



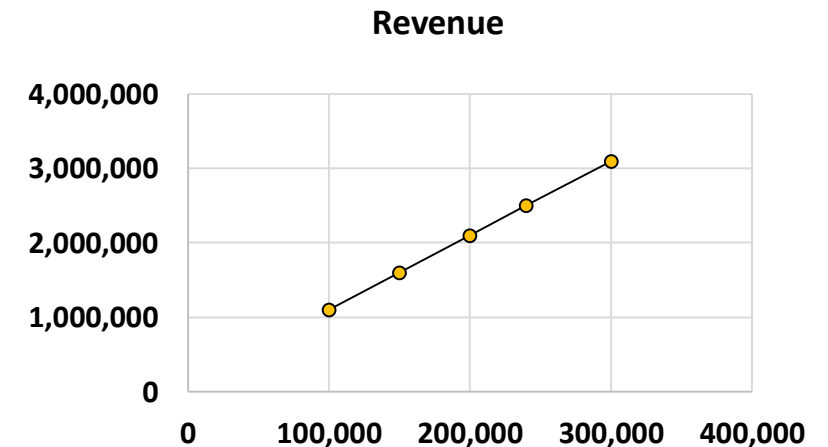
# Linear Regression

# Linear Regression

- A Production House in Mumbai wants to analyse the relationship between the promotion and sales for their daily serials. They want to know how much revenue would be generated if they spend Rs.4,00,000 on promotion of a new serial.
- In the Data set 1, we can see that the data is linearly distributed.

Data set 1

Serial	Promotion	Revenue
1	1,00,000	11,00,000
2	1,50,000	16,00,000
3	2,00,000	21,00,000
4	2,40,000	25,00,000
5	3,00,000	31,00,000

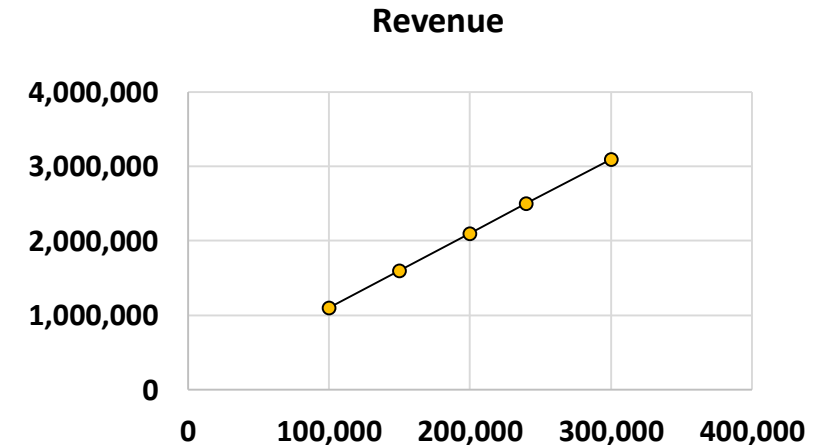


# Linear Regression

- For every 1 additional unit on promotional expenses, the revenue figure is increasing by 10. i.e. when the promotional expenses increased by 50,000, in serial no. 2, their revenue also increased by 5,00,000.
- In all the serials, you can also notice that revenue is  $10 * \text{Promotion} + 1,00,000$  i.e.  $1,00,000 * 10 + 1,00,000 = 11,00,000$  for serial 1 and  $1,50,000 * 10 + 1,00,000 = 16,00,000$  and so on. Therefore, you also have to add Rs.100,000 to every serial.
- **Therefore, if the promotion house spend Rs.4,00,000, then the revenue will be :  $4,00,000 * 10 + 1,00,000 = 41,00,000$**
- **This example is one of the simplest form of Regression**

Data set 1

Serial	Promotion	Revenue = $10 * \text{Promotion} + 100000$
Serial 1	1,00,000	1100000
Serial 2	1,50,000	1600000
Serial 3	2,00,000	2100000
Serial 4	2,40,000	2500000
Serial 5	3,00,000	3100000



# Linear Regression

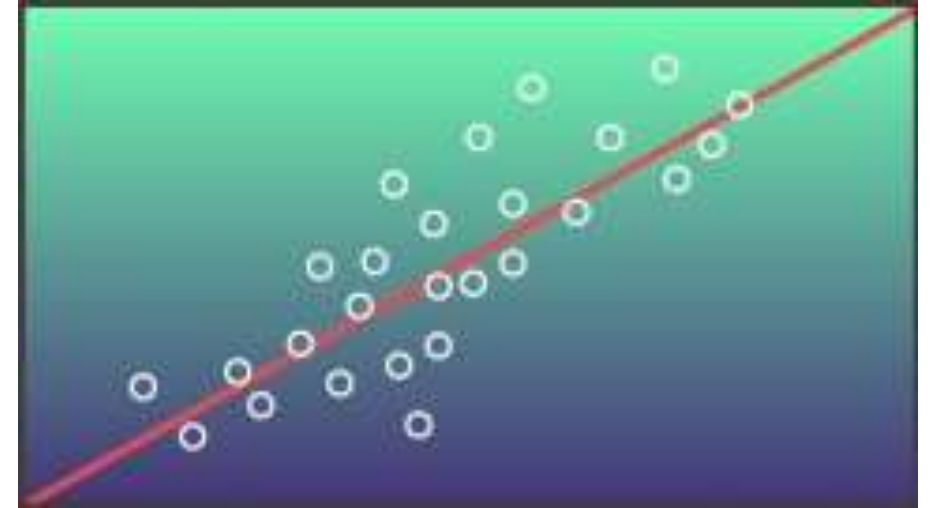
- If you get a complex data set ( Ref. Data 1 ) it's difficult to find out the relationship between the promotion and revenue just by observation.
- To know the relationship between the variables and predict the revenue , we need to understand regression in details.

