

# Mohamed Imran

-Data Scientist  
Ganit Inc.

# 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 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 is believed to be influenced by the value of another variable. It is by definition, dependent on another variable.

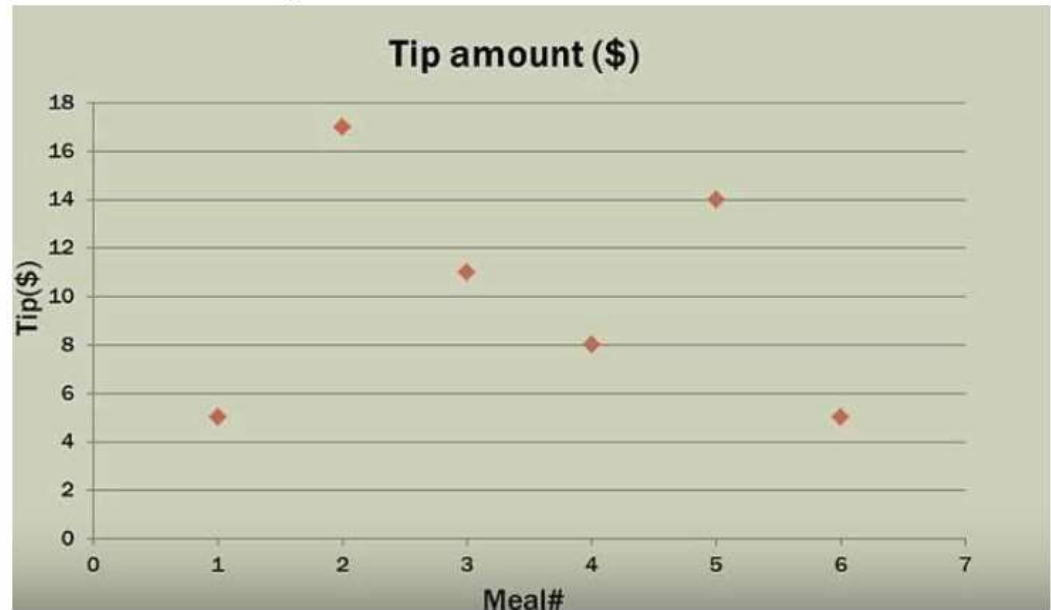
**Best-Fit Line** – Represents our model. It is the line that “best fits” our data points. The line represents the best estimate of the y value for every given input of x.

**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.

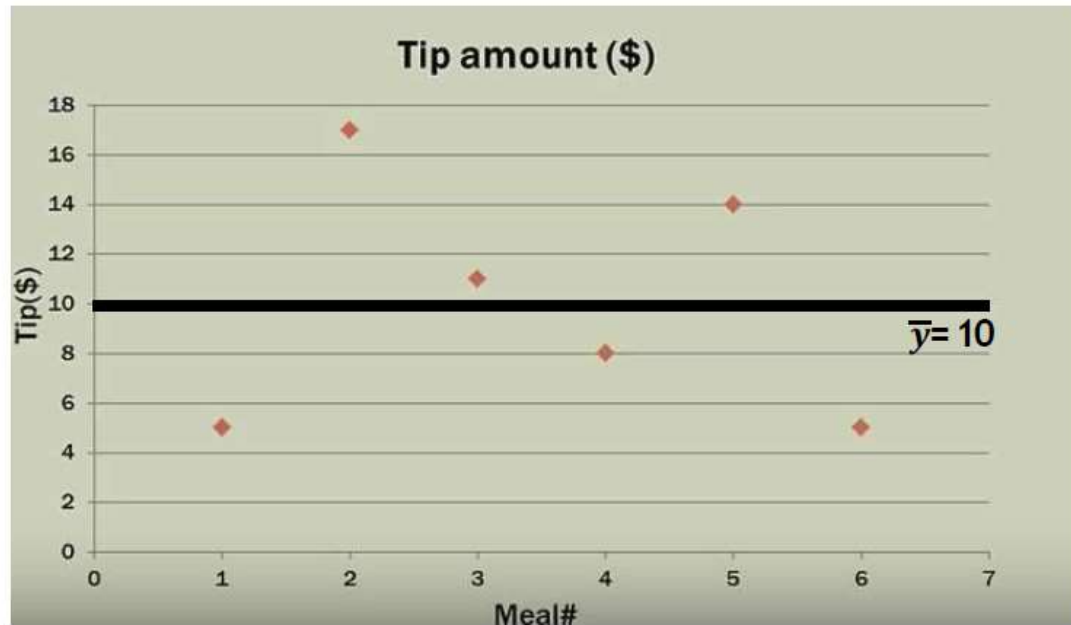
Meal#	Tip amount (\$)
1	5.00
2	17.00
3	11.00
4	8.00
5	14.00
6	5.00



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 axis. We would call the mean  $\bar{y}$ .
- The mean for the tip amounts is 10.
- The model for our problem is simply  $y = 10$ .
- $y = 10$  is our *best fit line* (represented by bold blackline).

