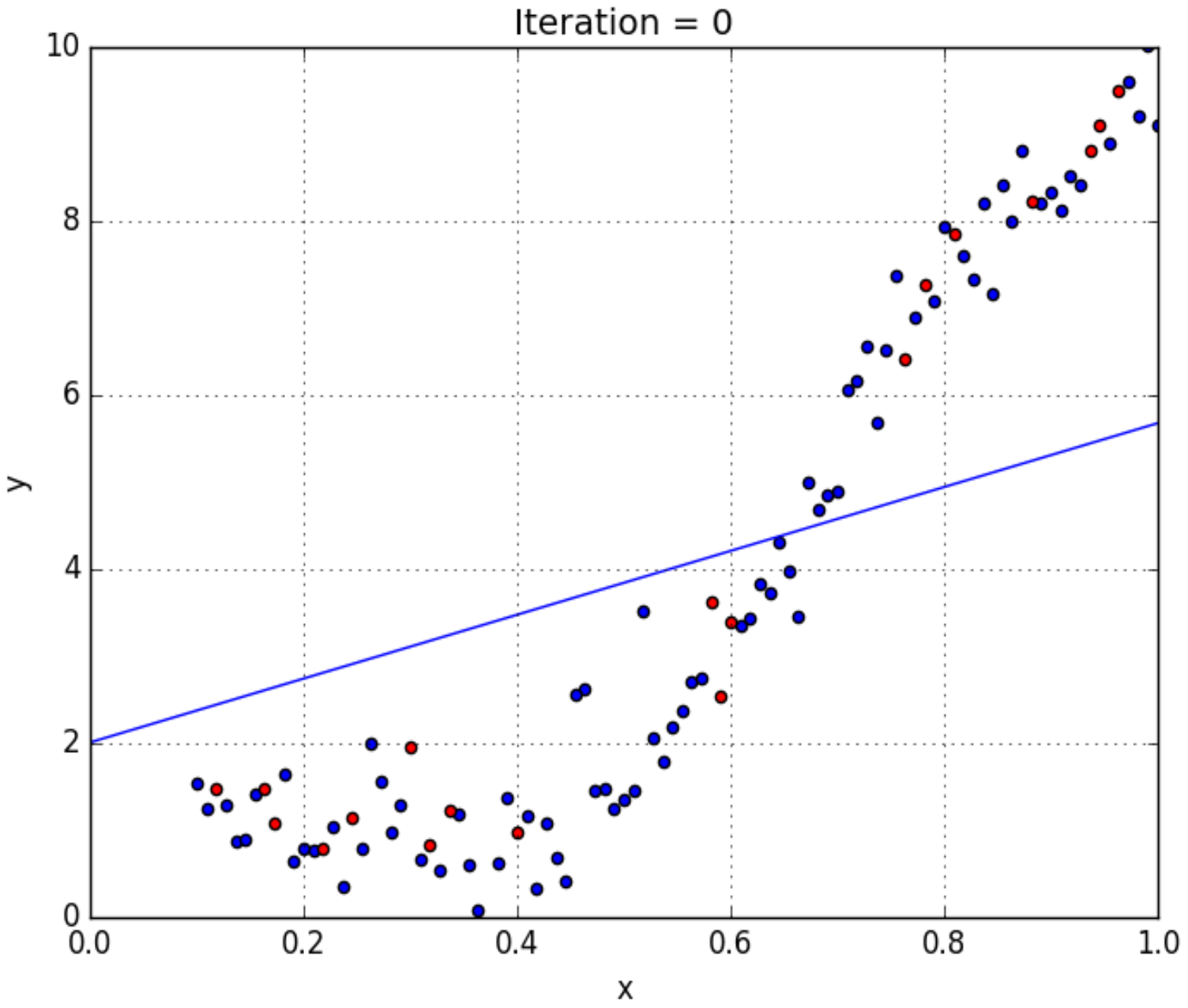
STEPS TO ESTABLISH A REGRESSION

By Dr. Ujwala Bharambe

Acknowledgement : Prof Anuradha Srinivasaraghavan
St. Francis Institute of Technology


Iteration = 0
10
8
6
4
2
0
y
0.0
0.2
0.4
0.6
0.8
1.0
x

TIPS FOR SERVICE

Let us Assume that you are waiter at nice restaurant. 'Tips' are very important part of the waiters pay. Most of the time amount of TIP is related to the Amount of Total Bill.

As a waiter , you would like to develop a model , that will allow you to prediction about what amount of tip to expect for any given Bill amount. Therefore one evening you collect data for six meals.

TIPS FOR SERVICE

Unfortunately when you begin to look at your data , you realize you only collected data for TIP amount and not for the meal amount also!

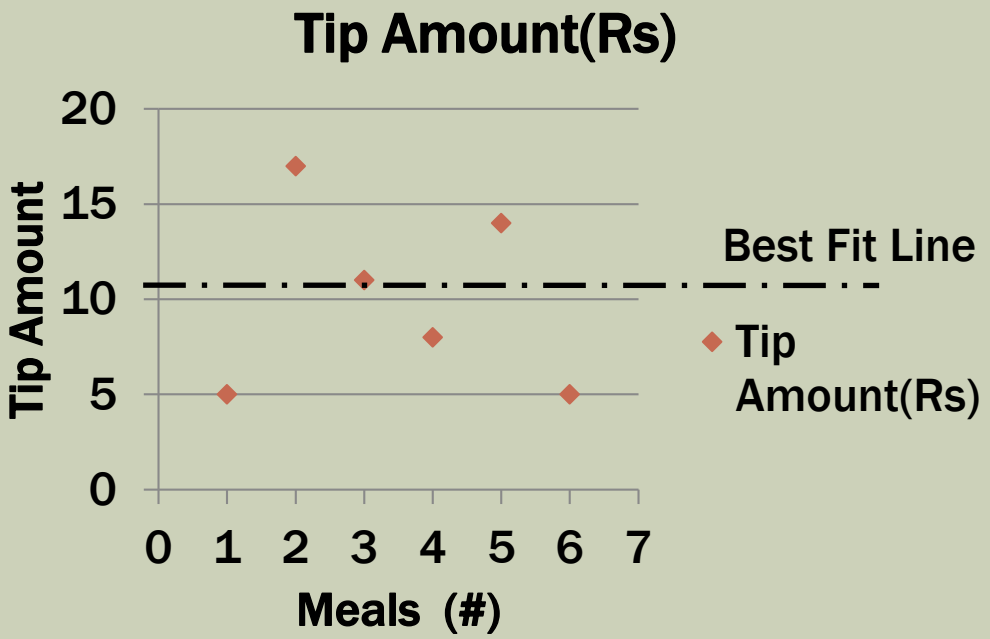
Meal(#)	TIP Amount (Rs.)
1	5.00
2	17.00
3	11.00
4	8.00
5	14.00
6	5.00

So this is the best data you have !

How might you predict the tip amount for future meals using only this data ?

TIPS FOR SERVICE

Meal(#)	TIP Amount (Rs.)
1	5.00
2	17.00
3	11.00
4	8.00
5	14.00
6	5.00

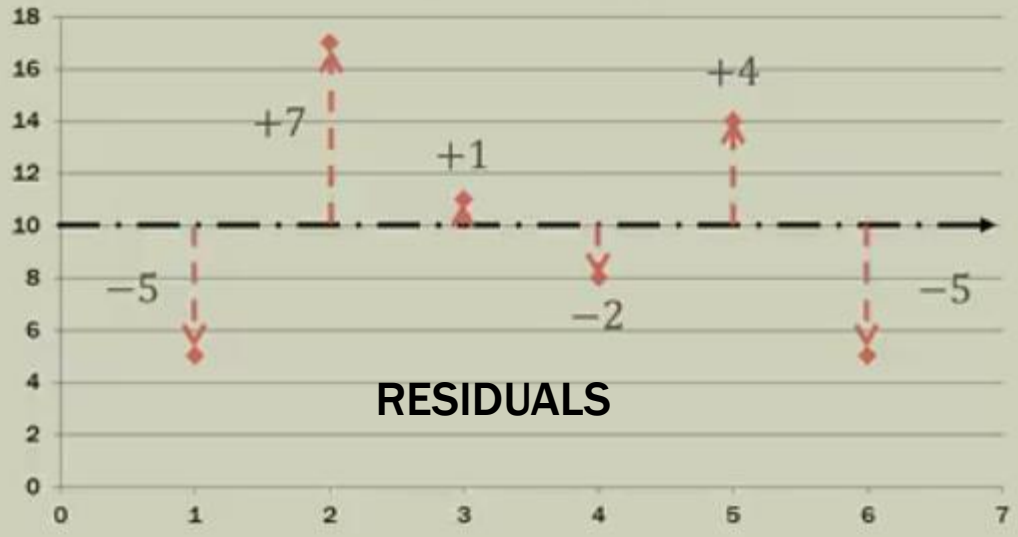


$$\bar{y} = 10$$

With Only one variable, and no other information, the best prediction for next measurement is the mean of sample itself. The variability in Tips amount can be explained by the tips them selves !

