Plotting the height and weight data:

```
1 plt.scatter(height,weight,color='black')
2 plt.xlabel("height")
3 plt.ylabel("weight")
```

Output:



Step 4

Declaring the linear regression function and call the `fit` method to learn from data:

```
reg=linear_model.LinearRegression()
reg.fit(height,weight)
```

Slope and intercept:

```
m=reg.coef_[0]
b=reg.intercept_
print("slope=",m, "intercept=",b)
```

Output:

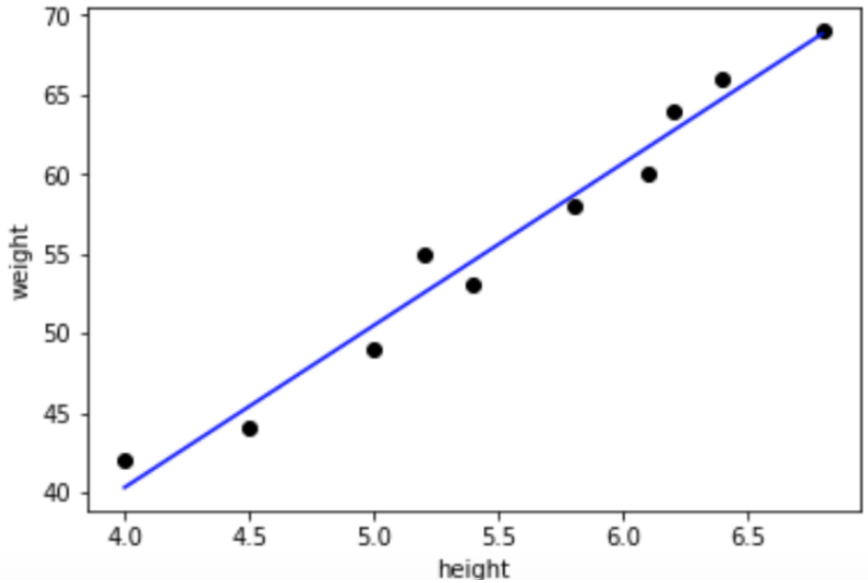
```
slope= 10.1936218679 intercept= -0.4726651480
```

Step 5

Using the values of slope and intercept to construct the line to fit our data points:

```
plt.scatter(height,weight,color='black')
predicted_values = [reg.coef_ * i + reg.intercept_ for i in height]
plt.plot(height, predicted_values, 'b')
plt.xlabel("height")
plt.ylabel("weight")
```

Output:



Error Metrics

Mean squared error	$\mathrm{MSE} = \frac{1}{n}\sum_{t=1}^{n} e_t^2$
Root mean squared error	$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{t=1}^{n} e_t^2}$
Mean absolute error	$\mathrm{MAE} = \frac{1}{n}\sum_{t=1}^{n} \lvert e_t \rvert$
Mean absolute percentage error	$\mathrm{MAPE} = \frac{100\%}{n}\sum_{t=1}^{n} \left\lvert \frac{e_t}{y_t} \right\rvert$

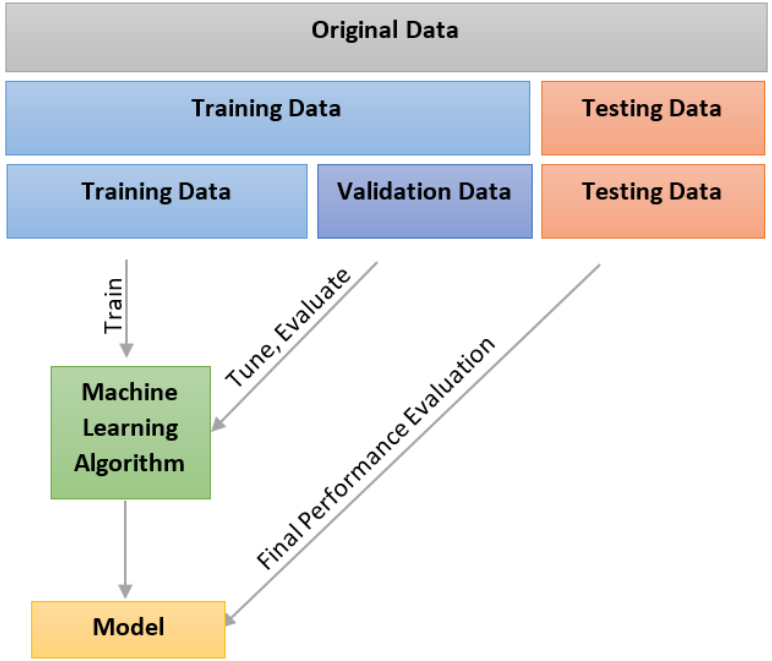
Advantage of Linear Regression

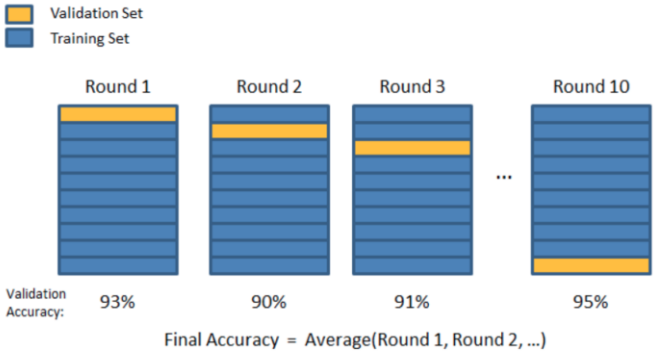
- Linear regression implements a statistical model that, when relationships between the independent variables and the dependent variable are almost linear, shows optimal results.
- Best place to understand the data analysis
- Easily Explicable

Disadvantages

- Linear regression is often inappropriately used to model non-linear relationships.
- Linear regression is limited to predicting numeric output.
- A lack of explanation about what has been learned can be a problem.
- Prone to bias variance problem

How to evaluate our model?