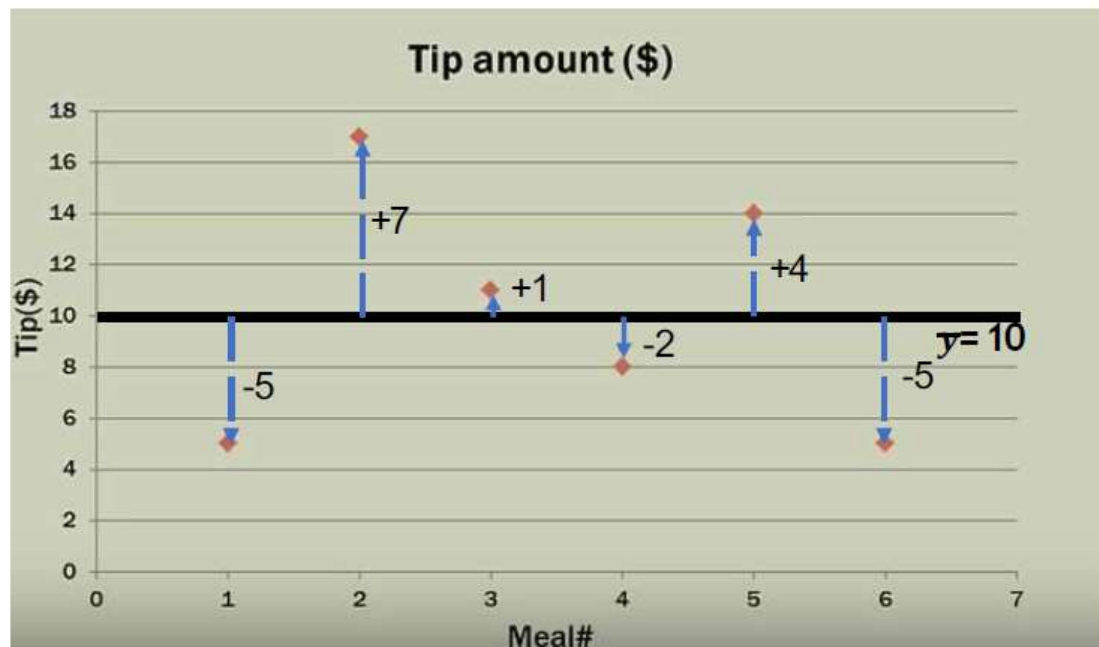
Meal#	Tip amount (\$)
1	5.00
2	17.00
3	11.00
4	8.00
5	14.00
6	5.00





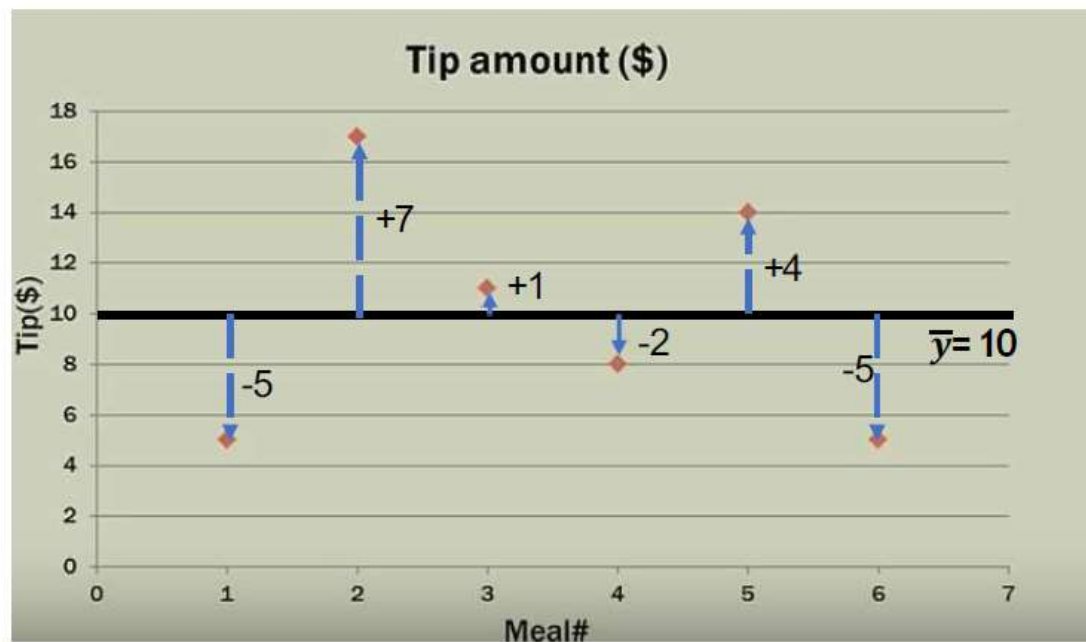
- Now, let's talk about goodness of fit. This will tell us how good our data points fit the line.
- We need to calculate the residuals (errors) for each point.

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  $x$ .

- **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)

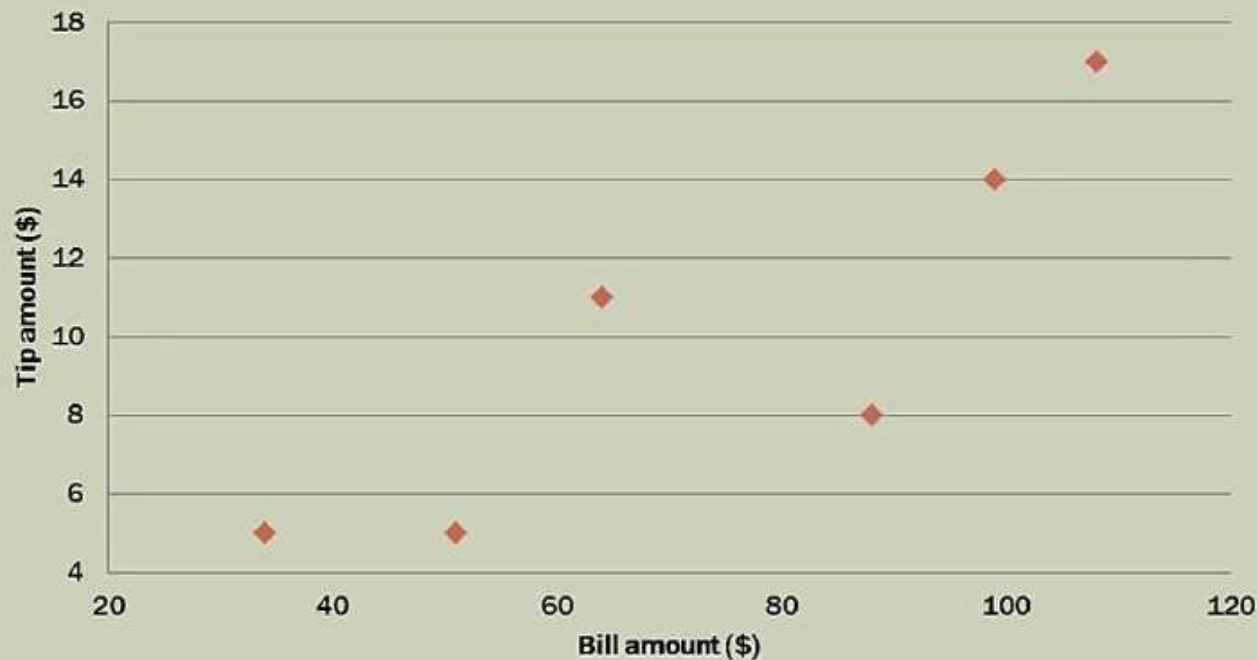
Dependent Variable (y)