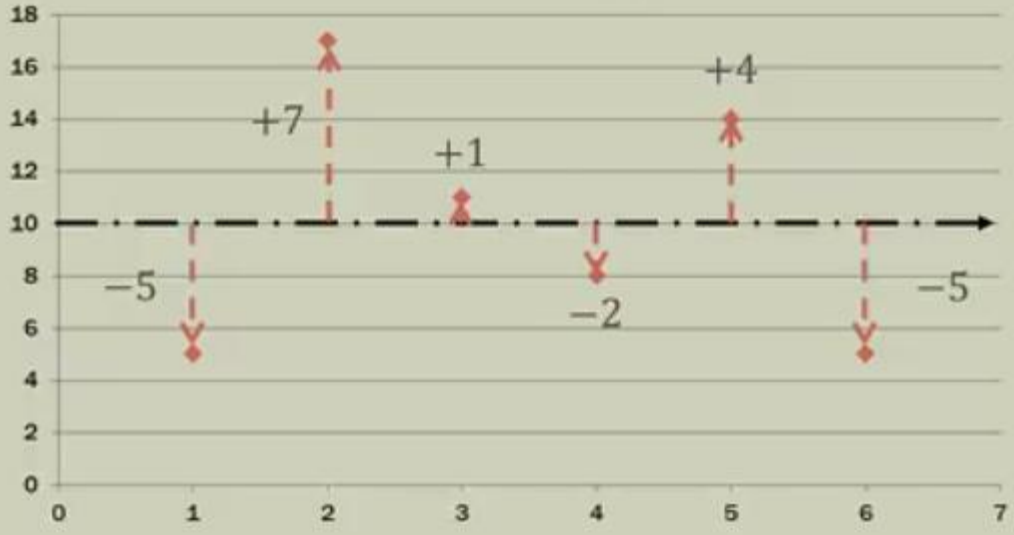
"GOODNESS OF FIT" FOR THE TIPS

Meal(#)	TIP Amount (Rs.)
1	5.00
2	17.00
3	11.00
4	8.00
5	14.00
6	5.00



$$\bar{y} = 10$$

SQUARING THE RESIDUALS(ERRORS)



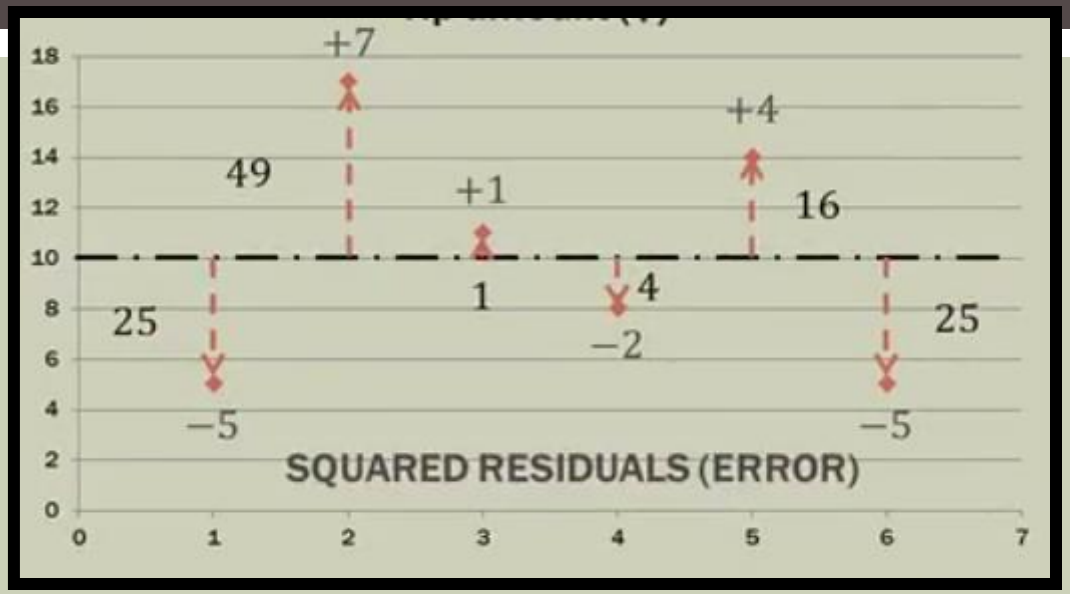
Meal#	Residual	Residual^2
1	−5	25
2	+7	49
3	+1	1
4	−2	4
5	+4	16
6	−5	25

Why square the residuals ?

1. they become positive
2. emphasizes larger **deviations**.

Sum of Squared Errors (SSE) = 120

SQUARING THE RESIDUALS(ERRORS)



The Goal of Simple Linear regression is to create a linear model that minimizes the sum of squares of residuals / Errors (SSE)

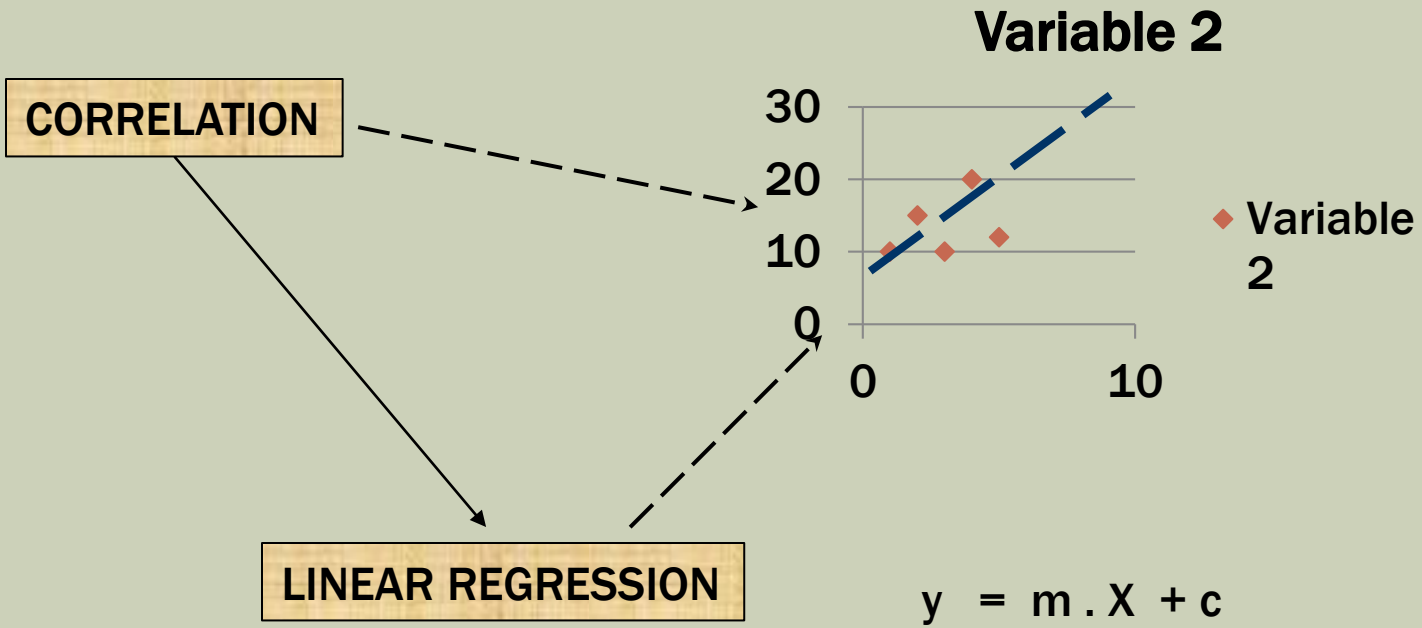
If Our regression model is Significant it will "Eat up" much of the raw SSE.

OUR AIM SHOULD BE TO DEVELOP A REGRESSION LINE WHICH WILL FIT BETTER WITH MINIMUM SSE

QUICK REVIEW

- Simple Linear regression is really a comparison of 2 Models
  - One is where independent variable does not exist.
  - And the other uses best fit regression line.
- If there is only one variable , the best prediction is mean.
- The difference between actual value and predicted value is RESIDUAL or ERROR.
- The RESIDUALS ARE SQUARED to Obtain SSE.
- Simple linear regression is designed to find the best fitting line through data to minimizes the SSE.

SIMPLE LINEAR REGRESSION:



Value of one variable , is a function of the other variable.
Value of *y* is a function of *x i.e. y = f(x)*
X is independent variable , Y is Dependent variable

CORRELATION

1. Relationship 
2. Movement together 

REGRESSION

One variable affects the other 

cause and effect 

learn.g2.com

Differences Between Correlation and Regression

Correlation

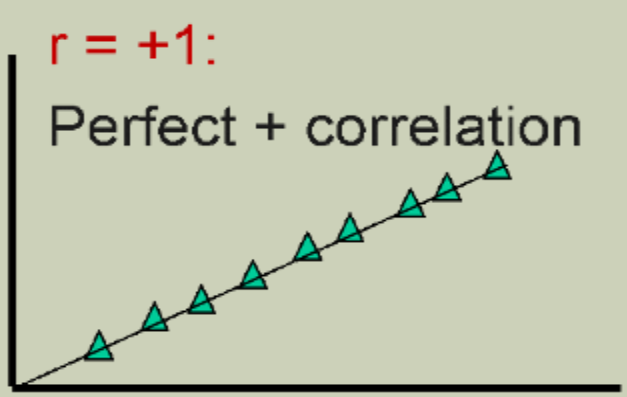
1. Relationship
2. Variables move together
3. x and y can be interchanged
4. Data represented in single point

