What is Linear Regression?

- Regression allows to model mathematically the relationship between two or more variables (specifically linear relationship with the help of algebra)

- **Dependent Variable (DV)**
- **Independent Variable (IV)**

Question 1:

Suppose you are an owner of a restaurant and interested to develop a model that will allow you to make a prediction about what amount of tip to expect for any given bill amount?

- collected data for six meals

Data for Meals

Meal (#)	Tip Amount (in Rs.)
1.	7
2.	19
3.	13
4.	10
5.	16
6.	7
7.	**?**

Interesting Fact: You have only tip amount data?

- How to predict tip amount?
- What is DV and IV variable?

Linear Regression

Finding the value of dependent variable
Independent variables
Mathematical model: $Y = f(X)$

Linear Regression

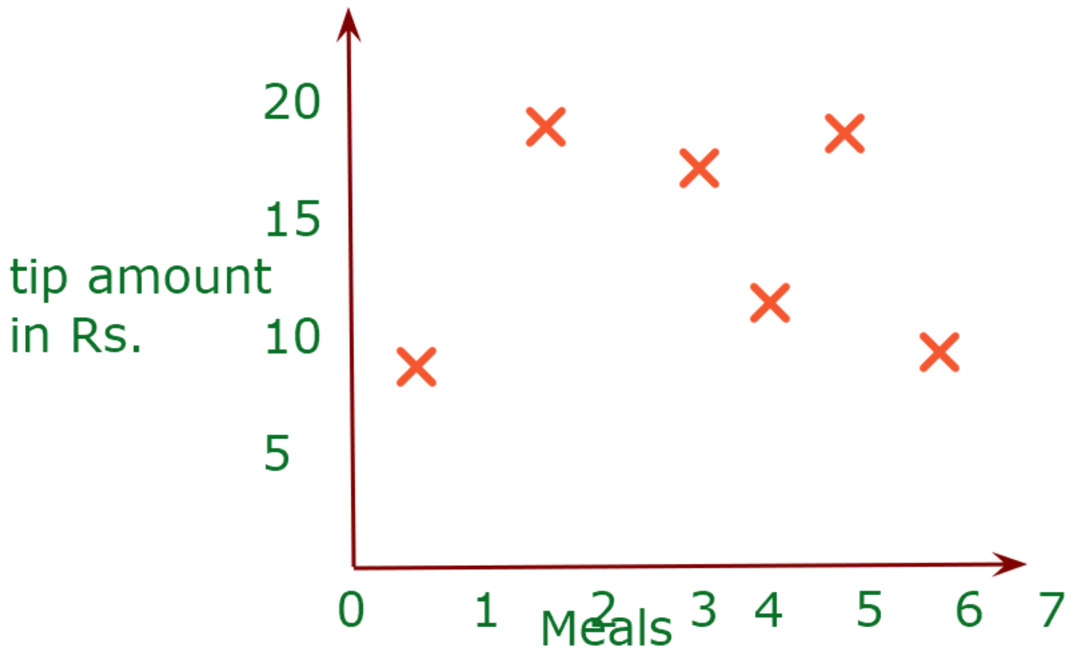
Finding the value of dependent variable
Independent variables
Mathematical model: $Y = f(X)$

Q1. X:=?
7, 19, 13, 10, 17, & 7: tip amount
tip amount for meal X=?
Can we predict the tip amount:?

Linear Regression

Finding the value of dependent variable
Independent variables
Mathematical model: Y = f(X)

Q1. X:=?
7, 19, 13, 10, 17, & 7: tip amount
tip amount for meal X=?
Can we predict the tip amount:?

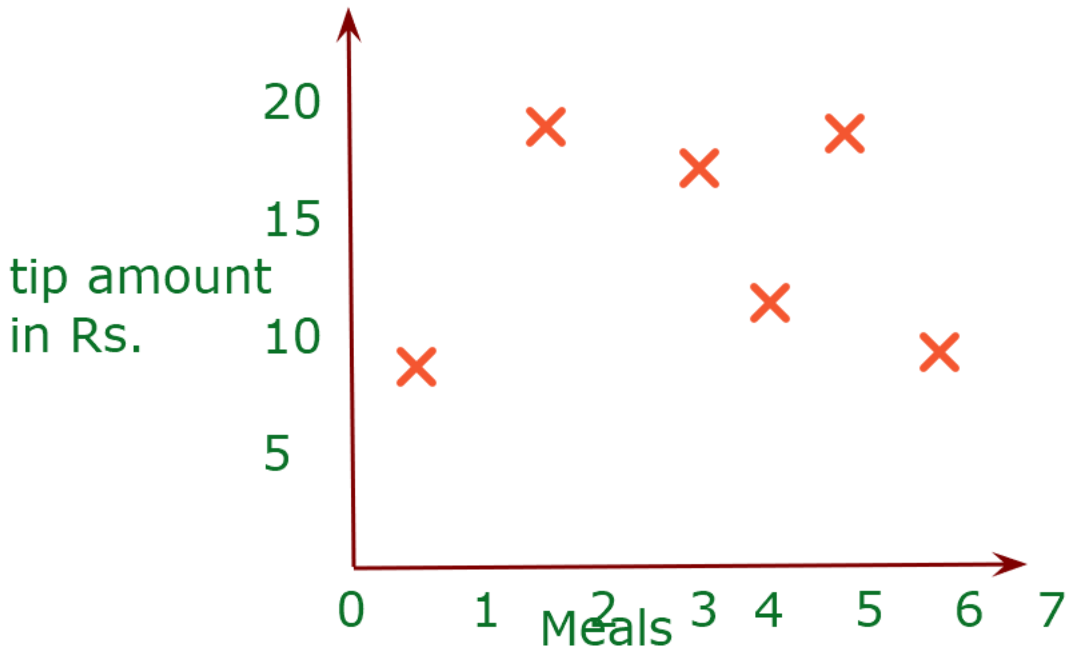


Linear Regression

Finding the value of dependent variable
Independent variables
Mathematical model: Y = f(X)

Q1. X:=?
7, 19, 13, 10, 17, & 7: tip amount
tip amount for meal X=?
Can we predict the tip amount:?