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

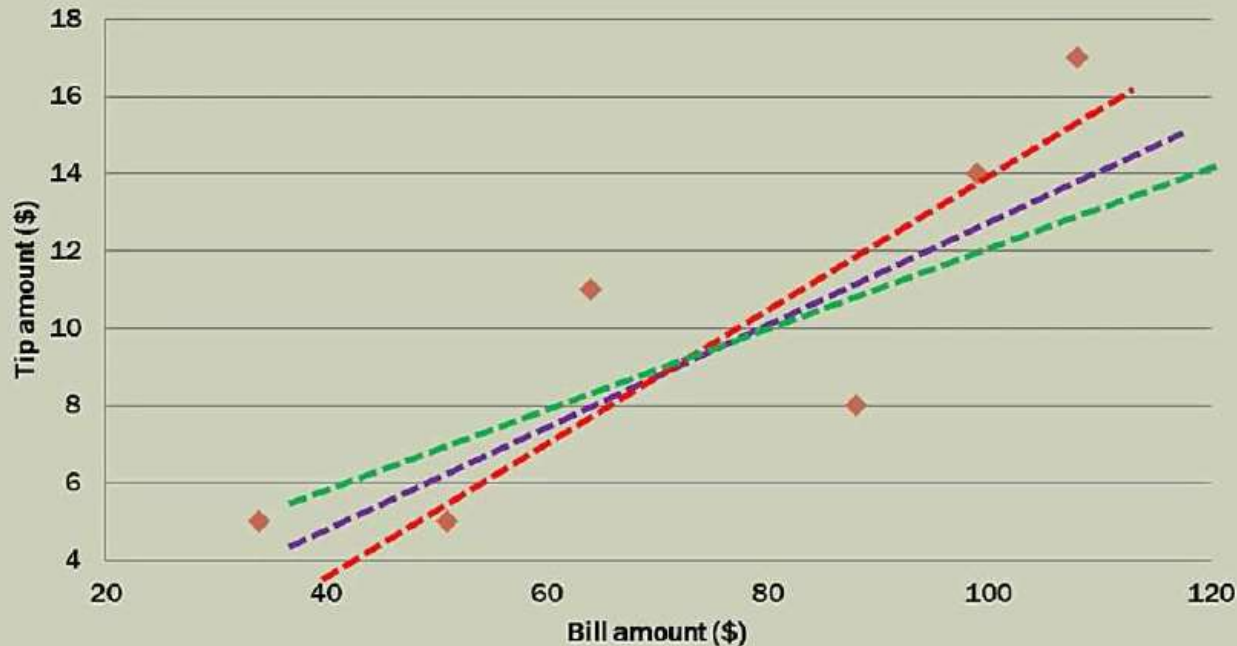
If we scale the graph according to the data points available, we can then plot the points.