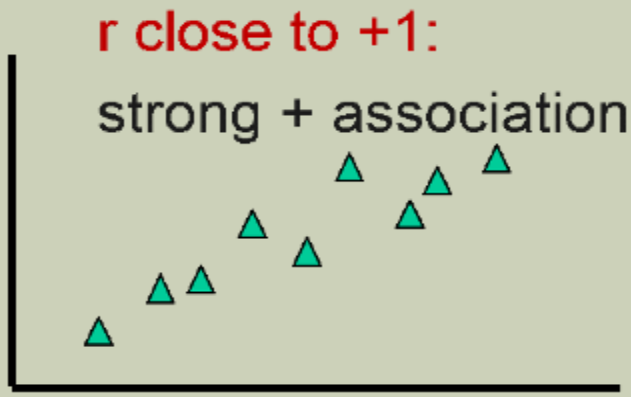
Regression

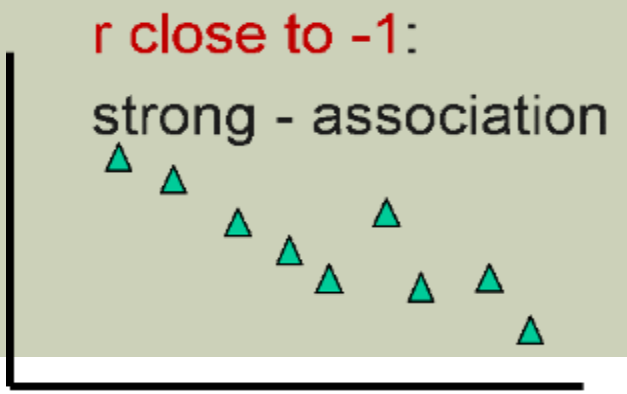
1. One affects the other
2. Cause and effect
3. x and y cannot be interchanged
4. Data represented by line

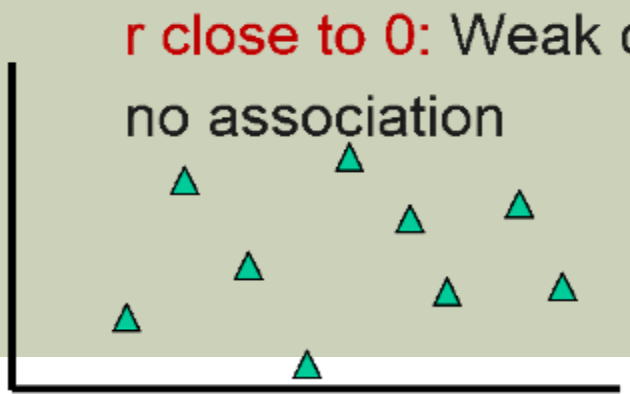
CORRELATION

- **Correlation** is a statistical term describing the degree to which two variables move in coordination with one another.

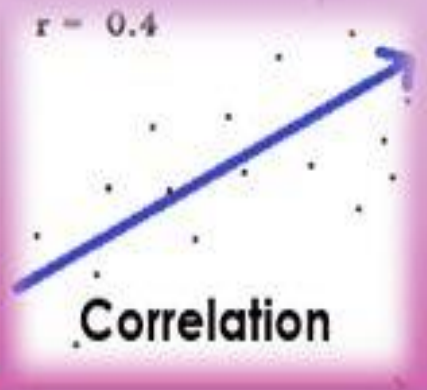








Differences Between Pearson Correlation and Linear Regression

	Use when...	Results	Assumptions (Requirements)
 Correlation	You want to know if there's **an association between two variables.**	**Strength & direction of relationship (r):** -1 = perfect negative, +1. = perfect positive, 0 = no correlation.	**Validity:** Valid measurements, a good sample, unconfounded comparisons. **Distribution:** Linear relationship between the two.
 Regression	You want to predict **how one variable will change with another.**	**Estimates of parameters for a regression equation:** The B0 and B1 in the linear regression equation $Y = B0 + B1 * X$	**All the above, plus:** 1. Quantitative Data 2. Outlier Condition 3. Independence of Errors 4. Homoscedasticity 5. Normality of Error Distribution

SIMPLE LINEAR REGRESSION MODEL

$y = m \cdot X + b$

$$y = \beta_0 + \beta_1 X + \varepsilon$$

Where

β_0 = y intercept population parameter

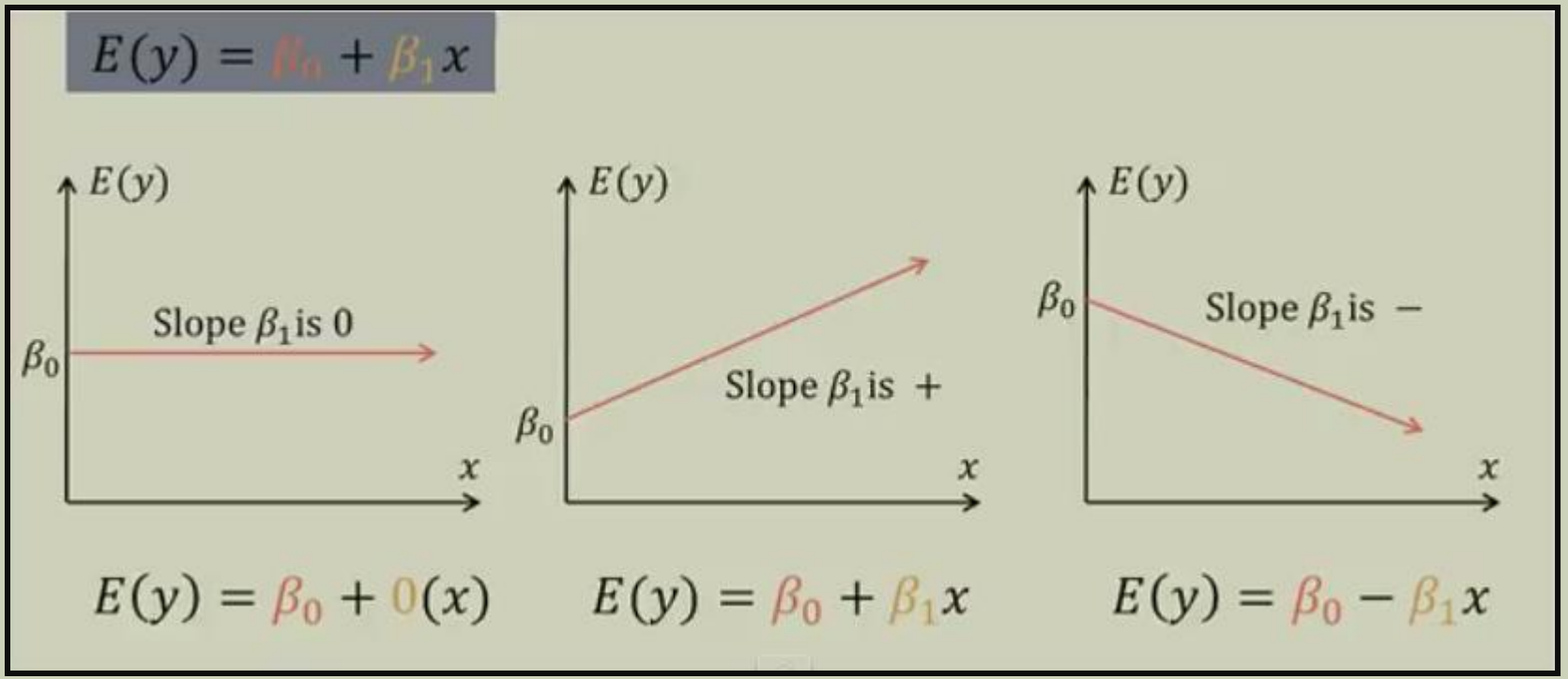
β_1 = slope population parameter

ε = error term in y variation (unknown reasons)

Simple regression model is $E(y) = \beta_0 + \beta_1 X$

GENERAL REGRESSION LINES

- Accuracy of **E(y)** values depends upon mean value of y and also on **distribution of y .**



REGRESSION EQUATION WITH ESTIMATES

If we actually knew the population parameters, β_0 and β_1, we could use the Simple Linear Regression Equation.

$$E(y) = \beta_0 + \beta_1 x$$

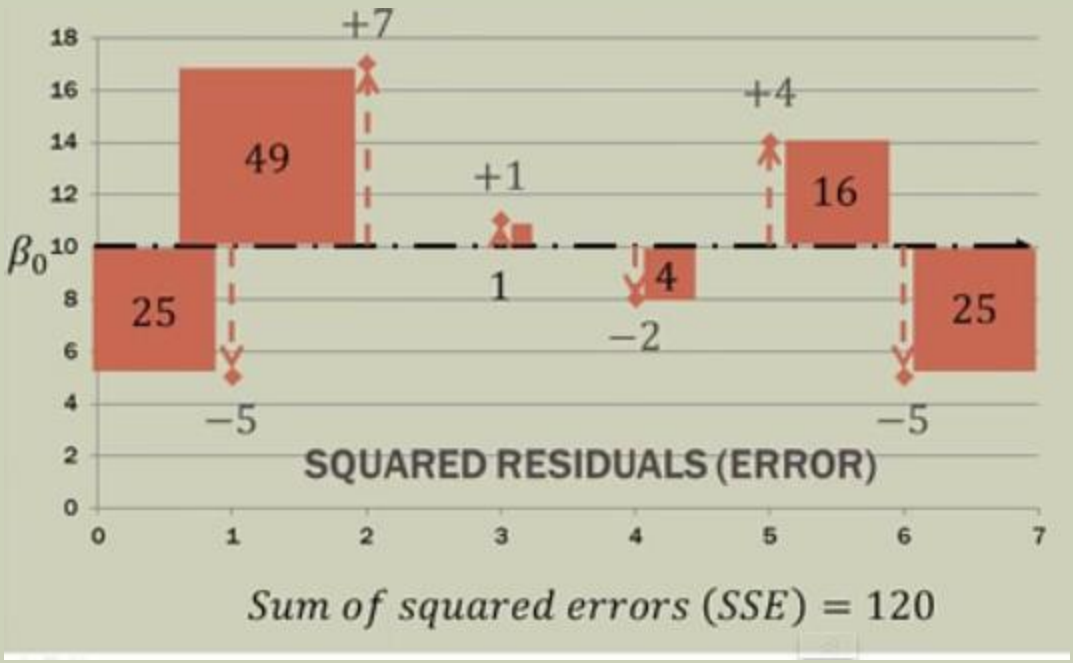
In reality we almost never have the population parameters. Therefore we will estimate them using sample data. When using sample data, we have to change our equation a little bit.