yes, tip amount for meal#7



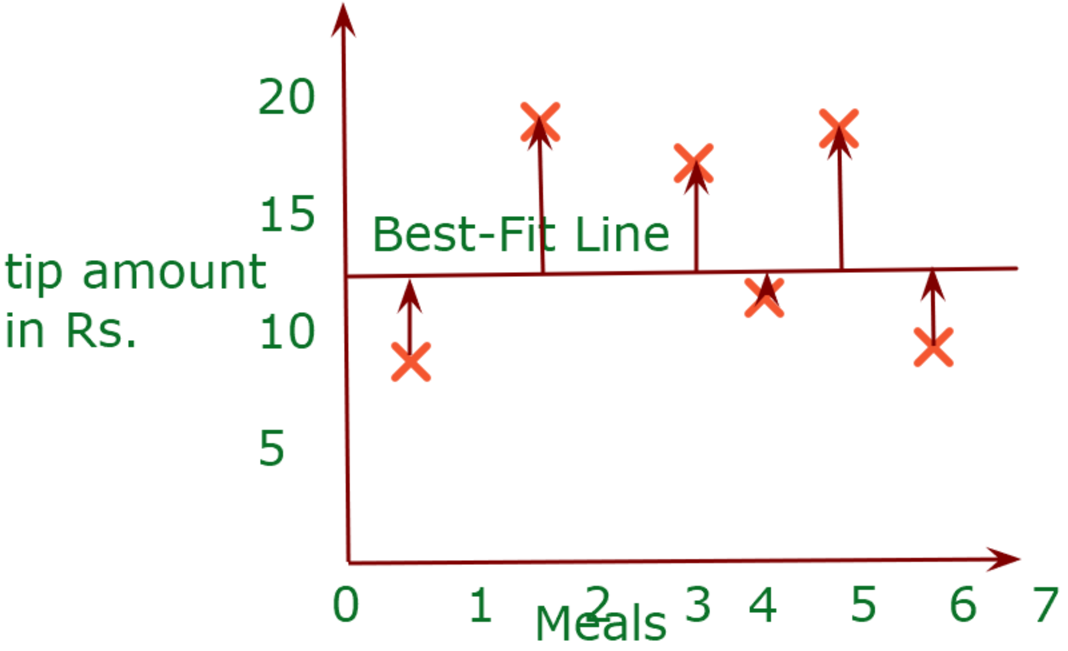
Linear Regression

Finding the value of dependent variable
Independent variables
Mathematical model: Y = f(X)

Q1. X:=?
7, 19, 13, 10, 17, & 7: tip amount
tip amount for meal X=?
Can we predict the tip amount:?

yes, tip amount for meal#7

Mean is the best predictor when you do not have any other information available in the dataset.



Mean = 12.16: 12

Errors

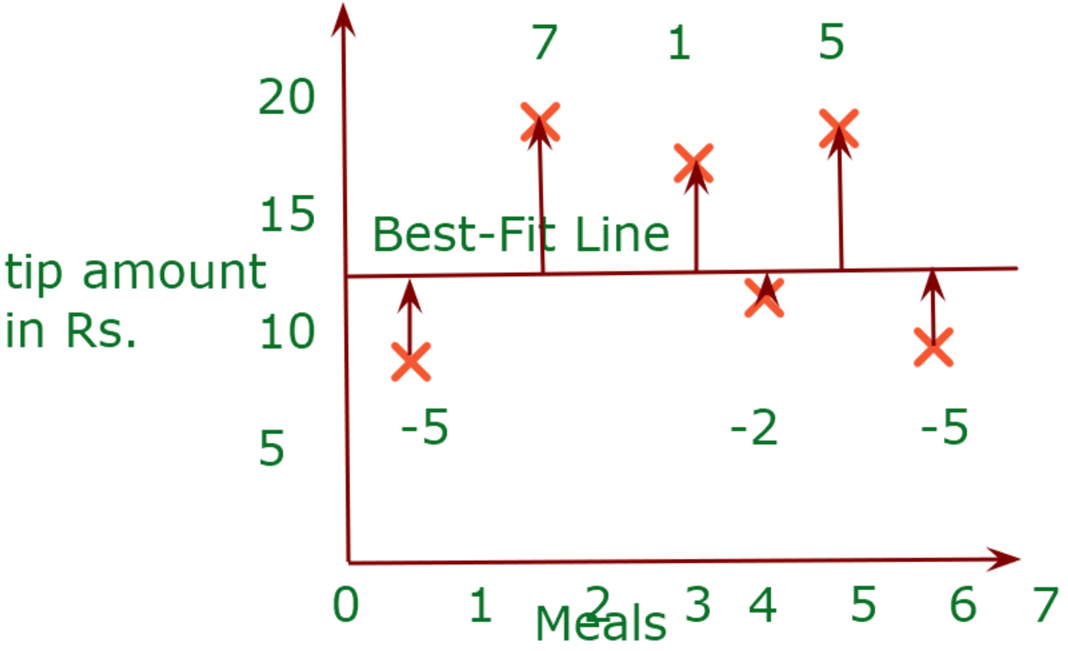
Linear Regression

Finding the value of dependent variable
Independent variables
Mathematical model: Y = f(X)

Q1. X:=?
7, 19, 13, 10, 17, & 7: tip amount
tip amount for meal X=?
Can we predict the tip amount:?

yes, tip amount for meal#7

Mean is the best predictor when you do not have any other information available in the dataset.



Mean = 12.16: 12

Errors

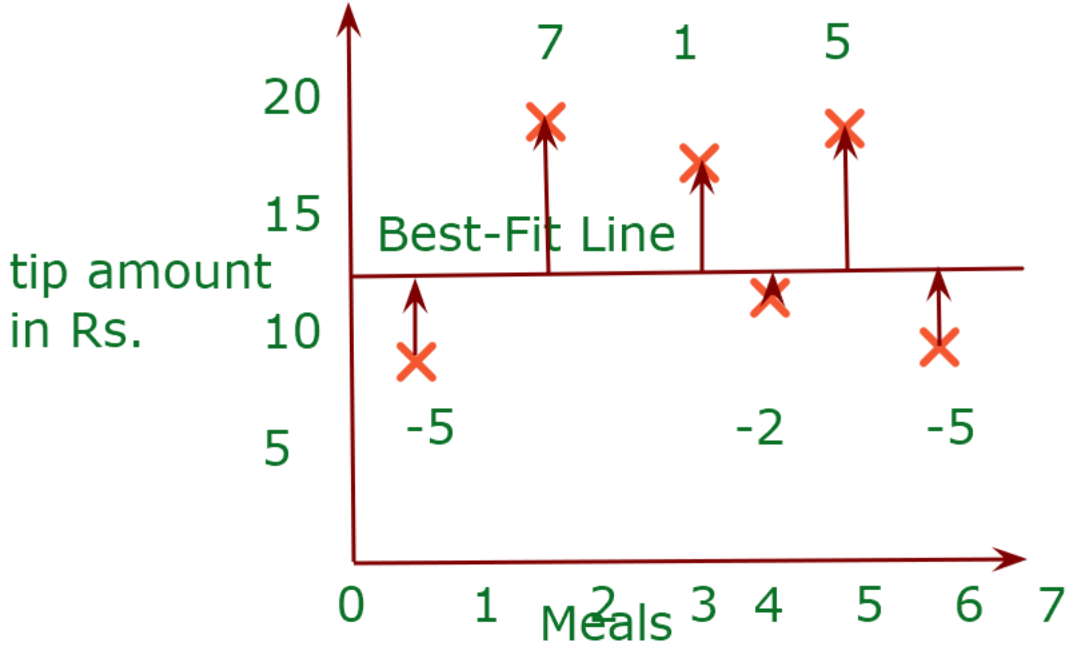
Linear Regression

Finding the value of dependent variable
Independent variables
Mathematical model: Y = f(X)

Q1. X:=?
7, 19, 13, 10, 17, & 7: tip amount
tip amount for meal X=?
Can we predict the tip amount:?

yes, tip amount for meal#7

Mean is the best predictor when you do not have any other information available in the dataset.