Data Set 2

Serial	Promotion	Revenue
1	111600	1197576
2	104400	1053648
3	97200	1124172
4	79200	987144
5	126000	1283616
6	108000	1295100
7	147600	1407444
8	104400	922416
9	169200	1272012
10	75600	1064856
11	133200	1269960
12	133200	1064760
13	176400	1207488
14	180000	1186284
15	133200	1231464
16	147600	1296708
17	122400	1320648
18	158400	1102704
19	165600	1184316
20	104400	1326360

# What is Linear Regression?

- A statistical measure that attempts to determine the strengths of the relationship between one dependent variable (usually denoted by  $Y$ ) and a series of the other changing variable (Known as independent variable)
- In another word: Regression is an mathematical technique used to estimate the cause and effect relationship among variables



# Regression?

- For example: what happens to the revenue coming from advertisement of a Television program, if the production house decide to increase the promotional budget by Rs.1,00,000

Here:

**Cause :** Money spend on Promotion

**Effect :** An increase in Advertisement Revenue

- We know that the effect of an increase in money spent in Promotion is an increase in Advertisement Revenue - If we also want to know by how much?
- The regression technique tells us what is the impact and what is the quantification of the impact when we have cause and effect relationship

# Applications of Linear Regression



Which consumer is likely to default?



Which promotion is more effective?



What is the risk associated with a consumer?