**Meal bill vs Tip amount (\$)**



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

## Meal bill vs Tip amount (\$)



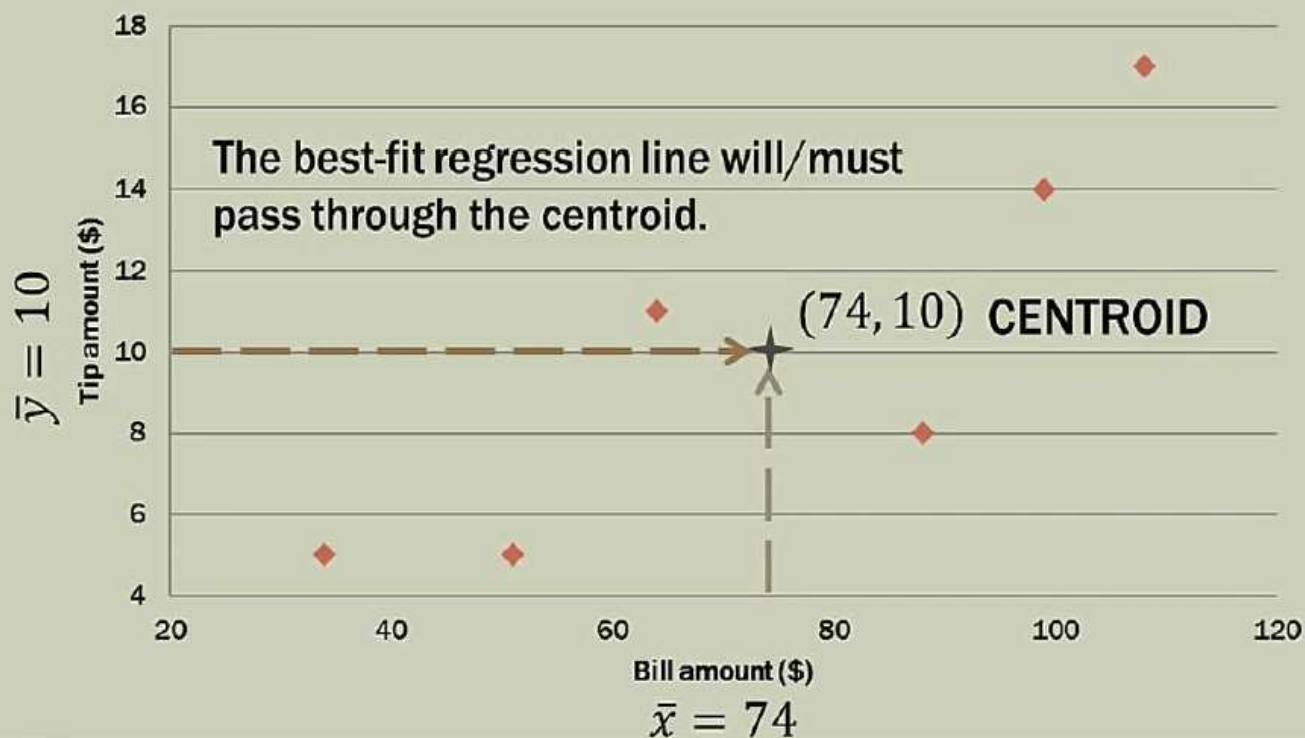
Does the data seem to fall along a line?

*In this case, YES! Proceed.*

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



## Meal bill vs Tip amount (\$)



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
$\bar{x} = 74$	$\bar{y} = 10$



# 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!

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

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

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

Derivation to this

$x_i$  = value of independent 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?

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

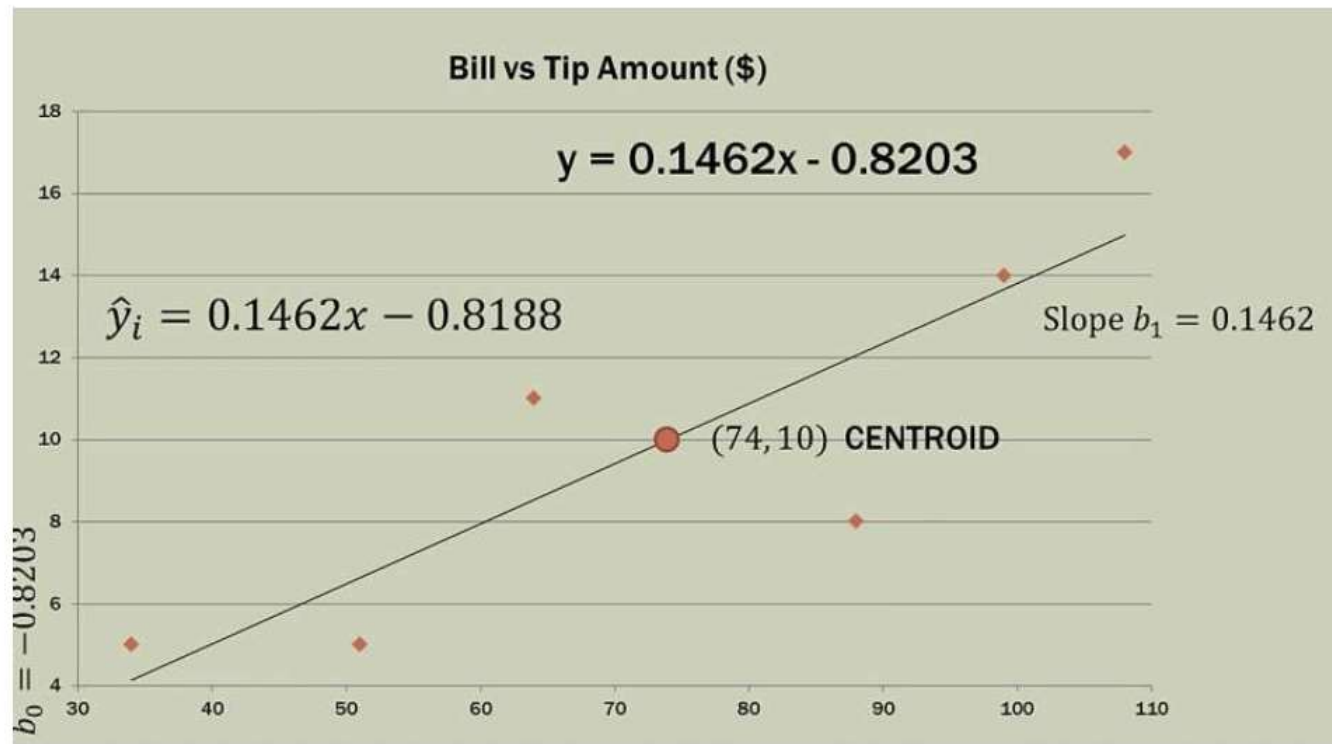
$$b_0 = 10 - 0.1462(74)$$

$$b_0 = 10 - 10.8188$$


$$b_0 = -0.8188$$

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.