$\hat{y}$, pronounced "y-hat" is the point estimator of $E(y)$

$$\hat{y} = b_0 + b_1 x$$

$\hat{y}$, is the **mean value** of y for a given value of x.

WHEN THE SLOPE , $B_1 = 0$



When conducting simple linear regression with TWO variables, we will determine how good the regression line "fits" the data by comparing it to THIS TYPE; where we pretend the second variable does not even exist; the slope, $\beta_1 = 0$.

In this situation, the value of $\hat{y}$ is 10 for every value of x.

$\hat{y} = b_0 + b_1 x$

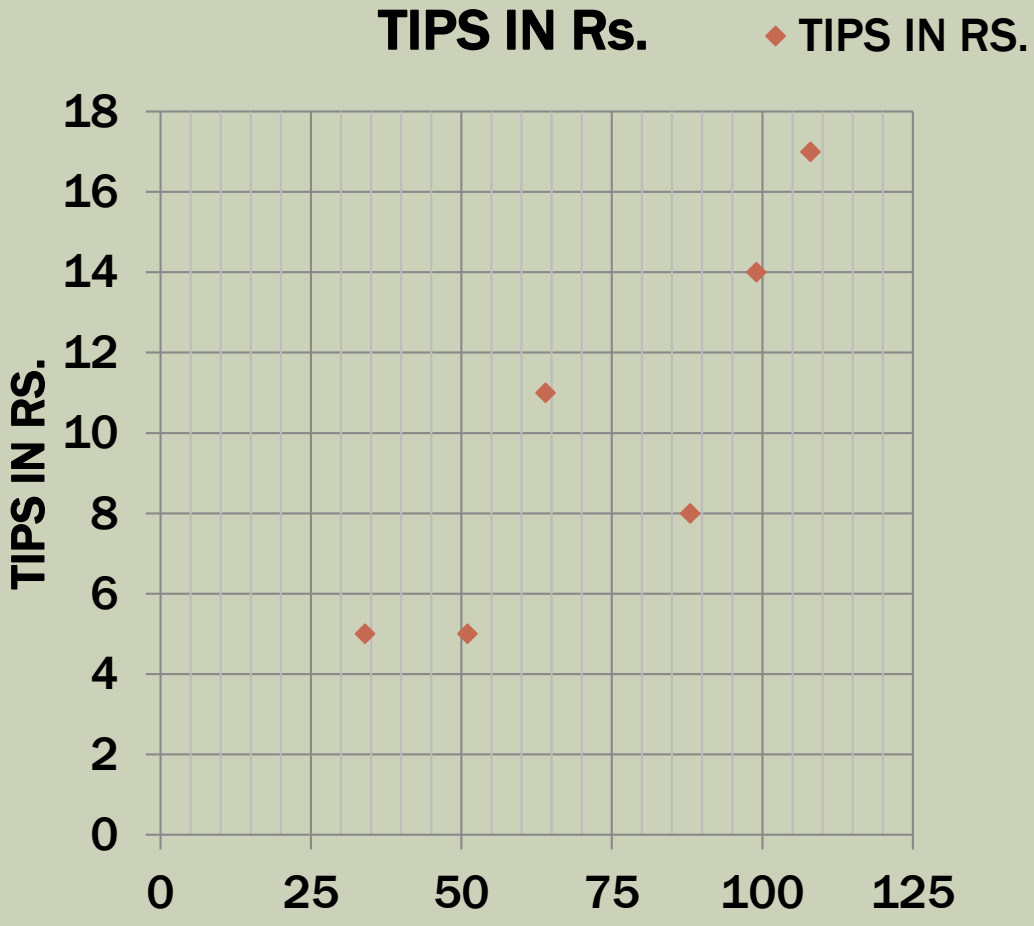
$\hat{y} = b_0 + (0)x$

$\hat{y} = b_0$

$b_0 = 10$

$\hat{y} = 10$

GETTING READY FOR LEAST SQUARES



Meal (#)	Bill Amount (Rs)	TIP Amount (Rs.)
1	34.00	5.00
2	108.00	17.00
3	64.00	11.00
4	88.00	8.00
5	99.00	14.00
6	51.00	5.00

We want to know what degree the tip amount can be predicted by the Bill.
TIP is Dependent variable
Bill is Independent variable

LEAST SQUARE CRITERIA

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

y_i = observed value of dependent variable (tip amount)

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

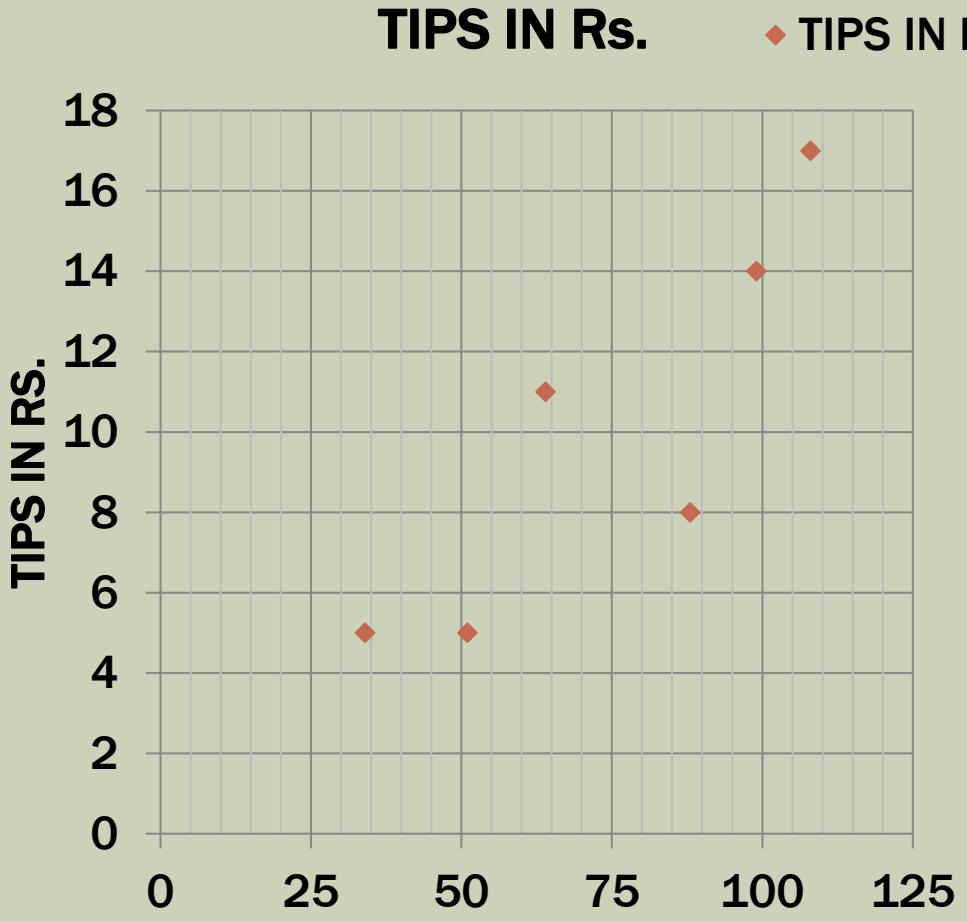
Plain English. The goal is to minimize the sum of the squared differences between the observed value for the dependent variable (y_i) and the estimated/predicted value of the dependent variable $(\hat{y}_i)$ that is provided by the regression line. Sum of the squared residuals.

Not only that, but the sum of the squared residuals should be much smaller than when we just used the dependent variable alone; $\beta_1 = 0, \hat{y} = 10$ for all values of x. That sum of squared residuals was 120.