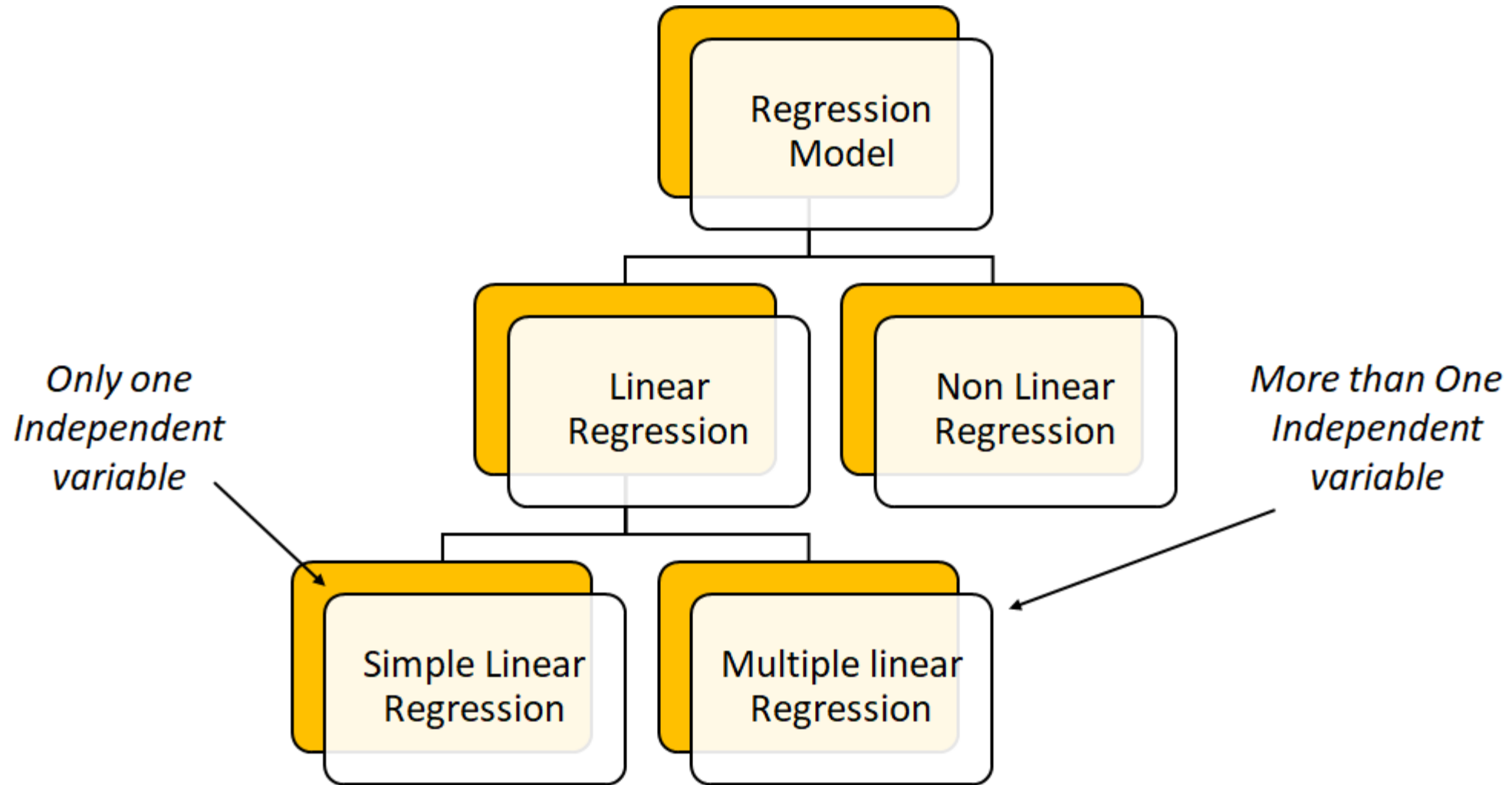


What percentage of loans is likely to result in a loss?



How to identify the most profitable customer?

# Types of Linear Regression





# Regression Analysis Forms and Types

- **Simple Linear Regression:** Single explanatory variable
  - ✓ Only one independent variable, X
  - ✓ In our Case: X = Promotion  
Y = Advertisement Revenue
  - ✓ Relationship between X and Y is described by a linear function
  - ✓ Changes in Y are assumed to be related to changes in X
- **Equation:**  $Y = a + bX$
- **Multiple Linear Regression:** Includes any number of explanatory variables
  - ✓ Two or more independent variables, X's
  - ✓ In our Case: X = Television Rating Points, Language of the program (Hindi or regional language), Money spent on Promotion.
  - ✓ Relationship between X's and Y is described by a linear function
  - ✓ Changes in Y are assumed to be related to changes in X
- **Equation:**  $Y = a + b_1 * X_1 + b_2 * X_2 + ... + b_p * X_p$
- **Non-linear:** Implies curved relationships
  - ✓ Logarithmic Relationships

- **Dependent variable:** The variable we wish to predict or explain. It can also be called Measured variable or Explained variable or Response variable.

Example : Advertising Revenue.

- **Independent variable:** The variable(s) used to predict or explain the dependent variable called as Explanatory variable(s) or Manipulated variable(s) or Controlled variable(s) or Predictor variable(s).

Example: Money spent on Promotion.

- **Coefficients:** The estimate of magnitude of impact of changes in the predictor(s) on the predicted variable.

Example: The coefficient is 10 in the equation  $(1,00,000 + 10x)$

- **Intercept:** The intercept of this line is the value of y at the point where the line crosses the y axis.

Example: The intercept is 1,00,000 in equation  $(1,00,000 + 10x)$

- **Error(e):** The impact of the unobserved variables on the dependent variable, usually calculated as the difference between the predicted value of Y given the estimated regression function and the actual value of Y

# Problem Statement

- A famous Television Production House in Mumbai wants to forecast the revenue generated during advertisement. To predict the Advertising Revenue, Television Production House has to look at money spent on Promotion.



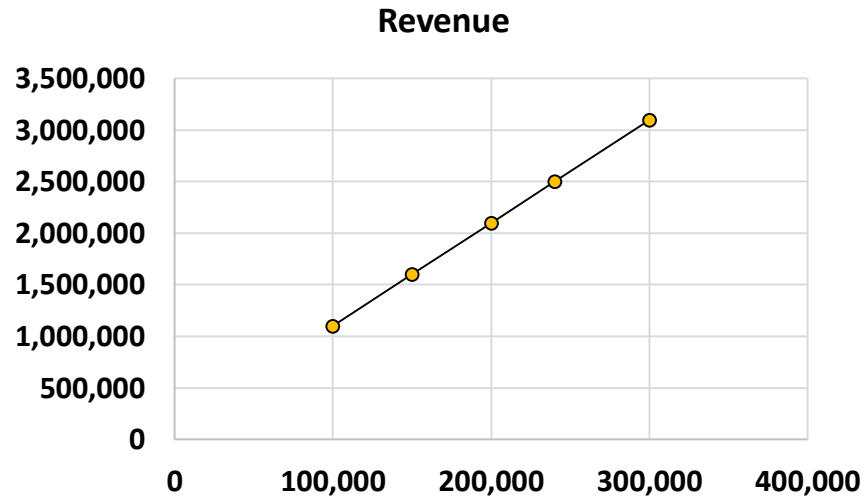
# Data set 2

Serial	Promotion	Revenue
1	111600	1197576
2	104400	1053648
3	97200	1124172
4	79200	987144
5	126000	1283616
6	108000	1295100
7	147600	1407444
8	104400	922416
9	169200	1272012
10	75600	1064856
11	133200	1269960
12	133200	1064760
13	176400	1207488
14	180000	1186284
15	133200	1231464
16	147600	1296708
17	122400	1320648
18	158400	1102704
19	165600	1184316
20	104400	1326360