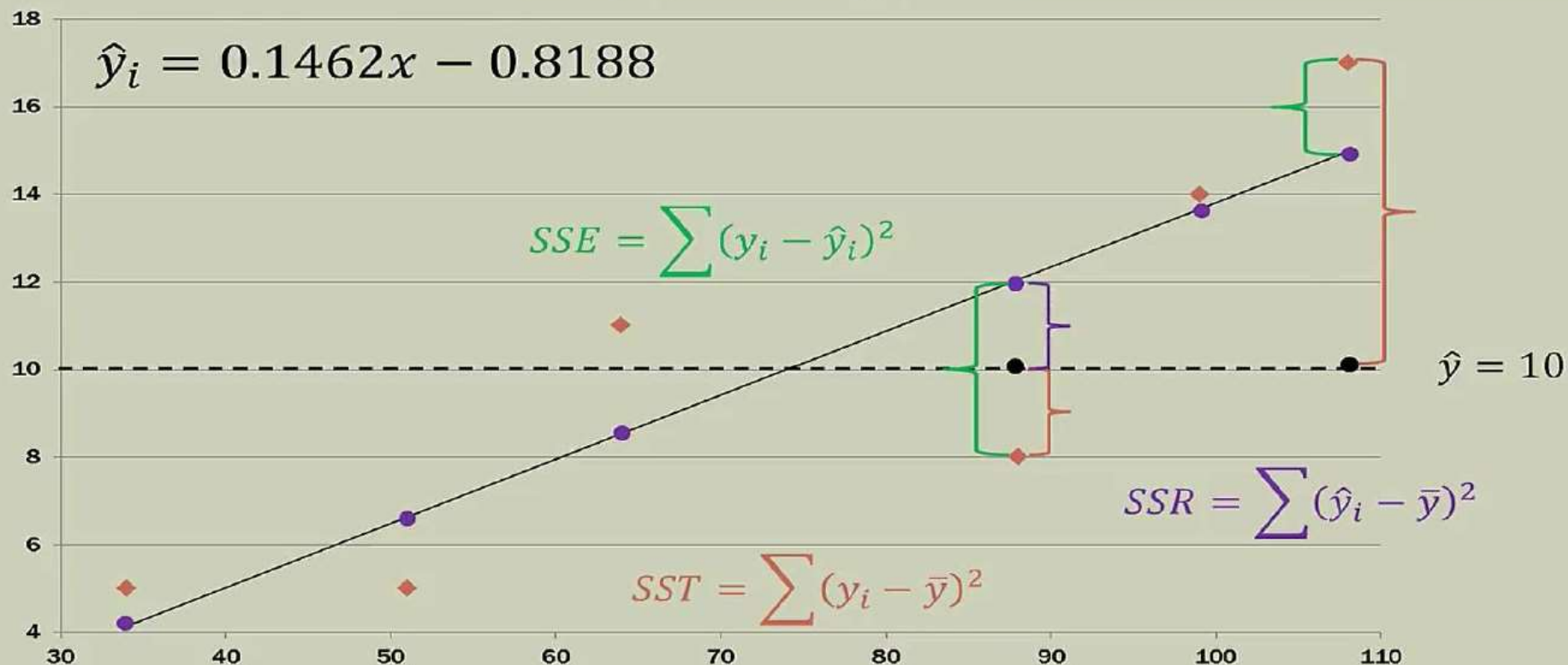
If the bill amount ( $x$ ) is zero, then the expected/predicted tip amount is \$-0.8188 or negative 82-cents! Does this make sense? NO. The intercept may or may not make sense in the “real world.”



# SST = SSR + SSE

Bill vs Tip Amount (\$)

3 Squared Differences



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
	$\bar{x} = 74$	$\bar{y} = 10$		<b>SSE =</b> $\sum = 30.075$	



# Error Metrics

---

Mean squared error

$$\text{MSE} = \frac{1}{n} \sum_{t=1}^n e_t^2$$

---

Root mean squared error

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{t=1}^n e_t^2}$$

---

Mean absolute error

$$\text{MAE} = \frac{1}{n} \sum_{t=1}^n |e_t|$$

---

Mean absolute percentage error

$$\text{MAPE} = \frac{100\%}{n} \sum_{t=1}^n \left| \frac{e_t}{y_t} \right|$$

---

## 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

# Logistic Regression

What type of target data was in Linear Regression?

Continuous

What if it is discrete?

We have to classify them

# Classification problems

- Email - spam/not spam?
- Online transactions - fraudulent?
- Tumour - Malignant/benign
- Gaming - Win vs Loss
- Sales - Buying vs Not buying
- Marketing – Response vs No Response
- Credit card & Loans – Default vs Non Default
- Operations – Attrition vs Retention
- Websites – Click vs No click
- Fraud identification –Fraud vs Non Frau
- Healthcare –Cure vs No Cure

# Learn from what we know.

- We would like to use something like what we know from linear regression:
- Continuous outcome =  $\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots$

# Transforming a proportion

- A proportion is a value between 0 and 1
- The odds are always positive:

$$\text{odds} = \left( \frac{p}{1-p} \right) \Rightarrow [0, +\infty)$$

- The log odds is continuous:

$$\text{Logodds} = \ln \left( \frac{p}{1-p} \right) \Rightarrow (-\infty, +\infty)$$