Mean = 12.16: 12

Errors

to make all the errors positive
for larger deviations

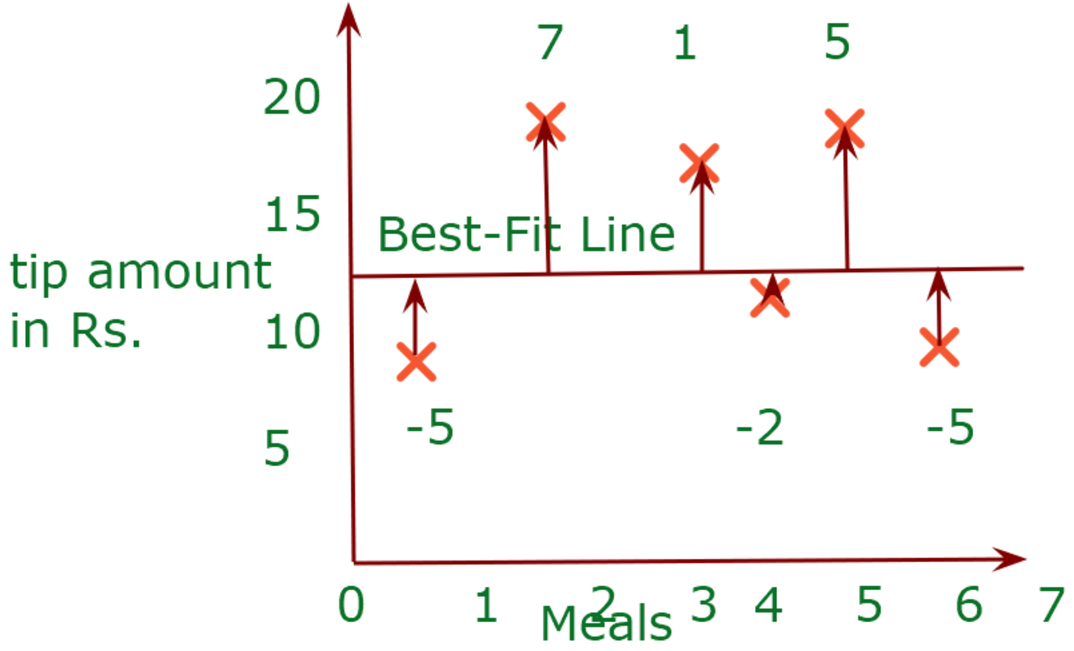
Linear Regression

Finding the value of dependent variable
Independent variables
Mathematical model: Y = f(X)

Q1. X:=?
7, 19, 13, 10, 17, & 7: tip amount
tip amount for meal X=?
Can we predict the tip amount:?

yes, tip amount for meal#7

Mean is the best predictor when you do not have any other information available in the dataset.



Mean = 12.16: 12

SSE= 129

Errors

to make all the errors positive for larger deviations

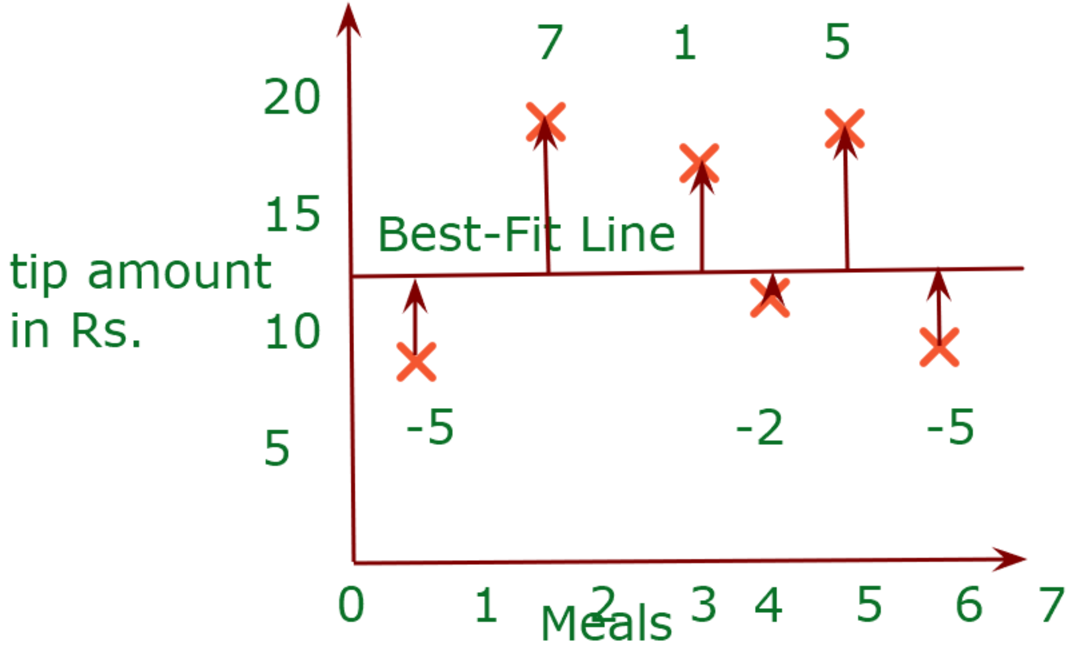
Variability in tip amount that we can observe, it is only going to be there because tip amount itself

Linear Regression

Finding the value of dependent variable
Independent variables
Mathematical model: Y = f(X)

Q1. X:=?
7, 19, 13, 10, 17, & 7: tip amount
tip amount for meal X=?
Can we predict the tip amount:?

yes, tip amount for meal#7



Mean is the best predictor when you do not have any other information available in the dataset.

Mean = 12.16: 12

SSE= 129

Errors

to make all the errors positive
for larger deviations

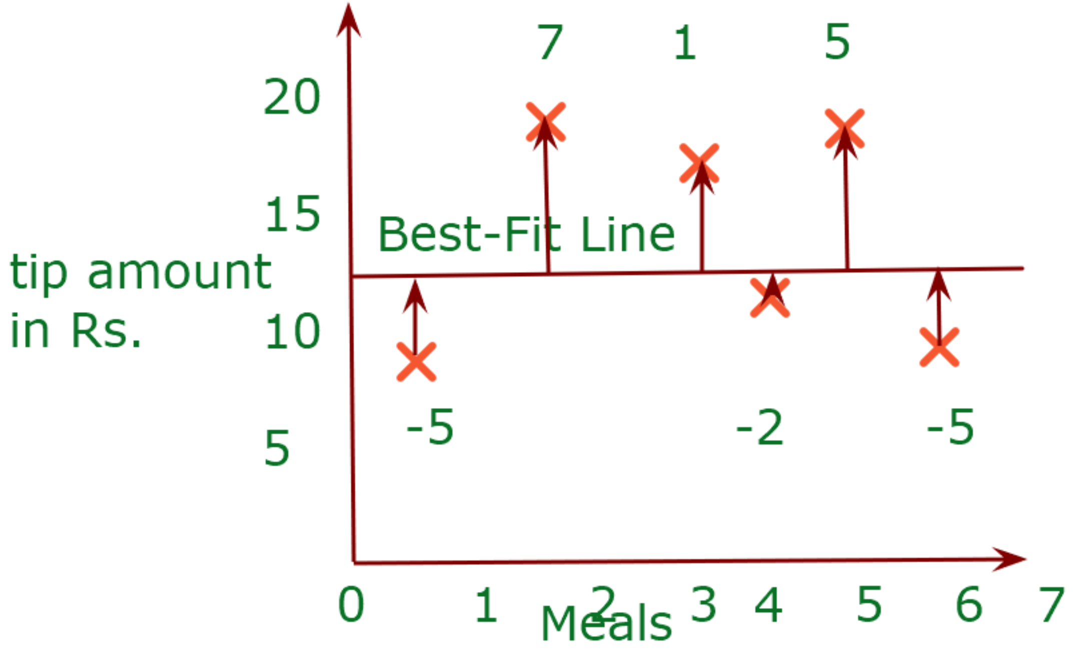
Variability in tip amount that we can observe, it is only going to be there because tip amount itself

Linear Regression

Finding the value of dependent variable
Independent variables
Mathematical model: Y = f(X)

Q1. X:=?
7, 19, 13, 10, 17, & 7: tip amount
tip amount for meal X=?
Can we predict the tip amount:?

yes, tip amount for meal#7



Mean is the best predictor when you do not have any other information available in the dataset.