- Lets start with Simple linear regression and analyze the relationship between the Advertising Revenue & Promotion
- Here, the effect is Advertising Revenue.
- The possible causes of Advertising Revenue in this dataset is Promotion.
- The possible ways of assessing these relationships are:
  - ✓ ***Graphical visualization (or Scatter Plot)***
  - ✓ ***Run regression model***
- Along with this the best way to find out the relationship is regression because it gives the statistical significance of the variable and the impact of the multiple factor on effect

# Graphical Visualization of Example 1 ( Ref: Dataset 1)

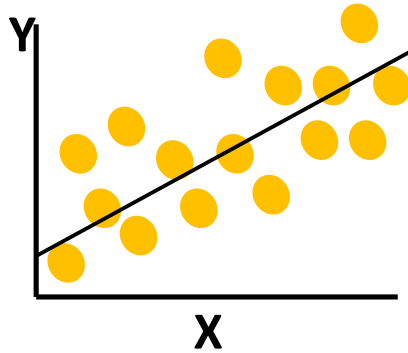
Data set 1



Serial	Promotion	Revenue
1	1,00,000	11,00,000
2	1,50,000	16,00,000
3	2,00,000	21,00,000
4	2,40,000	25,00,000
5	3,00,000	31,00,000

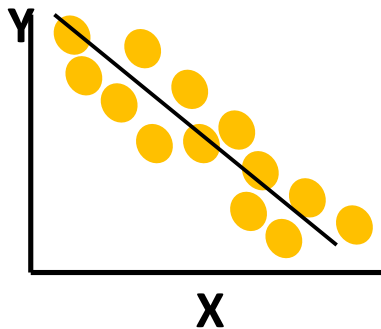
- In the above graph, it is seen that the revenue and promotion are positively linear.
- Lets discuss about the other linear relations in next slides

# Graphical Visualization Examples



**Positive  
Linear**

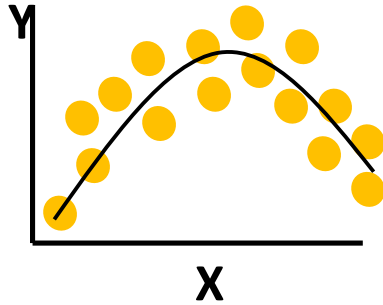
In Positive Linear Relationship, the values of Y are increasing linearly as X increases.



**Negative  
Linear**

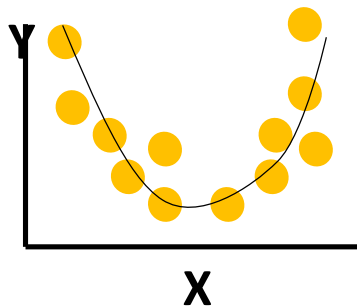
In Negative Linear Relationship as X increases, the values of Y decreases.

# Graphical Visualization Examples (Contd.)



## Positive Curvilinear

In positive curvilinear relationship between  $X$  and  $Y$  the values of  $Y$  increases as  $X$  increases, but this increase tapers off beyond certain values of  $X$



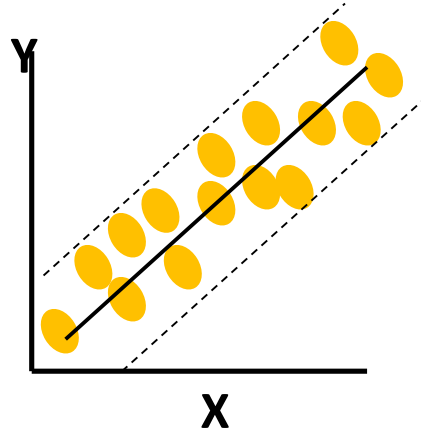
## Negative Curvilinear

In Negative Curvilinear Relationship it illustrate an exponential relationship between  $X$  and  $Y$ . In this case,  $Y$  decreases very rapidly as  $X$  first increases, but then it decreases much less rapidly as  $X$  increases further