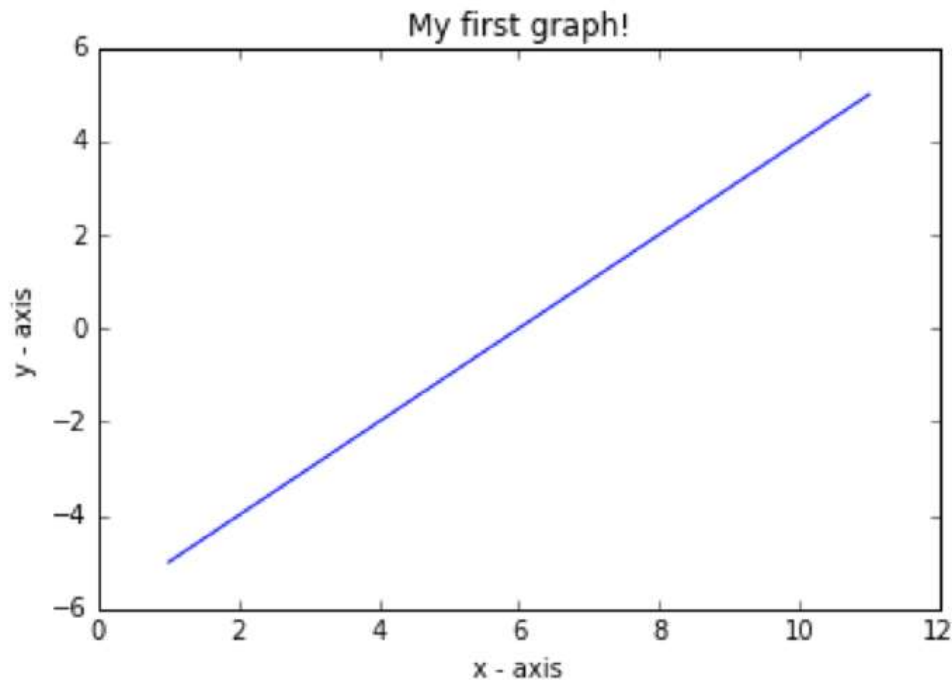


# “Logit” transformation of the probability

Measure	Min	Max	Name
$\Pr(Y = 1)$	0	1	“probability”
$\frac{\Pr(Y = 1)}{1 - \Pr(Y = 1)}$	0	$\infty$	“odds”
$\log\left(\frac{\Pr(Y = 1)}{1 - \Pr(Y = 1)}\right)$	$-\infty$	$\infty$	“log-odds” or “logit”

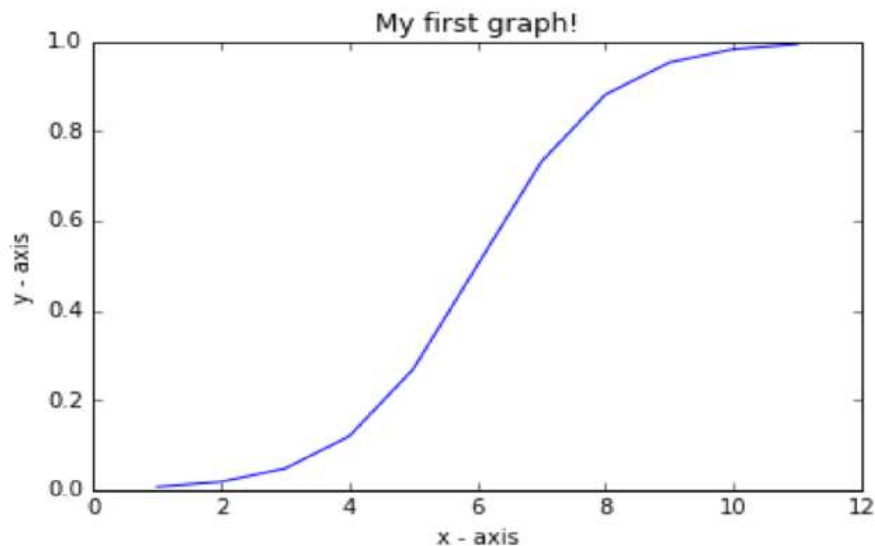
# Regression line

x	Y
1	-5
2	-4
3	-3
4	-2
5	-1
6	0
7	1
8	2
9	3
10	4
11	5

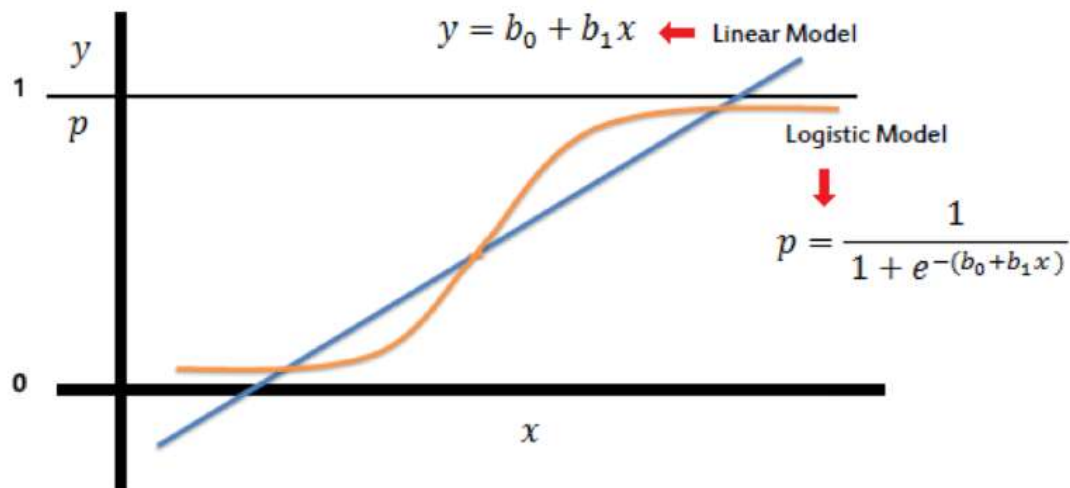


# Transformation to Classification

x	Sigmoid(Y)
1	0.006692850924
2	0.01798620996
3	0.04742587318
4	0.119202922
5	0.2689414214
6	0.5
7	0.7310585786
8	0.880797078
9	0.9525741268
10	0.98201379
11	0.9933071491



# Logistic Regression Equation



# Learning from Example

- In 1846 the Donner and Reed families left Springfield, Illinois, for California by covered wagon. In July, the Donner Party, as it became known, reached Fort Bridger, Wyoming. There its leaders decided to attempt a new and untested route to the Sacramento Valley. Having reached its full size of 87 people and 20 wagons, the party was delayed by a difficult crossing of the Wasatch Range and again in the crossing of the desert west of the Great Salt Lake. The group became stranded in the eastern Sierra Nevada mountains when the region was hit by heavy snows in late October. By the time the last survivor was rescued on April 21, 1847, 40 of the 87 members had died from famine and exposure to extreme cold.