Mean = 12.16: 12

SSE= 129

Errors

to make all the errors positive
for larger deviations

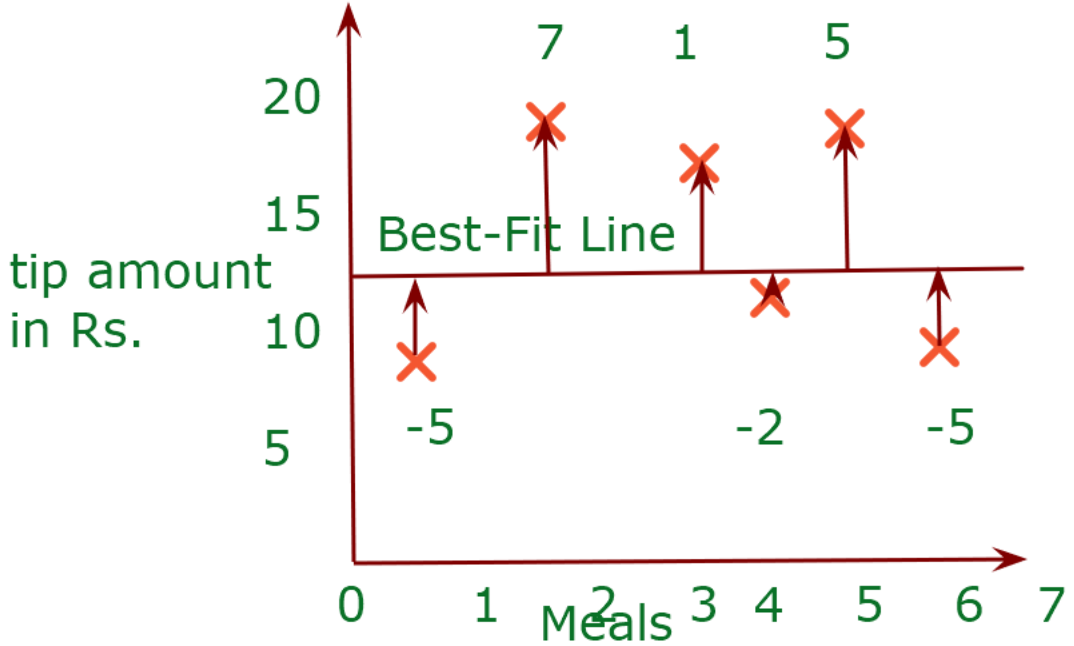
Variability in tip amount that we can observe, it is only going to be there because tip amount itself

Linear Regression

Finding the value of dependent variable
Independent variables
Mathematical model: Y = f(X)

Q1. X:=?
7, 19, 13, 10, 17, & 7: tip amount
tip amount for meal X=?
Can we predict the tip amount:?

yes, tip amount for meal#7



Mean is the best predictor when you do not have any other information available in the dataset.