STEPS TO ESTABLISH A REGRESSION

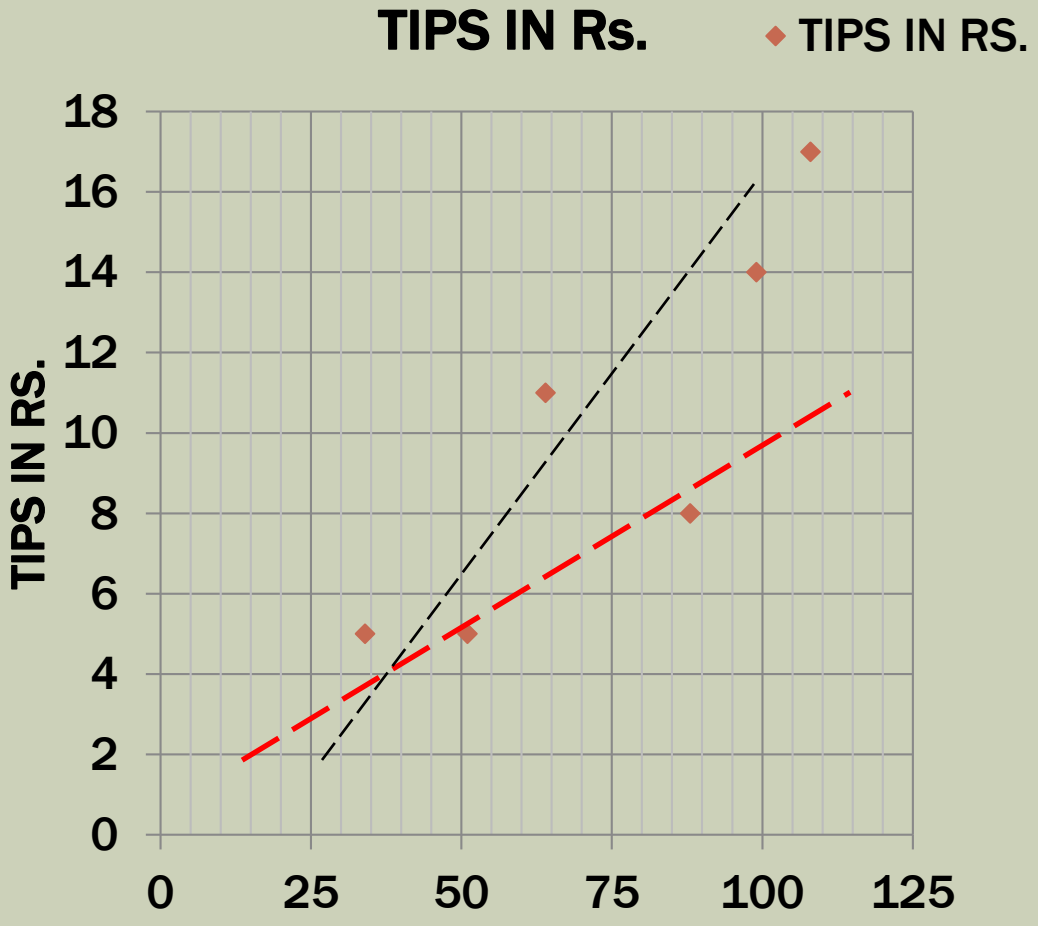
- STEP I : DRAW SCATTER PLOT
- STEP II : LOOK FOR A VISUAL LINE
- STEP III : CORREALATION(OPTINAL)
- STEP IV : Descriptive STATISTICS /Centroid
- STEP V : CALCULATIONS
  - B1 calculations
  - B0 calculation

STEP I : DRAW SCATTER PLOT

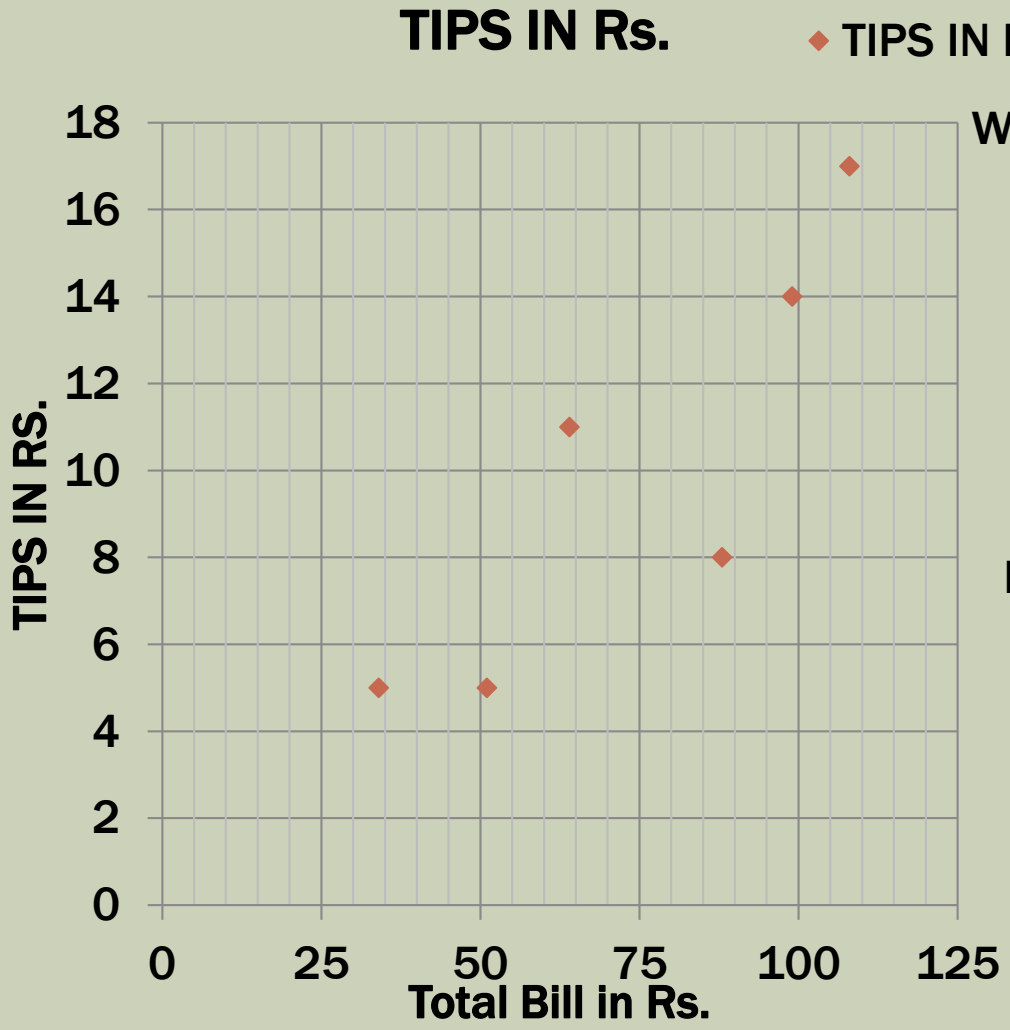


Meal (#)	Bill Amount (Rs)	TIP Amount (Rs.)
1	34.00	5.00
2	108.00	17.00
3	64.00	11.00
4	88.00	8.00
5	99.00	14.00
6	51.00	5.00

STEP II : LOOK FOR A VISUAL LINE



STEP III : CORREALATION(OPTINAL)



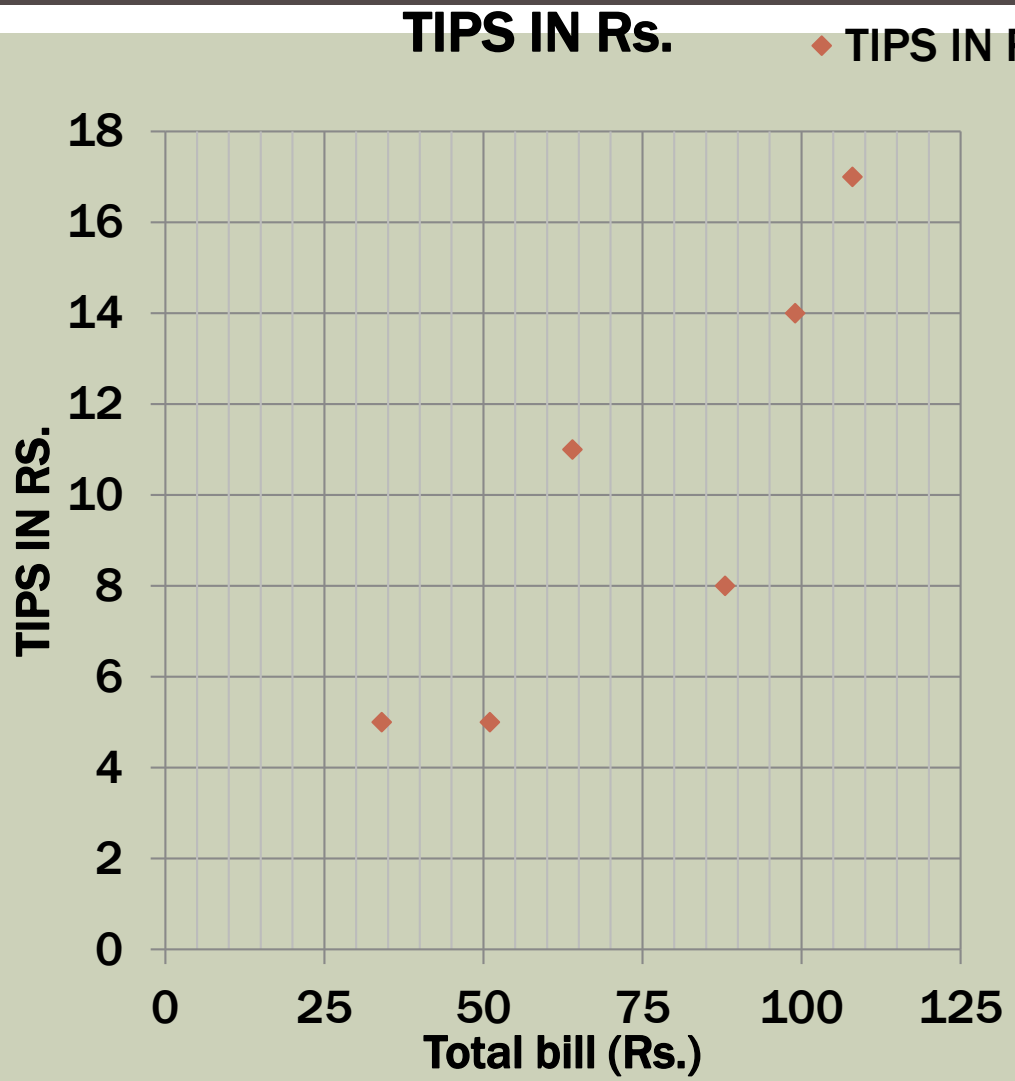
What is correlation coefficient r ?

In our example r = 0.866

Is relationship strong ?

YES in our case ! !!