# Dataset

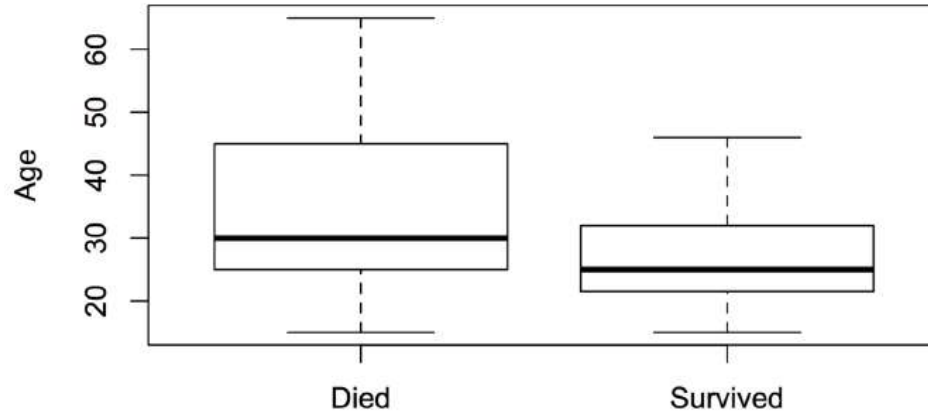
	Age	Sex	Status
1	23.00	Male	Died
2	40.00	Female	Survived
3	40.00	Male	Survived
4	30.00	Male	Died
5	28.00	Male	Died
⋮	⋮	⋮	⋮
43	23.00	Male	Survived
44	24.00	Male	Died
45	25.00	Female	Survived

# Exploratory Analysis

Status vs. Gender:

	Male	Female
Died	20	5
Survived	10	10

Status vs. Age:



# Exploratory Analysis

- It seems clear that both age and gender have an effect on someone's survival,
- how do we come up with a model that will let us explore this relationship?
- Even if we set Died to 0 and Survived to 1, this isn't something we can transform our way out of - we need something more.
- One way to think about the problem - we can treat Survived and Died as successes and failures arising from a binomial distribution where the probability of a success is given by a transformation of a linear model of the predictors.



- It turns out that this is a very general way of addressing this type of problem in regression, and the resulting models are called Logistic regression.

- All Logistic regression have the following three characteristics:

- A probability distribution describing the outcome variable

- A linear model

Linear regression

$$Y = b_0 + b_1 \times X_1 + b_2 \times X_2 + \dots + b_K \times X_K$$

- A link function that relates the linear model to the parameter of the outcome distribution

Linear regression

$$Y = b_0 + b_1 \times X_1 + b_2 \times X_2 + \dots + b_K \times X_K$$

Sigmoid Function

$$P = \frac{1}{1 + e^{-Y}}$$

$$\ln \left( \frac{P}{1 - P} \right) = b_0 + b_1 \times X_1 + b_2 \times X_2 + \dots + b_K \times X_K$$

## Odds / Probability of survival for a new-born (Age=0):

Model:

$$\log \left( \frac{p}{1-p} \right) = 1.8185 - 0.0665 \times \text{Age}$$

$$\log \left( \frac{p}{1-p} \right) = 1.8185 - 0.0665 \times 0$$

---

$$\frac{p}{1-p} = \exp(1.8185) = 6.16$$

$$p = 6.16 / 7.16 = 0.86$$

# Metrics for Performance Evaluation

- Focus on the predictive capability of a model
  - Rather than how fast it takes to classify or build models, scalability, etc.
- Confusion Matrix:

ACTUAL CLASS	PREDICTED CLASS		
		Class=Yes	Class=No
	Class=Yes	a: TP	b: FN
	Class=No	c: FP	d: TN

a: TP (true positive)

b: FN (false negative)

c: FP (false positive)

d: TN (true negative)

n=165	Predicted: NO	Predicted: YES
Actual: NO	50	10
Actual: YES	5	100

What can we learn from this matrix?

- There are two possible predicted classes: "yes" and "no". If we were predicting the presence of a disease, for example, "yes" would mean they have the disease, and "no" would mean they don't have the disease.
- The classifier made a total of 165 predictions (e.g., 165 patients were being tested for the presence of that disease).
- Out of those 165 cases, the classifier predicted "yes" 110 times, and "no" 55 times.
- In reality, 105 patients in the sample have the disease, and 60 patients do not.

Let's now define the most basic terms, which are whole numbers (not rates):

- **true positives (TP):** These are cases in which we predicted yes (they have the disease), and they do have the disease
- **true negatives (TN):** We predicted no, and they don't have the disease.
- **false positives (FP):** We predicted yes, but they don't actually have the disease. (Also known as a "Type I error.")
- **false negatives (FN):** We predicted no, but they actually do have the disease. (Also known as a "Type II error.")