Mean = 12.16: 12

SSE= 129

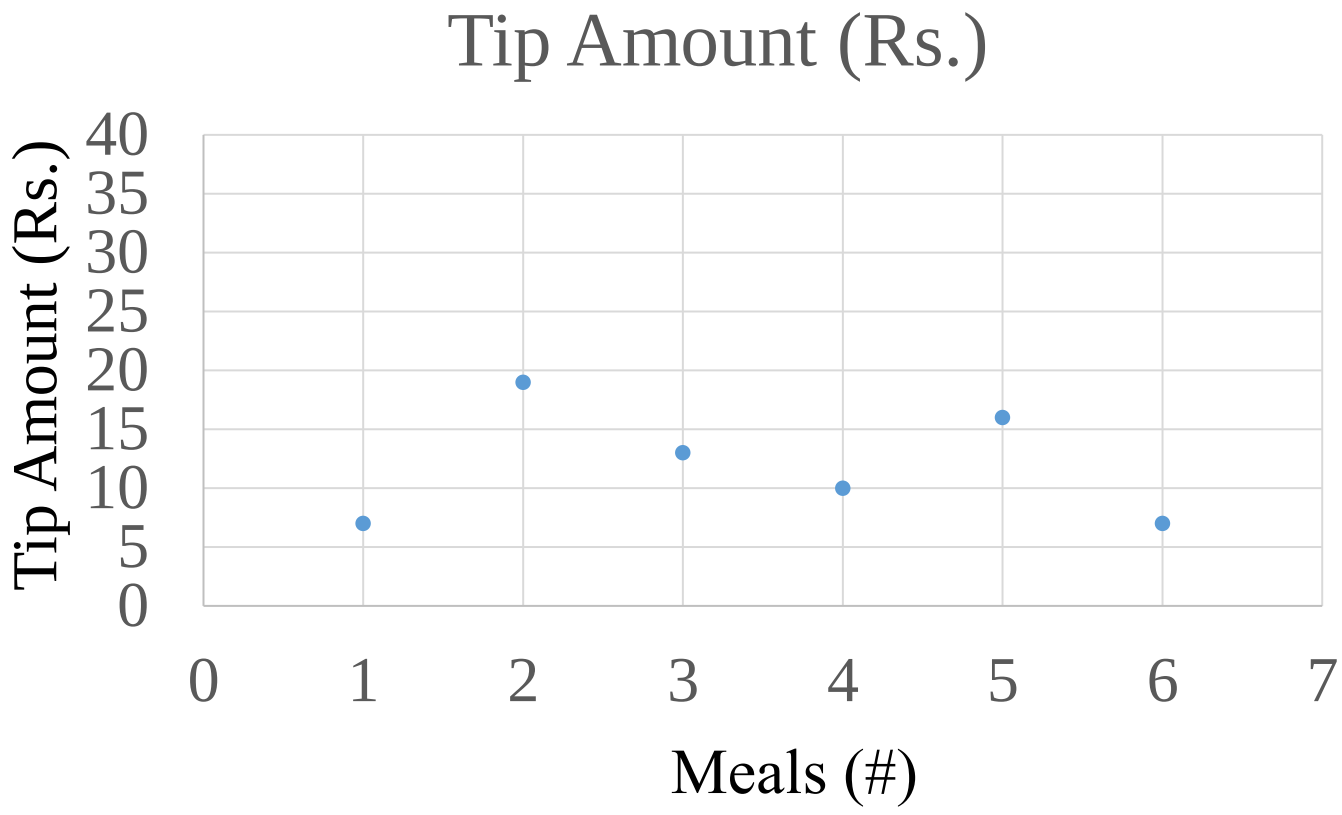
Errors

to make all the errors positive
for larger deviations

Collected Data for Service

Meal (#)	Tip Amount (in Rs.)
1.	7
2.	19
3.	13
4.	10
5.	16
6.	7
7.	**?**

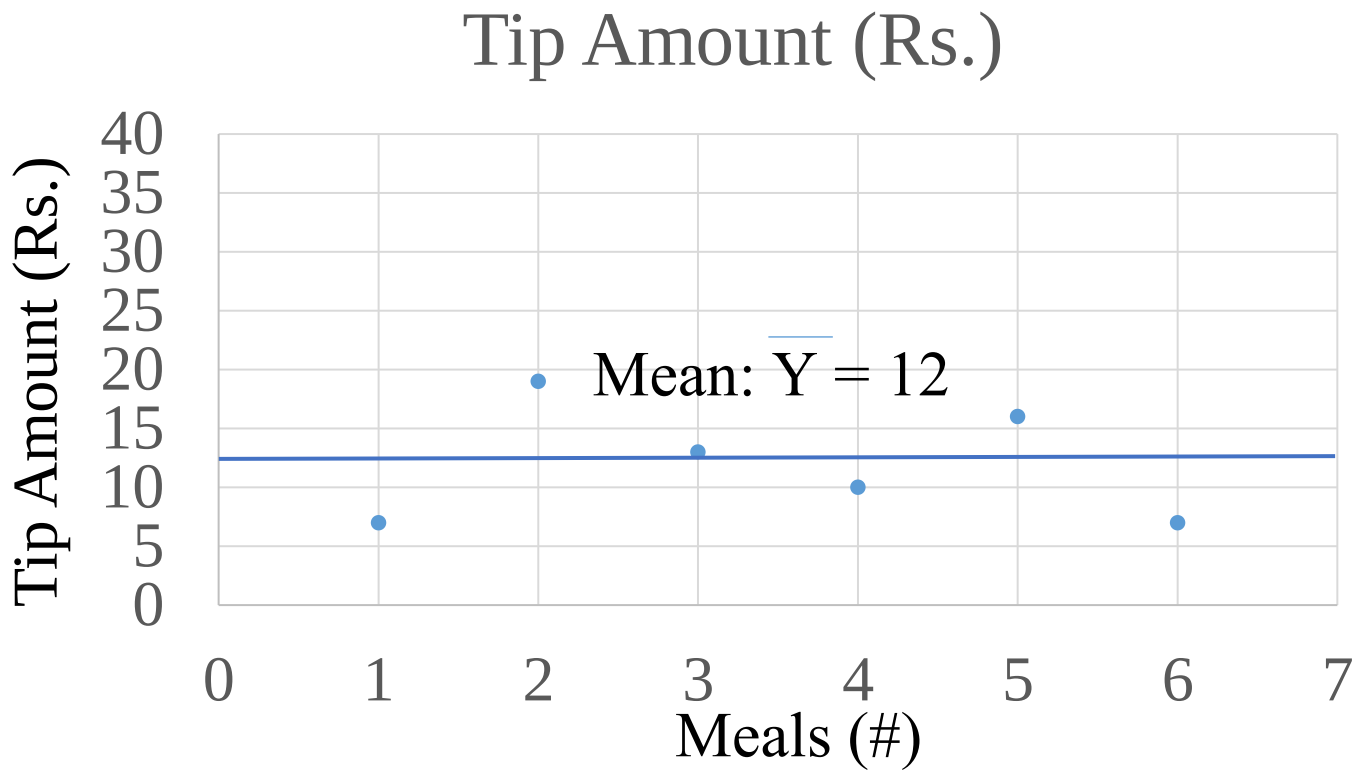
Scatter Plot: Visualize the data to observe the pattern



Collected Data for Service

Meal (#)	Tip Amount (in Rs.)
1.	7
2.	19
3.	13
4.	10
5.	16
6.	7
7.	**?**

Best Predictor?
Can we use Mean of the Tip Amount?

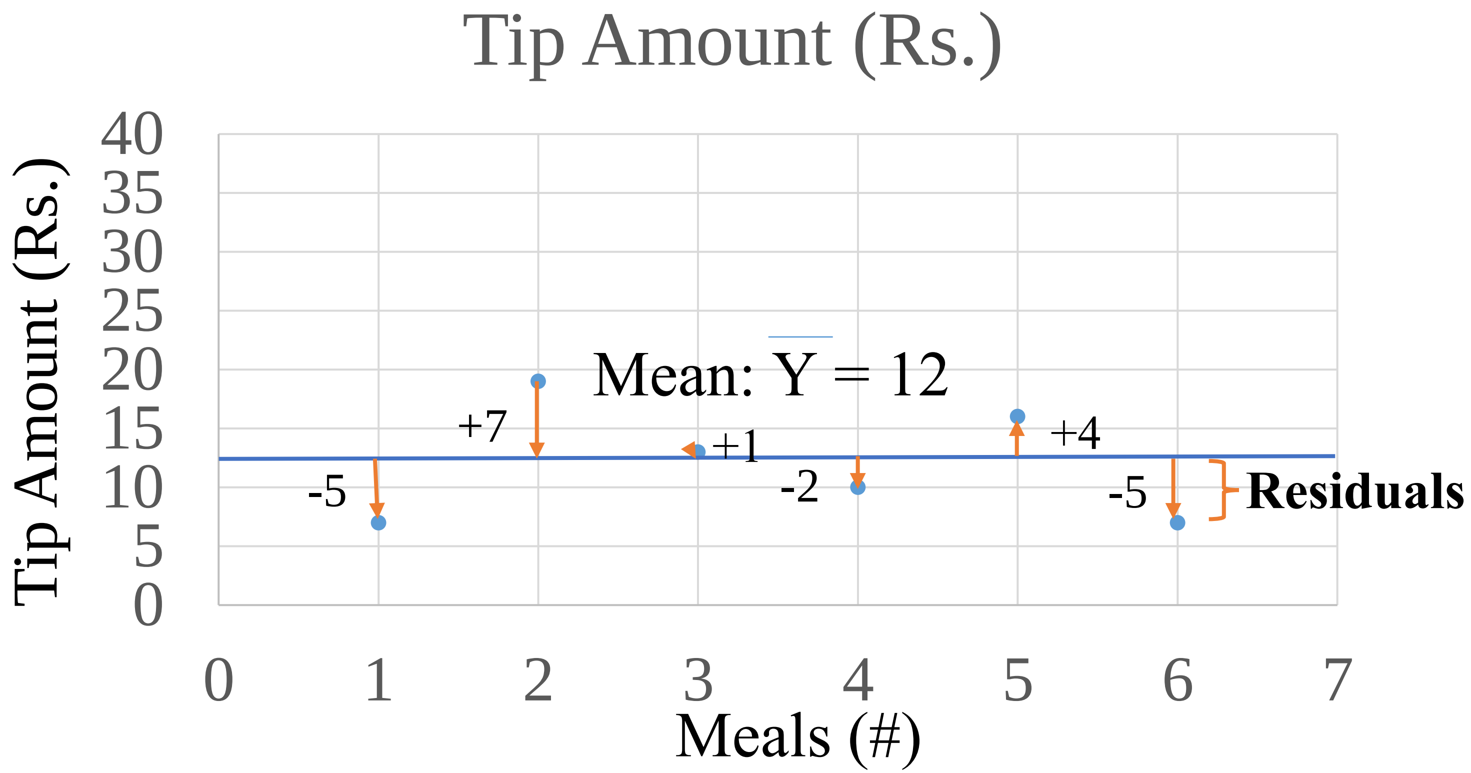


- Mean is the best estimate for predicting the tip amount when no other information is available with us, the variability in the tip amount can only be explained by the tips themselves

"Goodness of Fit" for the Tips

Meal (#)	Tip Amount (in Rs.)
1.	7
2.	19
3.	13
4.	10
5.	16
6.	7
7.	?