STEP IV : DESCRIPTIVE STATISTICS /CENTROID



Meal (#)	Bill Amount (Rs)	TIP Amount (Rs.)
1	34.00	5.00
2	108.00	17.00
3	64.00	11.00
4	88.00	8.00
5	99.00	14.00
6	51.00	5.00
	X bar = 74	Y bar = 10

STEP V : CALCULATIONS

Intercept

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

Slope

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

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

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

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

x_i = value of independent variable

y_i = value of dependent variable

CALCULATION

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

1. For each data point.
2. Take the x-value and subtract the mean of x.
3. Take the y-value and subtract the mean of y.
4. Multiply Step 2 and Step 3
5. Add up all of the products.

---

1. For each data point.
2. Take the x-value and subtract the mean of x.
3. Square Step 2
4. Add up all the products.

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

CALCULATIONS

MEAL	Total Bill	Tip Amt.	Bill Deviation	Tip Deviation	Deviation Products	Bill Deviation 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
	$\bar{x} = 74$	$\bar{y} = 10$			$\sum = 615$	$\sum = 4206$

B1 CALCULATIONS

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

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

$$b_1 = 0.1462$$

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
$\sum = 615$	$\sum = 4206$

B0 CALCULATION

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

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

$$b_0 = 10 - 10.8188$$

$$b_0 = -0.8188$$

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

YOUR REGRESSION LINE

$\hat{y}_i = b_0 + b_1 x_i$ $\quad b_0 = -0.8188$ intercept $\quad b_1 = 0.1462$ slope

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

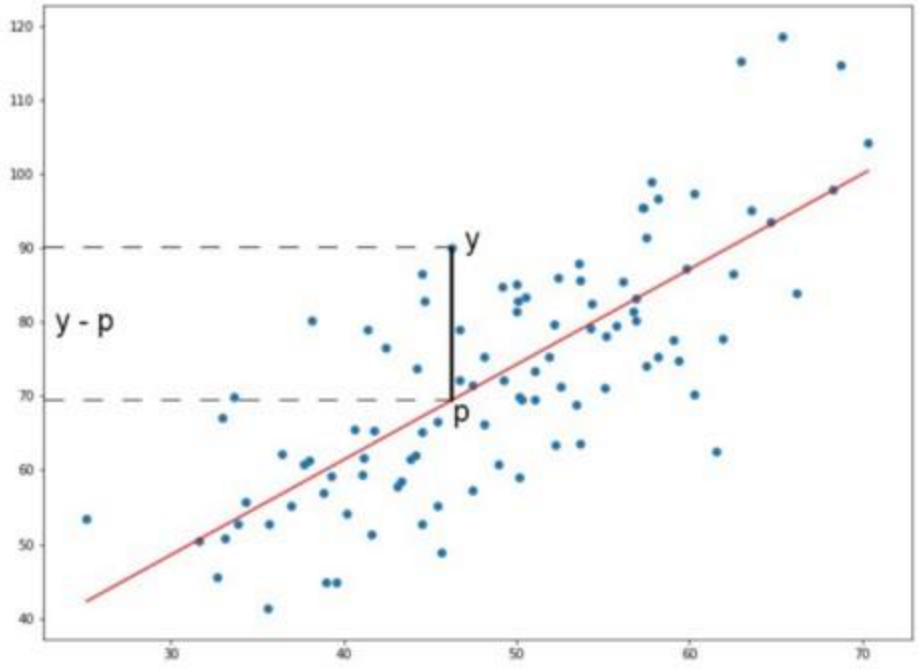
OR

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

STEPS TO ESTABLISH A REGRESSION

1. Carry out the experiment : E.g. gathering a sample of observed values of height and corresponding weight.
2. Create a relationship model.
3. Find the coefficients from the model created and create the mathematical equation .
4. Get a summary of the relationship model to know the average error in prediction.
5. Predict the weight of new persons.

$$Y = mX + c$$



Finding the Error :

$$L(x) = \sum_{i=1}^{n} (y_i - p_i)^2$$

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

$$c = \bar{y} - m\bar{x}$$

$\bar{x}$ is the mean of xi

$\bar{y}$ is the mean of all yi

- Sample data representing the observations:

Height	Weight
151	63
174	81
138	56
186	91
128	47
136	57
179	76
163	72
152	62
131	48

- Basic regression model is with only one predictor variable and the regression function is linear.
- The model can be stated as follows:

$$Y_i = \beta_0 + \beta_1 X_i$$

- where:
  - Y_i is the value of the response variable in the i^{th} trial
  - β_0 and β_1 are parameters
  - X_i is a known constant, namely, the value of the predictor variable in the i^{th} trial

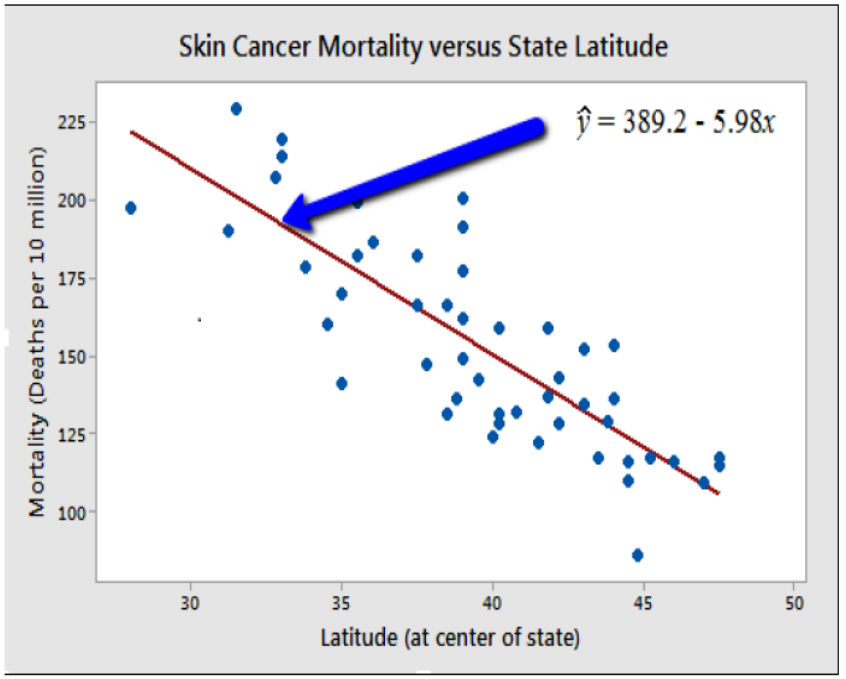
- To find “good” estimators of the regression parameters β_0 and β_1, we employ the method of least squares.
- For the observations (X_i , Y_i) for each case, the method of least squares considers the deviation of Y_i from its expected value:

$$Y_i - (\beta_0 + \beta_1 X_i)$$

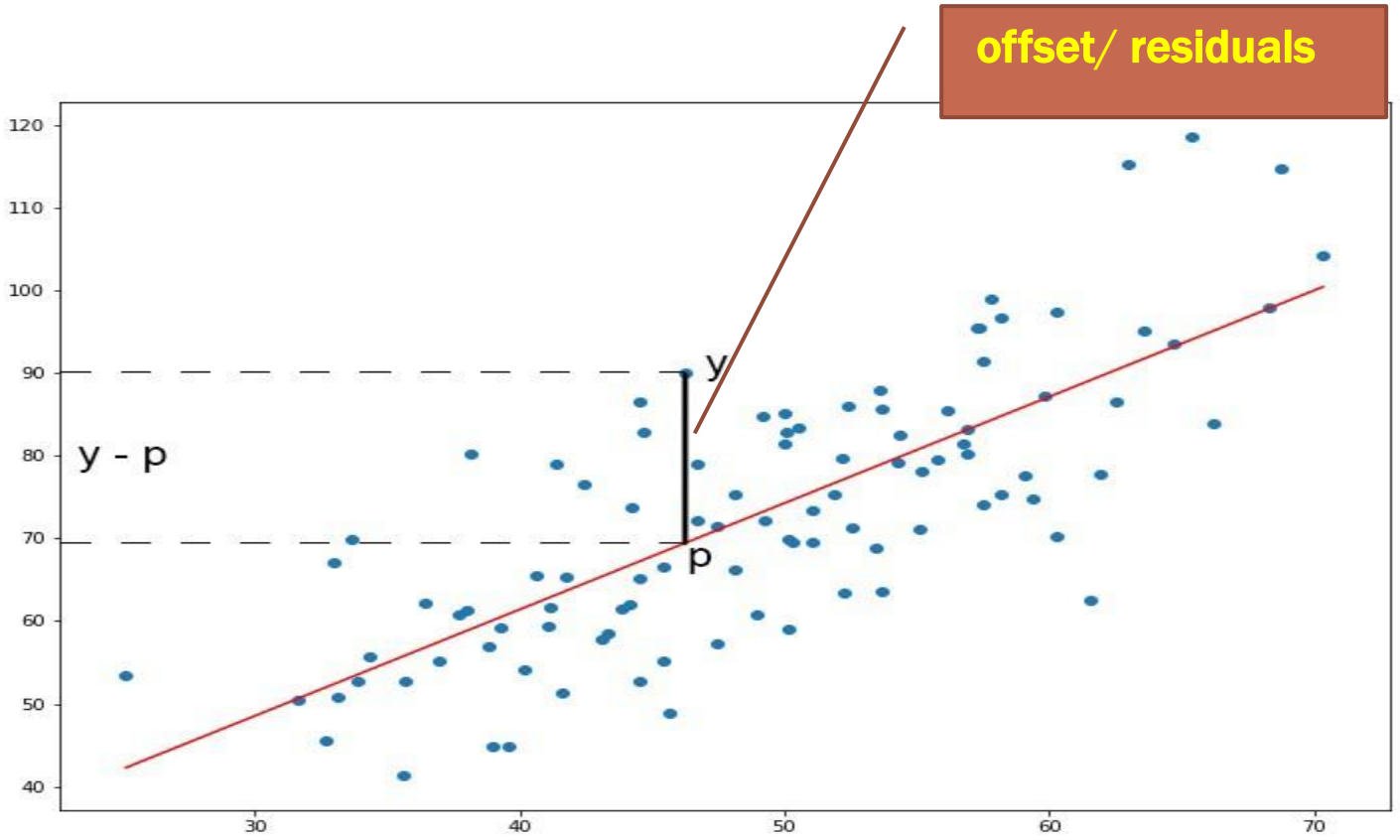
CREATING RELATIONSHIP MODEL

- The method of **least squares requires** that we consider the sum of the n squared deviations.
- This criterion is denoted by Q:

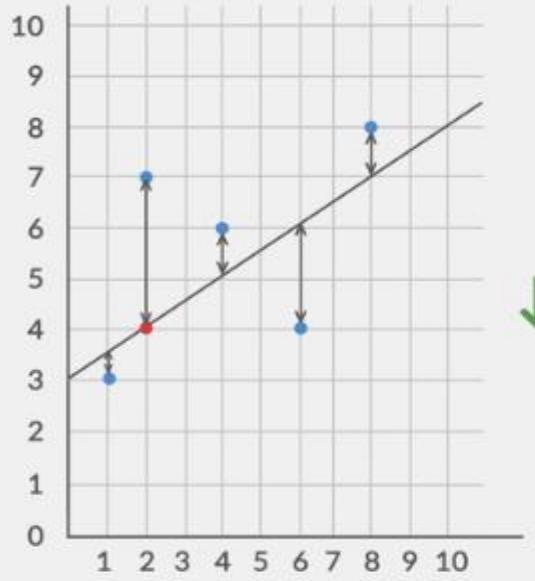
$$Q = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2$$



- What does the least squares method do exactly?
- The least-squares method is a statistical procedure **to find the best fit for a set of data points by minimizing the sum of the offsets or residuals of points from the plotted curve**. Least squares regression is used to predict the behavior of dependent variables.



- What does the least squares method do exactly?
- The least-squares method is a statistical procedure **to find the best fit for a set of data points by minimizing the sum of the offsets or residuals of points from the plotted curve**. Least squares regression is used to predict the behavior of dependent variables.



$$Y = \beta_0 + \beta_1 X$$

Intercept Slope

$$e_i = y_i - y_{pred}$$

Ordinary Least Squares Method:

$$e_1^2 + e_2^2 + \_\_\_ + e_n^2 = \text{RSS (Residual Sum Of Squares)}$$

$$RSS = (Y_1-\beta_0-\beta_1 X_1)^2 + (Y_2-\beta_0-\beta_1 X_2)^2 + \_\_\_\_ + (Y_n-\beta_0-\beta_1 X_n)^2$$

$$RSS = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2$$

WHAT IS ROLE OF INTERCEPT IN REGRESSION ANALYSIS?

- In regression analysis, the intercept is a constant term that represents the expected mean response when all predictors are equal to zero. It is an important component of a regression equation and it helps to determine the position of the regression line on the y-axis.
- In a simple linear regression model, the regression equation has the form:
- $y = \beta 0 + \beta 1x1$
- where y is the response variable, x1 is the predictor variable, β0 is the intercept, and β1 is the slope of the regression line.

- The intercept term has several important roles in regression analysis:
- **Setting the baseline**: The intercept sets the baseline or the expected value of the response when the predictor is equal to zero.
- **Adjusting for bias**: The intercept helps to adjust for any systematic bias in the data and ensures that the regression line passes through the mean of the data.
- **Interpretation of the regression coefficients**: The intercept and the slope coefficients have different interpretations when the predictors are transformed or when multiple predictors are used in a regression model.
- In summary, the intercept plays a crucial role in regression analysis as it helps to set the baseline, adjust for bias, and interpret the regression coefficients.
- .

WHAT IS ROLE OF SLOP REGRESSION ANALYSIS?

- In regression analysis, the slope (also known as the regression coefficient) is a measure of the relationship between the predictor variable and the response variable. It represents the change in the expected value of the response for a unit change in the predictor. The slope of the regression line indicates the direction and strength of the relationship between the two variables.

WHAT IS ROLE OF SLOP REGRESSION ANALYSIS?

- The slope term has several important roles in regression analysis:
  - **Measuring the strength of the relationship**: The slope indicates the strength of the relationship between the predictor and the response. A positive slope indicates a positive relationship, while a negative slope indicates a negative relationship. The magnitude of the slope indicates the strength of the relationship.
  - **Estimating the expected change**: The slope can be used to estimate the expected change in the response for a unit change in the predictor. For example, if the slope is 0.5, then we would expect the response to increase by 0.5 units for every unit increase in the predictor.
  - **Making predictions**: The slope and the intercept can be used to make predictions about the response for a given predictor value. For example, if we have the values of the slope and the intercept, we can predict the value of the response for a given value of the predictor.
- In summary, the slope is an important component of regression analysis as it measures the strength of the relationship between the predictor and the response, estimates the expected change, and helps to make predictions.

METHOD OF LEAST SQUARES

- According to the **method of least squares**, the estimators of β_0 and β_1 are those values b_0 and b_1, respectively, **that minimize the criterion Q for the given sample observations** (X_1, Y_1), (X_2, Y_2), . . . , (X_n, Y_n).
- Objective of the method : to find estimates b_0 and b_1 for β_0 and β_1, respectively, for which Q is a minimum.
- These estimates will provide a “good” fit of the linear regression function.

WHY METHOD NAME IS LEAST SQUARE FOR REGRESSION ANALYSIS?

- he method used in regression analysis is called "least squares" because it seeks to minimize the sum of the squared differences between the observed and predicted values of the response variable.
- In regression analysis, the goal is to fit a line (or a more complex model) to the data such that it best predicts the response variable.
- The residuals, which are the differences between the observed and predicted values of the response, are used to measure the goodness of fit of the model.
- The least squares method seeks to find the line (or model) that minimizes the sum of the squared residuals.

LEAST SQUARES ESTIMATORS

- The estimators b_0 and b_1 that satisfy the least squares criterion can be found in two basic ways:
  1. **Numerical search procedures** :Find Q for different estimates b_0 and b_1 until the minimum is found.
  2. **Analytical procedures**: find the values of b_0 and b_1 that minimize Q.

LEAST SQUARES ESTIMATORS

- Using the analytical approach, the values b_0 and b_1 that minimize Q for any particular set of sample data are given by the following equations:

$$\sum Y_i = nb_0 + b_1 \sum X_i$$

$$\sum X_i Y_i = b_0 \sum X_i + b_1 \sum X_i^2$$

DERIVATION

- Taking derivative

$$Q = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2$$

- Equating to zero

$$\frac{\partial Q}{\partial \beta_0} = -2 \sum (Y_i - \beta_0 - \beta_1 X_i)$$

$$\frac{\partial Q}{\partial \beta_1} = -2 \sum X_i (Y_i - \beta_0 - \beta_1 X_i)$$

$$-2 \sum (Y_i - b_0 - b_1 X_i) = 0$$

$$-2 \sum X_i (Y_i - b_0 - b_1 X_i) = 0$$

$$\sum_{i=1}^{n}(Y_i - b_0 - b_1 X_i) = 0$$

$$\sum_{i=1}^{n} X_i(Y_i - b_0 - b_1 X_i) = 0$$

Expanding, we have:

$$\sum Y_i - nb_0 - b_1 \sum X_i = 0$$

$$\sum X_i Y_i - b_0 \sum X_i - b_1 \sum X_i^2 = 0$$

Rearranging we get

$$\sum Y_i = nb_0 + b_1 \sum X_i$$

$$\sum X_i Y_i = b_0 \sum X_i + b_1 \sum X_i^2$$

LEAST SQUARES ESTIMATORS

- The normal equations can be solved simultaneously for b_0 and b_1:

$$b_1 = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2}$$

$$b_0 = \frac{1}{n}\left(\sum Y_i - b_1 \sum X_i\right) = \bar{Y} - b_1 \bar{X}$$

- where $\overline{X}$ and $\overline{Y}$ are the means of the X_i and the Y_i observations, respectively.

- Create the relationship model for the data given above for finding the relation of height on weight of students.