I've added these terms to the confusion matrix, and also added the row and column totals:

n=165	Predicted: NO	Predicted: YES	
Actual: NO	TN = 50	FP = 10	60
Actual: YES	FN = 5	TP = 100	105
	55	110	

This is a list of rates that are often computed from a confusion matrix for a binary classifier:

- **Accuracy:** Overall, how often is the classifier correct?
  - $(TP+TN)/total = (100+50)/165 = 0.91$
- **Misclassification Rate:** Overall, how often is it wrong?
  - $(FP+FN)/total = (10+5)/165 = 0.09$
  - equivalent to 1 minus Accuracy
  - also known as "Error Rate"
- **True Positive Rate:** When it's actually yes, how often does it predict yes?
  - $TP/actual\ yes = 100/105 = 0.95$
  - also known as "Sensitivity" or "Recall"
- **False Positive Rate:** When it's actually no, how often does it predict yes?
  - $FP/actual\ no = 10/60 = 0.17$
- **Specificity:** When it's actually no, how often does it predict no?
  - $TN/actual\ no = 50/60 = 0.83$
  - equivalent to 1 minus False Positive Rate
- **Precision:** When it predicts yes, how often is it correct?
  - $TP/predicted\ yes = 100/110 = 0.91$
- **Prevalence:** How often does the yes condition actually occur in our sample?
  - $actual\ yes/total = 105/165 = 0.64$



**Type I error**  
(false positive)



**Type II error**  
(false negative)



Figure 3.1 Type I and Type II errors

## Error Metrics

Metric	Formula
True positive rate, recall	$\frac{TP}{TP+FN}$
False positive rate	$\frac{FP}{FP+TN}$
Precision	$\frac{TP}{TP+FP}$
Accuracy	$\frac{TP+TN}{TP+TN+FP+FN}$
F-measure	$\frac{2 \cdot \text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$

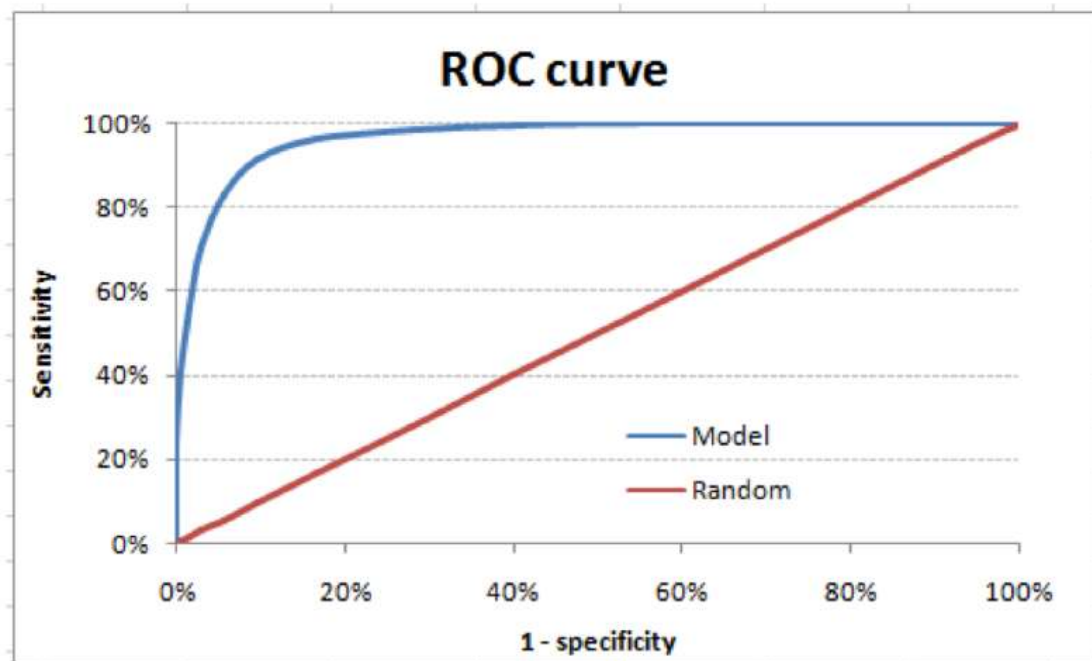


# Error Metrics

Metrics	Formula	Evaluation Focus
Accuracy (acc)	$\frac{tp + tn}{tp + fp + tn + fn}$	In general, the accuracy metric measures the ratio of correct predictions over the total number of instances evaluated.
Error Rate (err)	$\frac{fp + fn}{tp + fp + tn + fn}$	Misclassification error measures the ratio of incorrect predictions over the total number of instances evaluated.
Sensitivity (sn)	$\frac{tp}{tp + fn}$	This metric is used to measure the fraction of positive patterns that are correctly classified
Specificity (sp)	$\frac{tn}{tn + fp}$	This metric is used to measure the fraction of negative patterns that are correctly classified.
Precision (p)	$\frac{tp}{tp + fp}$	Precision is used to measure the positive patterns that are correctly predicted from the total predicted patterns in a positive class.
Recall (r)	$\frac{tp}{tp + tn}$	Recall is used to measure the fraction of positive patterns that are correctly classified
F-Measure (FM)	$\frac{2 * p * r}{p + r}$	This metric represents the harmonic mean between recall and precision values
Geometric-mean (GM)	$\sqrt{tp * tn}$	This metric is used to maximize the $tp$ rate and $tn$ rate, and simultaneously keeping both rates relatively balanced

# The ROC curve

- In a Receiver Operating Characteristic (ROC) curve the true positive rate (Sensitivity) is plotted in function of the false positive rate (100-Specificity) for different cut-off points.



- 90-1 = excellent (A)
- .80-.90 = good (B)
- .70-.80 = fair (C)
- .60-.70 = poor (D)
- .50-.60 = fail (F)

## Logistic Regression Merits

- Simple and efficient.
- Low variance.
- It provides **probability** score for observations.

## Logistic Regression Demerits

- Doesn't handle **large** number of categorical features/variables well.
- It requires transformation of non-linear features.