Mean of Tips: Some data points are below and some are above it

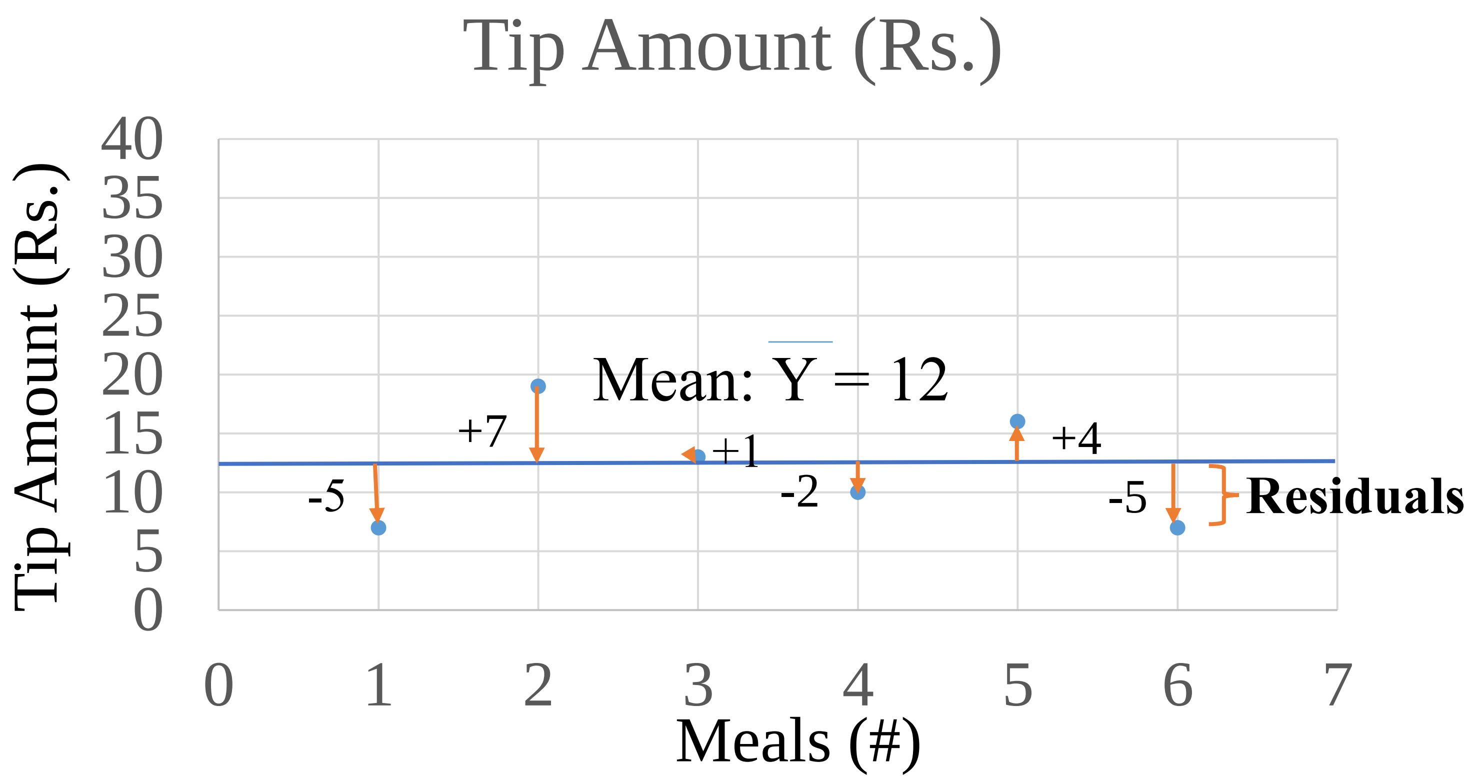


- ➢ The distance every data point is from mean is called as **residuals**
- ➢ The distance from the "**best fit line**" to the observed values are called as "**residuals or errors**"

“Goodness of Fit” for the Tips

Meal (#)	Tip Amount (in Rs.)	Predicted Tip amount (in Rs.)	Difference or error
1.	7	12	25
2.	19	12	49
3.	13	12	1
4.	10	12	4
5.	16	12	16
6.	7	12	25

This is the “best fit line”



- **Residuals** will add up & gives **zero: (i.e. -5 +7 +1 -2 +4 -5 = 0)**
- Residual square: **make them positive** and to **emphasize on larger deviations**
- **Sum of Squared errors (SSE) = 120/-**

Basic Algebra

Slope – intercept form of a line

$y = mx + b$?