Sr.No.	Height(x)	Weight(y)				
1	151	63				
2	174	81				
3	138	56				
4	186	91				
5	128	47				
6	136	57				
7	179	76				
8	163	72				
9	152	62				
10	131	48				
	$\bar{x} =$	$\bar{y} =$				

Sr.No.	Height(x)	Weight(y)				
1	151	63				
2	174	81				
3	138	56				
4	186	91				
5	128	47				
6	136	57				
7	179	76				
8	163	72				
9	152	62				
10	131	48				
	$\bar{x}$ = 153.8	$\bar{y}$ = 65.3				

Sr.No.	Height(x)	Weight(y)	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$
1	151	63				
2	174	81				
3	138	56				
4	186	91				
5	128	47				
6	136	57				
7	179	76				
8	163	72				
9	152	62				
10	131	48				
	$\bar{x}$ = 153.8	$\bar{y}$ = 65.3				

Sr.No.	Height(x)	Weight(y)	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$
1	151	63	-2.8			
2	174	81	20.2			
3	138	56	-15.8			
4	186	91	32.2			
5	128	47	-25.8			
6	136	57	-17.8			
7	179	76	25.2			
8	163	72	9.2			
9	152	62	-1.8			
10	131	48	-22.8			
	$\bar{x}$ = 153.8	$\bar{y}$ = 65.3				

Sr.No.	Height(x)	Weight(y)	$x_i - \bar{x}$	$y_i - \bar{y}$		
1	151	63	-2.8	-2.3		
2	174	81	20.2	15.7		
3	138	56	-15.8	-9.3		
4	186	91	32.2	25.7		
5	128	47	-25.8	-18.3		
6	136	57	-17.8	-8.3		
7	179	76	25.2	10.7		
8	163	72	9.2	6.7		
9	152	62	-1.8	-3.3		
10	131	48	-22.8	-17.3		
	$\bar{x} =$ 153.8	$\bar{y} =$ 65.3				

Sr.No.	Height(x)	Weight(y)	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$	
1	151	63	-2.8	-2.3	6.44	
2	174	81	20.2	15.7	317.14	
3	138	56	-15.8	-9.3	146.94	
4	186	91	32.2	25.7	827.54	
5	128	47	-25.8	-18.3	472.14	
6	136	57	-17.8	-8.3	147.74	
7	179	76	25.2	10.7	269.64	
8	163	72	9.2	6.7	61.64	
9	152	62	-1.8	-3.3	5.94	
10	131	48	-22.8	-17.3	394.44	
	$\bar{x} =$ 153.8	$\bar{y} =$ 65.3			$\Sigma =$ 2649.6	

Sr.No.	Height(x)	Weight(y)	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$
1	151	63	-2.8	-2.3	6.44	7.84
2	174	81	20.2	15.7	317.14	408.04
3	138	56	-15.8	-9.3	146.94	249.64
4	186	91	32.2	25.7	827.54	1036.8
5	128	47	-25.8	-18.3	472.14	665.64
6	136	57	-17.8	-8.3	147.74	316.84
7	179	76	25.2	10.7	269.64	635.04
8	163	72	9.2	6.7	61.64	84.64
9	152	62	-1.8	-3.3	5.94	3.24
10	131	48	-22.8	-17.3	394.44	519.84
	$\bar{x} =$ 153.8	$\bar{y} =$ 65.3			$\Sigma =$ 2649.6	$\Sigma =$ 3927.6

ESTIMATE

$$b_1 = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2}$$

$$b_0 = \frac{1}{n}\left(\sum Y_i - b_1 \sum X_i\right) = \bar{Y} - b_1 \bar{X}$$

ESTIMATE

$$b_1 = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2}$$

$$b_0 = \frac{1}{n}\left(\sum Y_i - b_1 \sum X_i\right) = \bar{Y} - b_1 \bar{X}$$

$b_1 = 2649.6/3927.6 = 0.6746$

ESTIMATE

$$b_1 = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2}$$

$$b_0 = \frac{1}{n}\left(\sum Y_i - b_1 \sum X_i\right) = \bar{Y} - b_1 \bar{X}$$

b_1 = 2649.6/3927.6 = 0.6746

b_0 =65.3-(0.6746 x 153.8) = -38.4551

RELATIONSHIP MODEL

Y_i (Weight) = 0.6746 X_i (Height) - 38.4551

EXAMPLE OF REGRESSION

- Consider a system to predict weight of a person when his height is known.
- To do this we need to have the relationship between height and weight of a person.

RELATIONSHIP MODEL

Y_i (Weight) = 0.6746 X_i (Height) - 38.4551

MODEL PERFORMANCE

- Once you build the model, we must know whether your model is good enough to predict
- For this purpose there are various metrics which we use.

KARL PEARSON'S CORRELATION COEFFICIENT

$$r = \frac{N\sum xy - \sum x \sum y}{\sqrt{[N\sum x^2 - (\sum x)^2][N\sum y^2 - (\sum y)^2]}}$$

- It has a value between +1 and −1
  - 1 is total positive linear correlation
  - 0 is no linear correlation, and
  - −1 is total negative linear correlation.

PROBLEM

- Compute the correlation coefficient for the example for the data given above for finding the relation of height on weight of students.

Sr.No.	Height(x)	Weight(y)	xy	x^2	y^2
1	151	63			
2	174	81			
3	138	56			
4	186	91			
5	128	47			
6	136	57			
7	179	76			
8	163	72			
9	152	62			
10	131	48			
	Σ=1538	**Σ=653**			
	$\bar{x}$ = 153.8	$\bar{y}$ =65.3			

Sr.No.	Height(x)	Weight(y)	xy	x^2	y^2
1	151	63	9513		
2	174	81	14094		
3	138	56	7728		
4	186	91	16926		
5	128	47	6016		
6	136	57	7752		
7	179	76	13604		
8	163	72	11736		
9	152	62	9424		
10	131	48	6288		
	Σ=1538	**Σ=653**	Σ		
	$\bar{x}$ = 153.8	$\bar{y}$ =65.3	103081		

Sr.No.	Height(x)	Weight(y)	xy	x^2	
1	151	63	9513	22801	
2	174	81	14094	30276	
3	138	56	7728	19044	
4	186	91	16926	34596	
5	128	47	6016	16384	
6	136	57	7752	18496	
7	179	76	13604	32041	
8	163	72	11736	26569	
9	152	62	9424	23104	
10	131	48	6288	17161	
	Σ=1538	Σ=653	Σ	Σ	
	$\bar{x}$ = 153.8	$\bar{y}$ =65.3	103081	240472	

Sr.No.	Height(x)	Weight(y)	xy	x^2	y^2
1	151	63	9513	22801	3969
2	174	81	14094	30276	6561
3	138	56	7728	19044	3136
4	186	91	16926	34596	8281
5	128	47	6016	16384	2209
6	136	57	7752	18496	3249
7	179	76	13604	32041	5776
8	163	72	11736	26569	5184
9	152	62	9424	23104	3844
10	131	48	6288	17161	2304
	Σ=1538	Σ=653	Σ	Σ	Σ
	$\bar{x}$ = 153.8	$\bar{y}$ =65.3	103081	240472	44513

$$r = \frac{N \sum xy - \sum x \sum y}{\sqrt{[N \sum x^2 - (\sum x)^2][N \sum y^2 - (\sum y)^2]}}$$

$$r = \frac{N\sum xy - \sum x \sum y}{\sqrt{[N\sum x^2 - (\sum x)^2][N\sum y^2 - (\sum y)^2]}}$$

$$r = \frac{(10 \times 103081) - (1538 \times 653)}{\sqrt{[(10 \times 240272) - 1538^2][(10 \times 44513) - 653^2]}}$$

$$r = \frac{N\sum xy - \sum x \sum y}{\sqrt{[N\sum x^2 - (\sum x)^2][N\sum y^2 - (\sum y)^2]}}$$

$$r = \frac{(10 \times 103081) - (1538 \times 653)}{\sqrt{[(10 \times 240272) - 1538^2][(10 \times 44513) - 653^2]}}$$

$$r = 0.9771$$

- For each observation in our data we can compute

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

- These are called fitted values.
- For the ith observation, we can compute **ordinary least squares residuals** as

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

- One of the properties of the residuals is that their sum is zero.

- Compute the following quantities:

$$\text{SST} = \sum (y_i - \bar{y})^2$$

$$\text{SSR} = \sum (\hat{y}_i - \bar{y})^2$$

$$\text{SSE} = \sum (y_i - \hat{y}_i)^2.$$

- SST : viations in Y from its mean
- SSR : sum of squares due to regression
- SSE : sum of squared residuals (errors)

- The ratio
- $R^2 = SSR/SST$
- can be interpreted as the proportion of the total variation in Y by X.
- The high value of R^2 indicates a strong linear relationship.

Height	Weight	ỹ	SSR	SST	SSE
151	63				
174	81				
138	56				
186	91				
128	47				
136	57				
179	76				
163	72				
152	62				
131	48				
Mean of y:	65.3				

Height	Weight	ỹ	SSR	SST	SSE
151	63	63.411	3.56828322		
174	81	78.927	185.696219		
138	56	54.641	113.612576		
186	91	87.022	471.860924		
128	47	47.895	302.934721		
136	57	53.292	144.195426		
179	76	82.3	289.00306		
163	72	71.506	38.5185321		
152	62	64.086	1.47471878		
131	48	49.919	236.581006		
Mean of y:	65.3		1787.44547		

Height	Weight	ỹ	SSR	SST	SSE
151	63	63.411	3.56828322	5.29	
174	81	78.927	185.696219	246.49	
138	56	54.641	113.612576	86.49	
186	91	87.022	471.860924	660.49	
128	47	47.895	302.934721	334.89	
136	57	53.292	144.195426	68.89	
179	76	82.3	289.00306	114.49	
163	72	71.506	38.5185321	44.89	
152	62	64.086	1.47471878	10.89	
131	48	49.919	236.581006	299.29	
Mean of y:	65.3		1787.44547	1872.1	

Height	Weight	ỹ	SSR	SST	SSE
151	63	63.411	3.56828322	5.29	0.16892922
174	81	78.927	185.696219	246.49	4.29716316
138	56	54.641	113.612576	86.49	1.84666357
186	91	87.022	471.860924	660.49	15.82162
128	47	47.895	302.934721	334.89	0.8009892
136	57	53.292	144.195426	68.89	13.7503023
179	76	82.3	289.00306	114.49	39.691134
163	72	71.506	38.5185321	44.89	0.24371007
152	62	64.086	1.47471878	10.89	4.34981078
131	48	49.919	236.581006	299.29	3.68183182
Mean of y:	65.3		1787.44547	1872.1	84.6521541

R-SQUARE

- R^2 = SSR/SST
  = 1784.45/1872.1
  = 0.9548