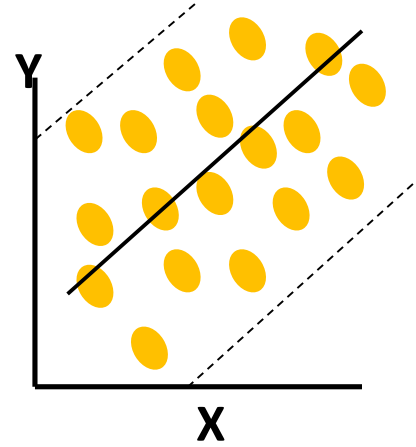


# Graphical Visualization Examples (Contd.)

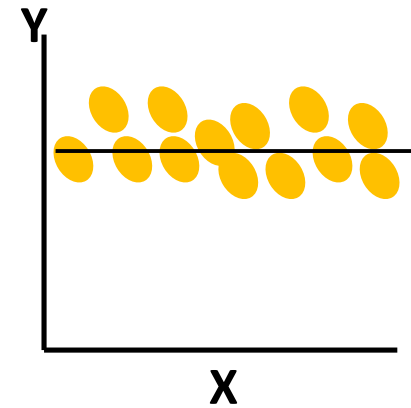
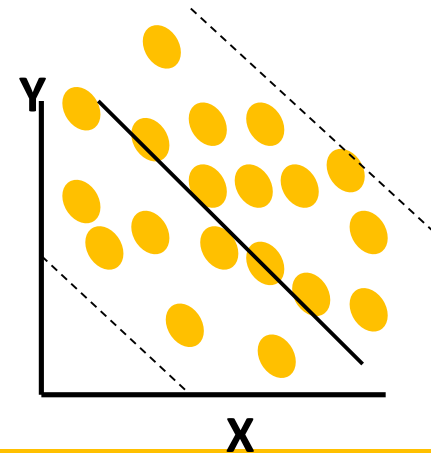
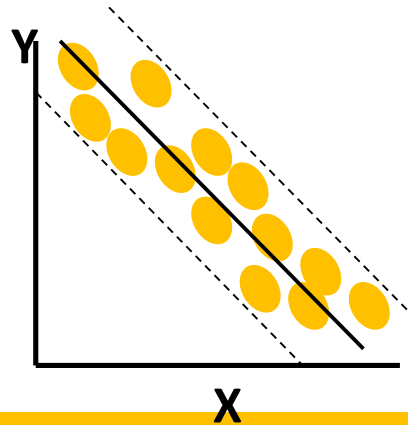
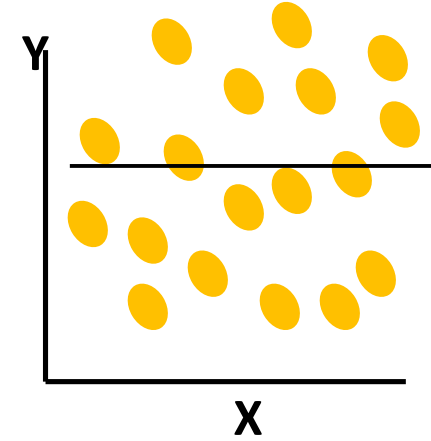
Strong  
Relationship



Weak  
Relationship



No  
Relationship



# Straight Line Relationship

- The case of regression is to find out the relationship between advertising revenue and promotion.
- Mathematically:
- $Y = f(X)$
- Where, Y is Dependent variable and X is an Independent Variable
- Advertising Revenue =  $f(\text{Promotion})$
- Where,  $f$  is the functional form
- We are currently reviewing a linear regression model-
- A linear relationship between the two variable is essentially a straight line relationship.

# Straight Line Relationship

- The mathematical equation that denotes a linear (straight line) relationship between two variables, x and y?

$$y = mx + c$$

Where,

Slope (m) = The rate of change of Y when X changes

Intercept(c) = The intercept is the value of Y when X = 0.