x = random variable

m = slope of the line rise/run

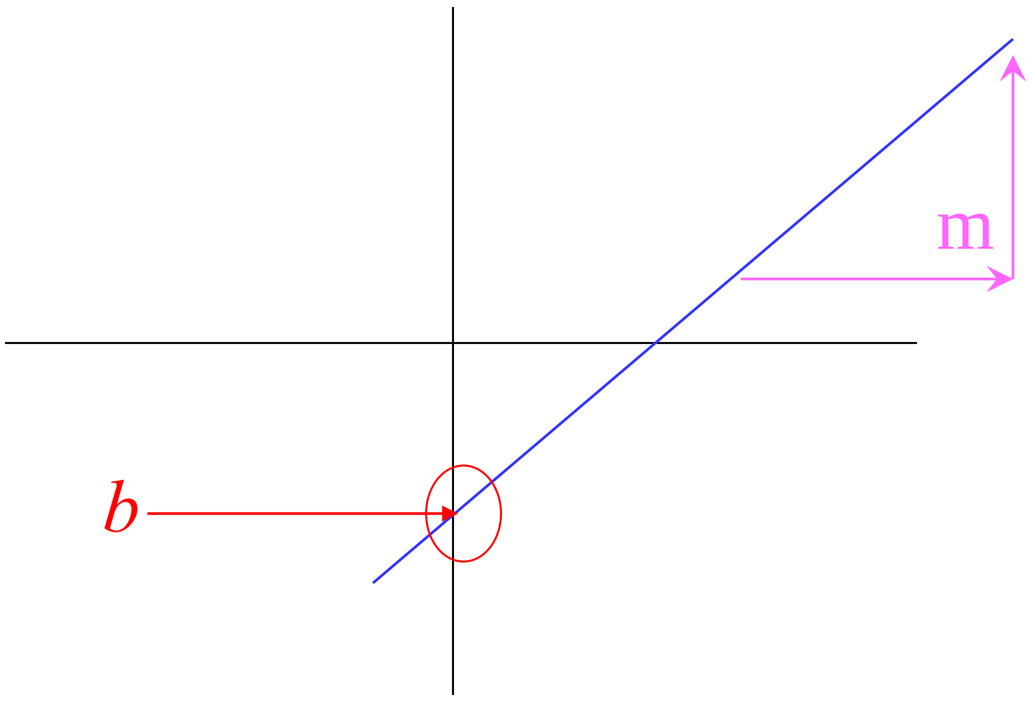
B= intercept where x = 0

$Y = \beta 0 + \beta 1x + €$

$\hat{y}_i = \quad b0 + b^*x_i + \text{error}$

$E(Y) = \beta 0 + \beta 1x$

Mean or expected value of y

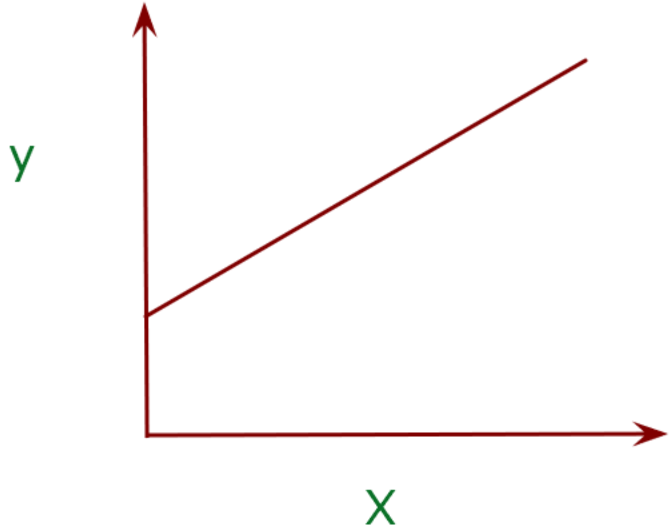


$Y = mx + b$

relationship between the regression models

y hat = beta sub 0 + beta sub 1 (xi) + errors

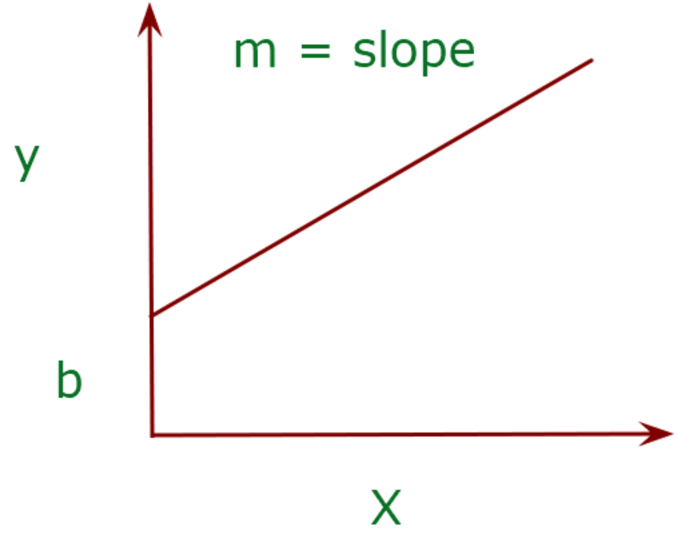
Y = mx + b
relationship between the regression models
y hat = beta sub 0 + beta sub 1 (xi) + errors



$Y = mx + b$
relationship between the regression models
y hat = beta sub 0 + beta sub 1 (xi) + errors

$Y = 3X + 6$

Slope of the line = 3

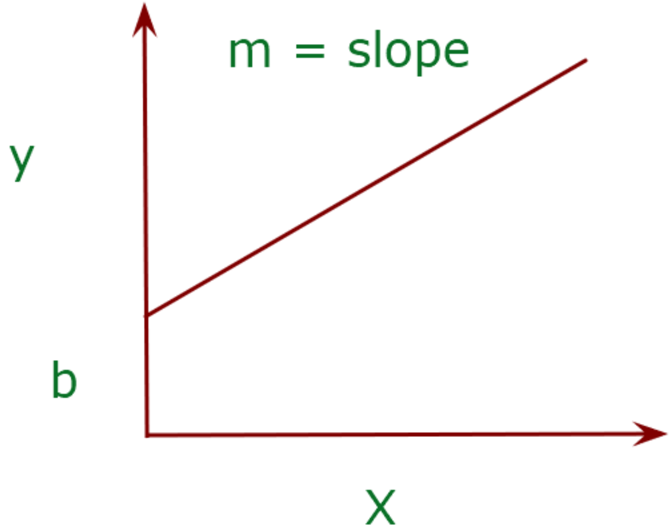


$Y = mx + b$
relationship between the regression models
y hat = beta sub 0 + beta sub 1 (xi) + errors

$Y = 3X + 6$

Slope of the line = 3

Y = dependent variable; X = independent variable



$Y = mx + b$

relationship between the regression models

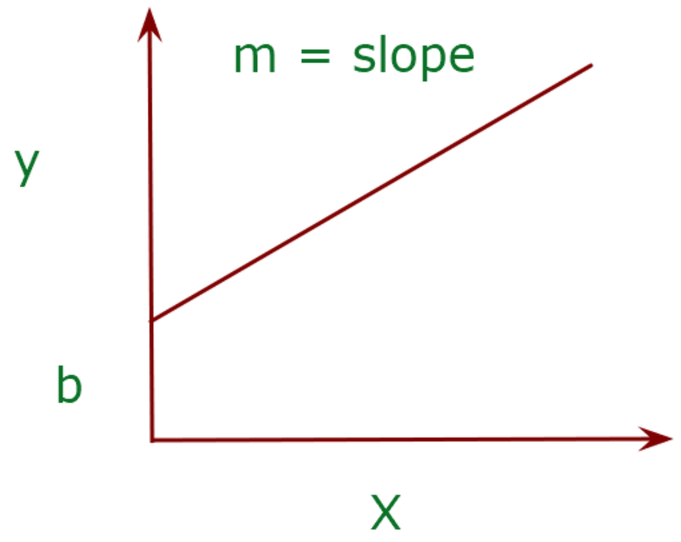
y hat = beta sub 0 + beta sub 1 (xi) + errors

$Y = 3X + 6$

Slope of the line = 3

Y = dependent variable; X = independent variable

errors= unexplained variation in variable 'y'



$Y = mx + b$
relationship between the regression models
y hat = beta sub 0 + beta sub 1 (xi) + errors

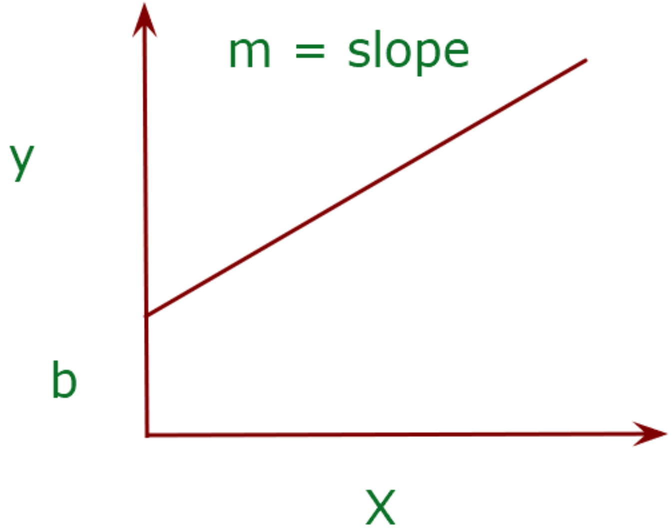
$Y = 3X + 6$

Slope of the line = 3

Y = dependent variable; X = independent variable

errors= unexplained variation in variable 'y'

y hat = expected value of y for a given value of x



Y = mx + b
relationship between the regression models
y hat = beta sub 0 + beta sub 1 (xi) + errors