Overfitting vs Underfitting

Training Data(Less Error)

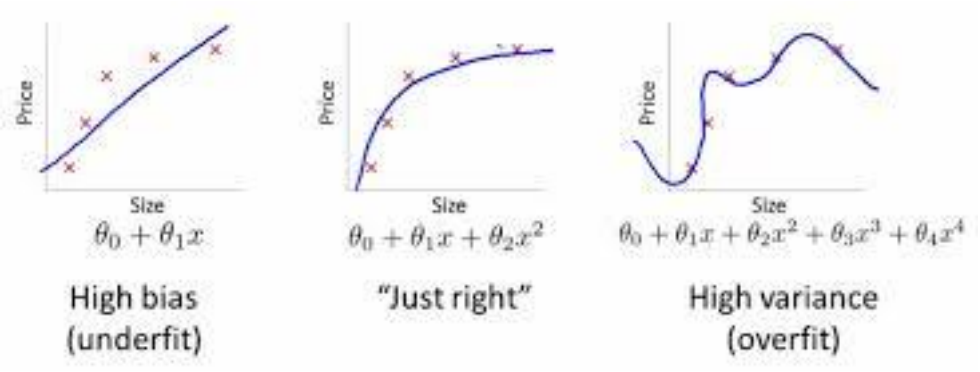
Testing Data (More Error)

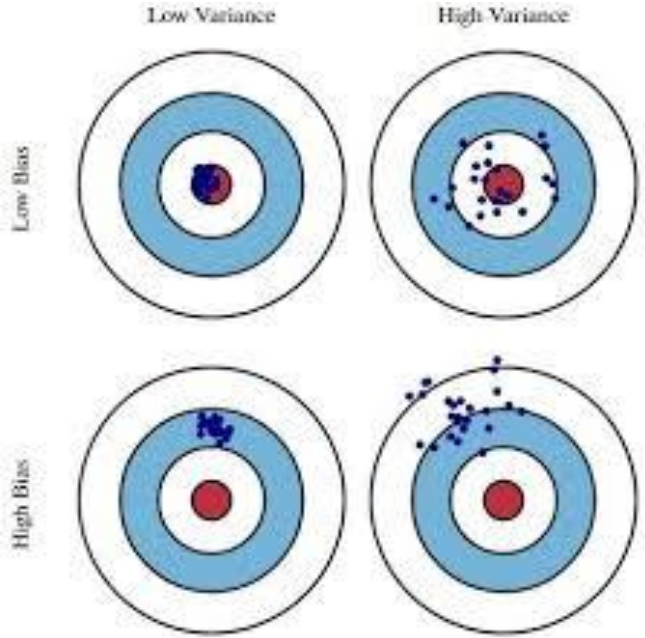
Overfitting vs Underfitting

Training Data (More Error)

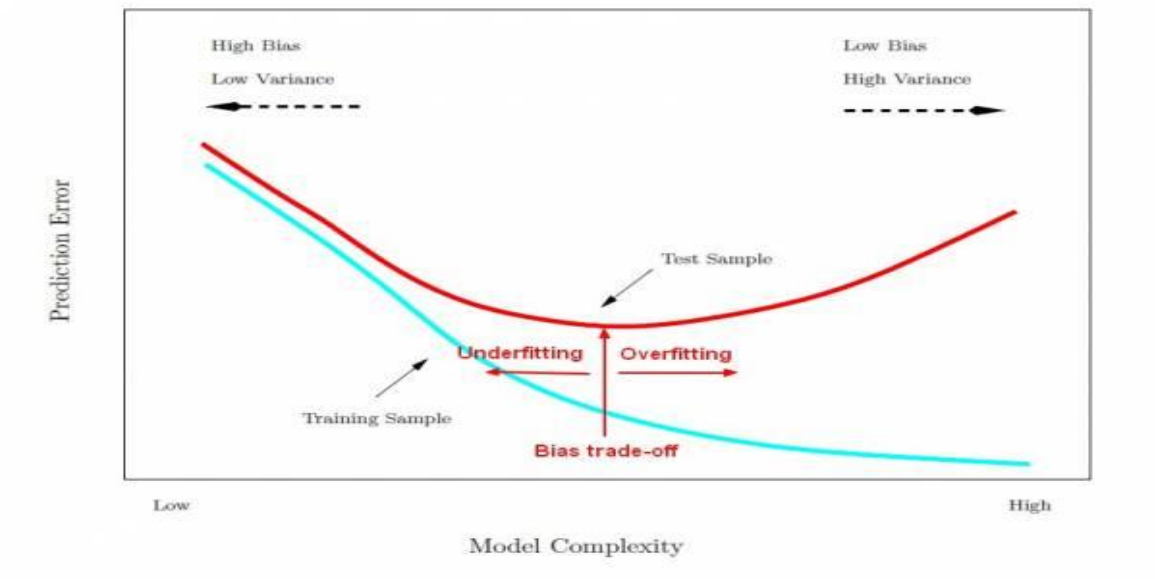
Testing (Still More Error)

Variance and Bias Trade off





Ideal Model should have Low varinance and Low Bias


High Bias
Low Variance
Low Bias
High Variance
Prediction Error
Test Sample
Underfitting
Overfitting
Training Sample
Bias trade-off
Low
High
Model Complexity