y = Dependent variable,

x = Independent variable

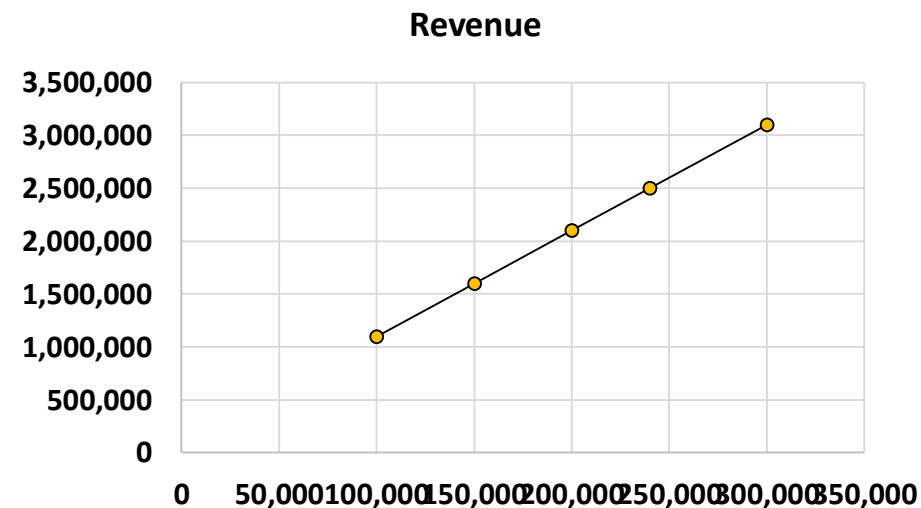
# Straight Line Relationship

- In a straight line relationship, the equation is

$$y = 10x + 1,00,000.$$

Here, Slope = 10 and Intercept = 1,00,000

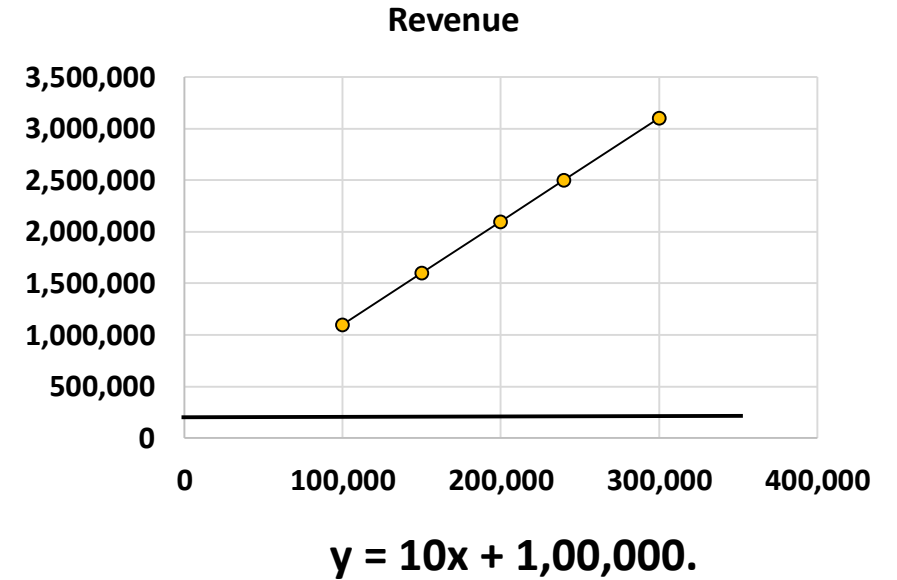
- If x is 1,00,000 then y is 11,00,000. If x is 1,50,000 then y is 16,00,000. So, when x changes from 1,00,000 to 1,50,000 the y changes from 11,00,000 to 16,00,000, so the change in y is 5,00,000 relative to a change in x and that is slope.
- In a straight line relationship, irrespective of level of x changes the rate of change of y is always constant.
- In Non linear relationship, The rate of change of Y relative to x is not constant then you have a curved line.
- In non linear case, slope will not constant due to curved line.



$$y = 10x + 1,00,000.$$

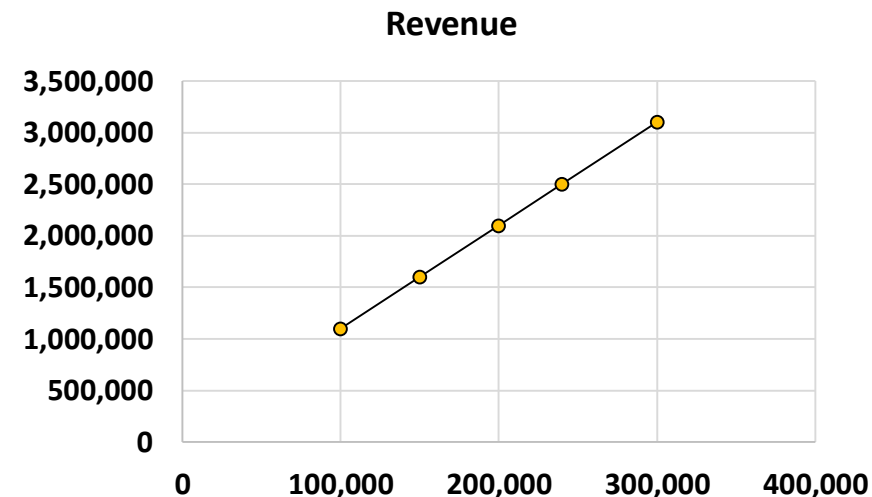
# Straight Line Relationship

- **Case 1 : Slope = 0**
  - ✓ The straight line is parallel to the x axis.
  - ✓ If slope is 0 then the value of x is 0
  - ✓ The value of y in this case is 1,00,000.
- **Case 2: Intercept = 0**
  - ✓ The straight line is passes through the centre.
  - ✓ If intercept is 0 then the value of y changes with the value of x, but it passes through the centre.



# Straight Line Relationship

- In our example of regression, the intercept and the slope of x i.e. promotion is 10.
- Intercept = 1,00,000 (Intercept is the value of y when x is 0)
- Slope = 10
- For a unit change in X, Y Changes by a constant amount (10)