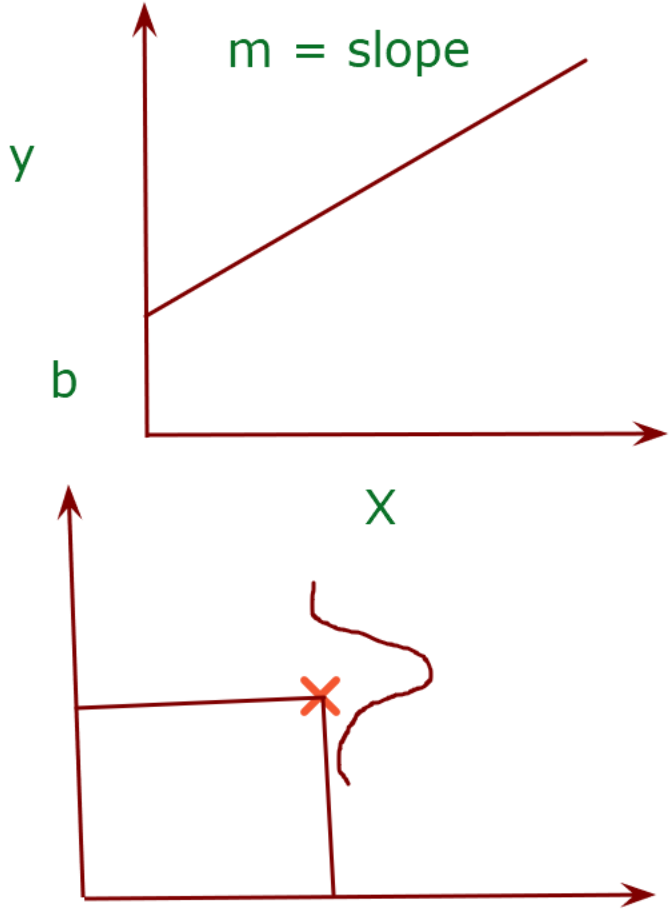
Y = 3X + 6

Slope of the line = 3

Y = dependent variable; X = independent variable

errors= unexplained variation in variable 'y'

y hat = expected value of y for a given value of x



Y = mx + b
relationship between the regression models
y hat = beta sub 0 + beta sub 1 (xi) + errors

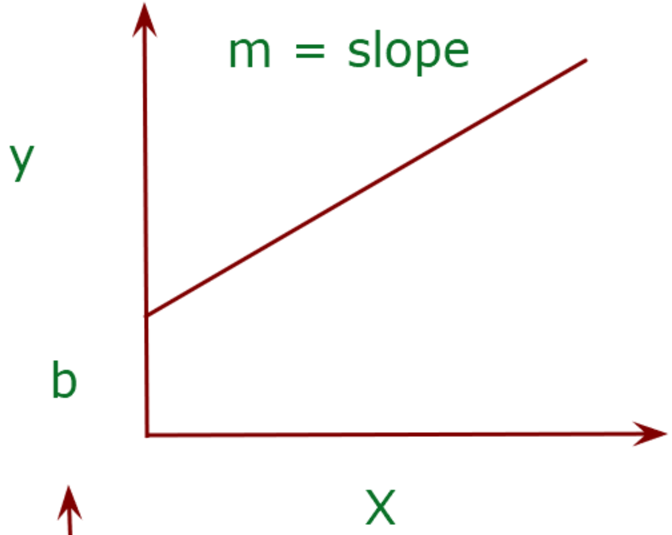
$Y = 3X + 6$

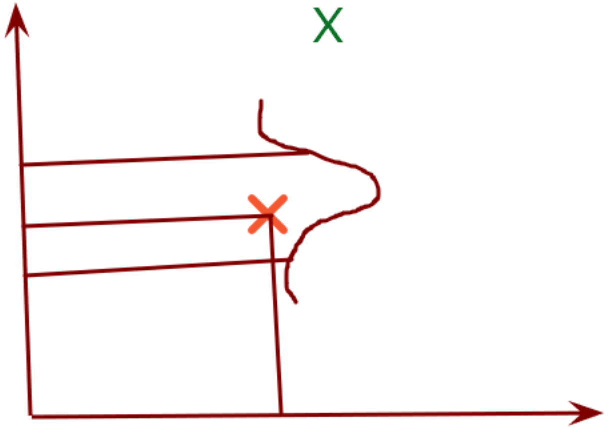
Slope of the line = 3

Y = dependent variable; X = independent variable

errors= unexplained variation in variable 'y'

y hat = expected value of y for a given value of x





$Y = mx + b$
relationship between the regression models
y hat = beta sub 0 + beta sub 1 (xi) + errors

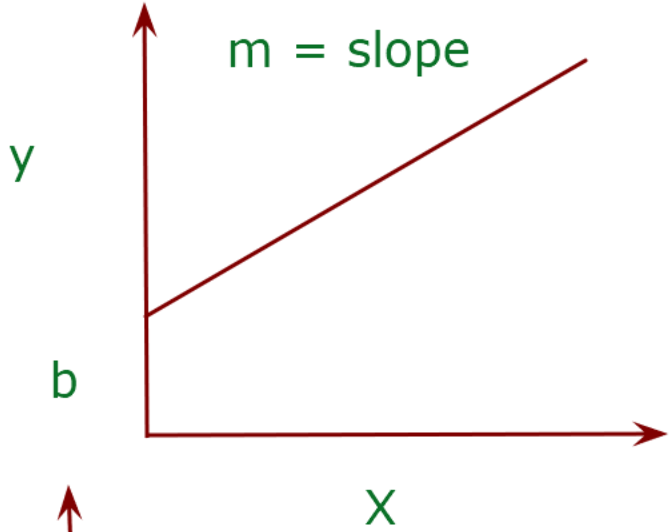
$Y = 3X + 6$

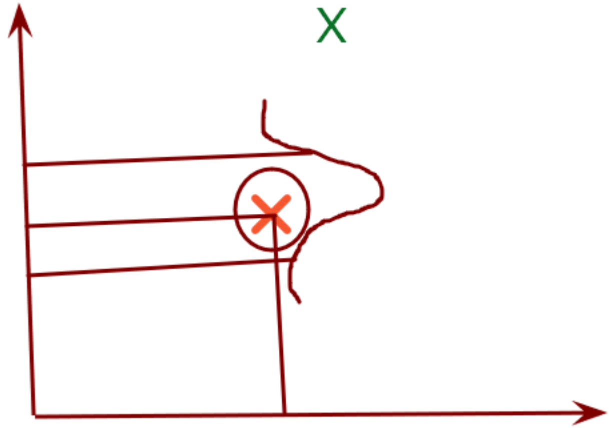
Slope of the line = 3

Y = dependent variable; X = independent variable

errors= unexplained variation in variable 'y'

y hat = expected value of y for a given value of x





$Y = mx + b$
relationship between the regression models
y hat = beta sub 0 + beta sub 1 (xi) + errors

$Y = 3X + 6$

Slope of the line = 3

Y = dependent variable; X =
independent variable

errors= unexplained variation in variable 'y'

y hat = expected value of y for a given value of x

mean value of the distribution