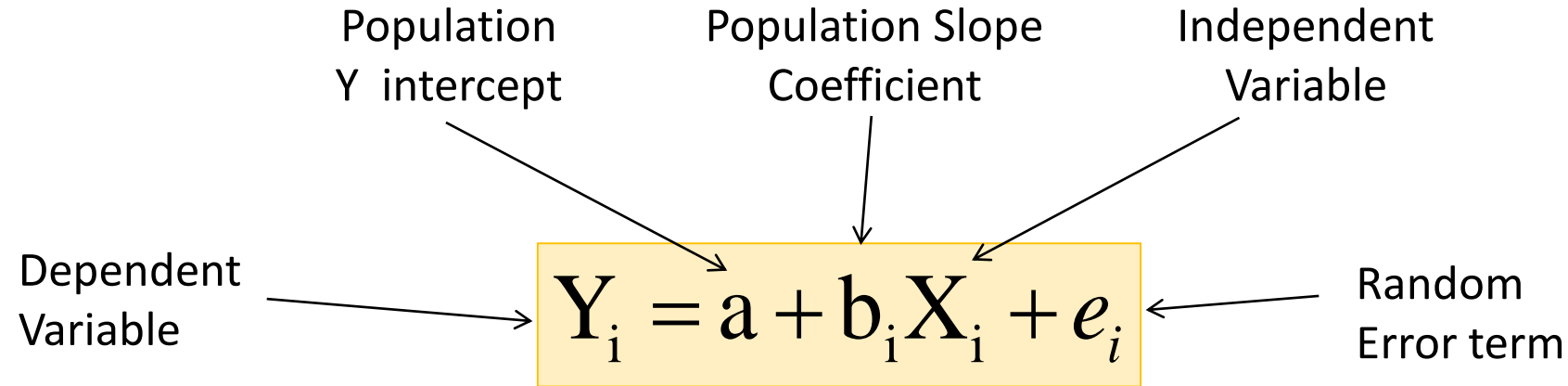


# Simple Linear Regression

- Lets continue with data set 2, Television production house have many factors to find out the revenue generated by advertisement. So production house decides to start by taking only one factor i.e. promotion to find out the advertising revenue.
- Here, the promotion and the advertising revenue are the only variables, where
  - $X = \text{Promotion}$
  - $Y = \text{Advertising Revenue}$
- So, This is the form of Simple linear regression
- Let's see how to calculate the advertising revenue based on promotion using simple linear regression

# Simple Linear Regression Model

- Linear Regression Equation:



The diagram shows the linear regression equation  $Y_i = a + b_i X_i + e_i$  enclosed in a yellow box. Arrows point from labels to the components of the equation: 'Dependent Variable' points to  $Y_i$ , 'Population Y intercept' points to  $a$ , 'Population Slope Coefficient' points to  $b_i$ , 'Independent Variable' points to  $X_i$ , and 'Random Error term' points to  $e_i$ .

$$Y_i = a + b_i X_i + e_i$$

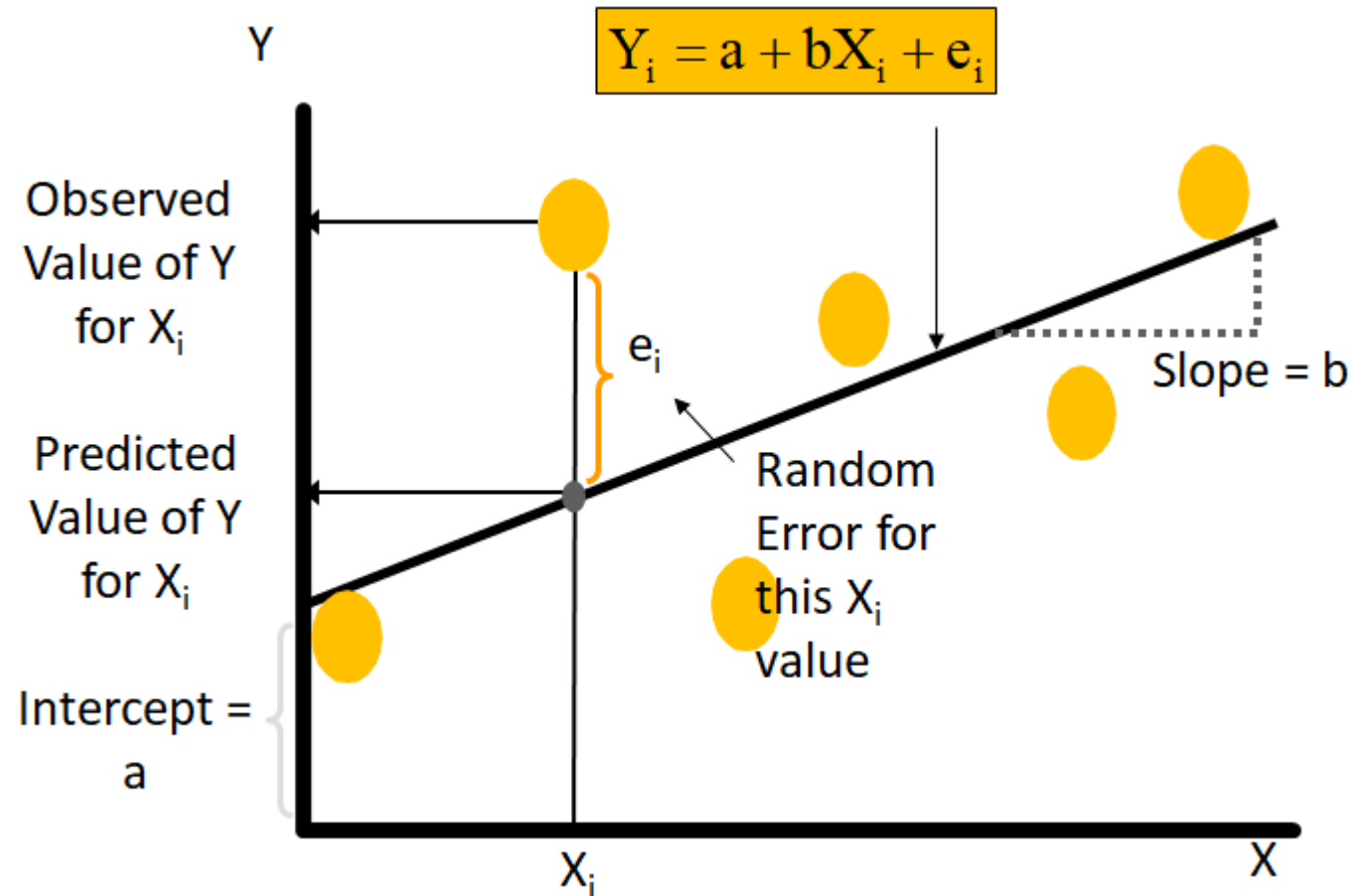
where:

- ✓  $Y$ : the variable that we are trying to predict
- ✓  $X_i$ : the variables that we are using to predict  $Y$
- ✓  $a$ : intercept
- ✓  $b_i$ : slope
- ✓  $e_i$  = the regression residual or the error term

- $a$  is the estimated **average value of  $Y$**  when the value of  $X$  is zero
- $b$  is the estimated change in the average value of  $Y$  as a result of a one-unit increase in  $X$



# Simple Linear Regression Model

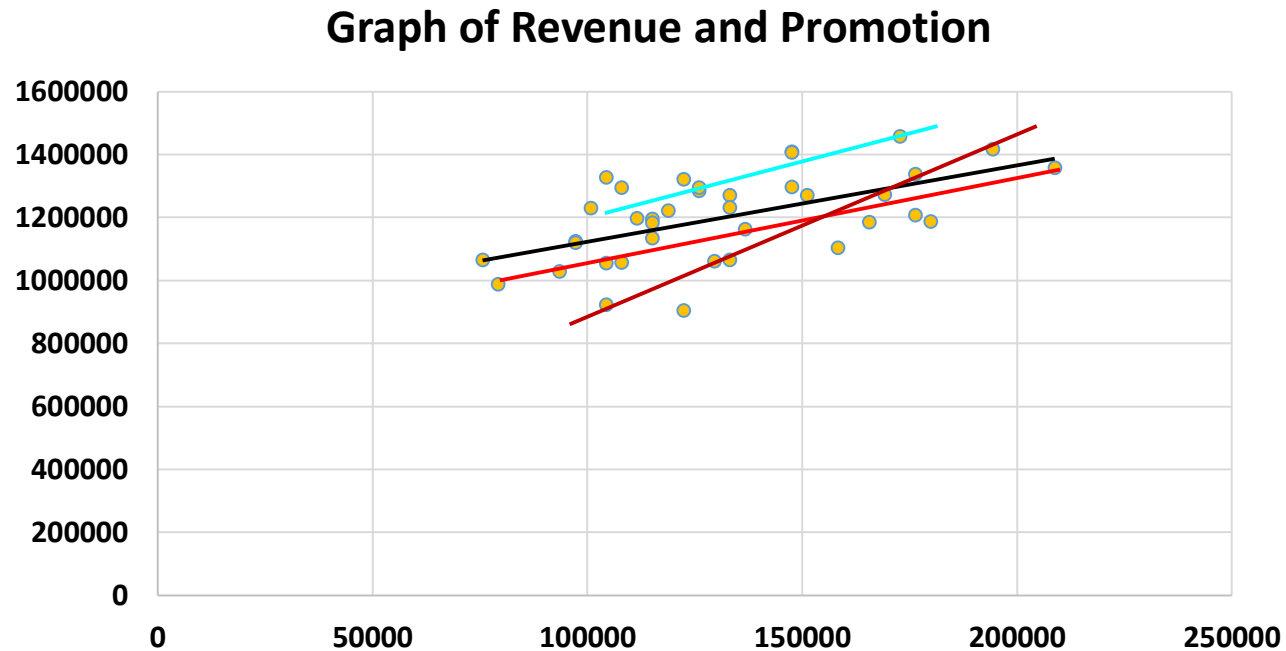


## Now lets consider a Television Production House example (data set 2)

- In order to understand the Television Production house dataset, we will take a sample of 38 observations as shown in the table.
  - Dependent variable (y): Advertising Revenue
  - Independent variable (x): Promotion
- In the Television Production House example , we believe:
- $Advertising\ Revenue = a + b_i * Promotion + e_i$
- Now we need to estimate what the coefficients values are, from the data available, that will best capture the relationship between Advertising Revenue and Promotion.

# Ordinary Least Square

- In statistics, **ordinary least squares (OLS)** is a method for estimating the unknown parameters in a linear regression model, with the goal of minimizing the sum of the squares of the differences between the observed values and predicted values which are calculated by a linear function of a set of independent variable.



# Best Fit Line

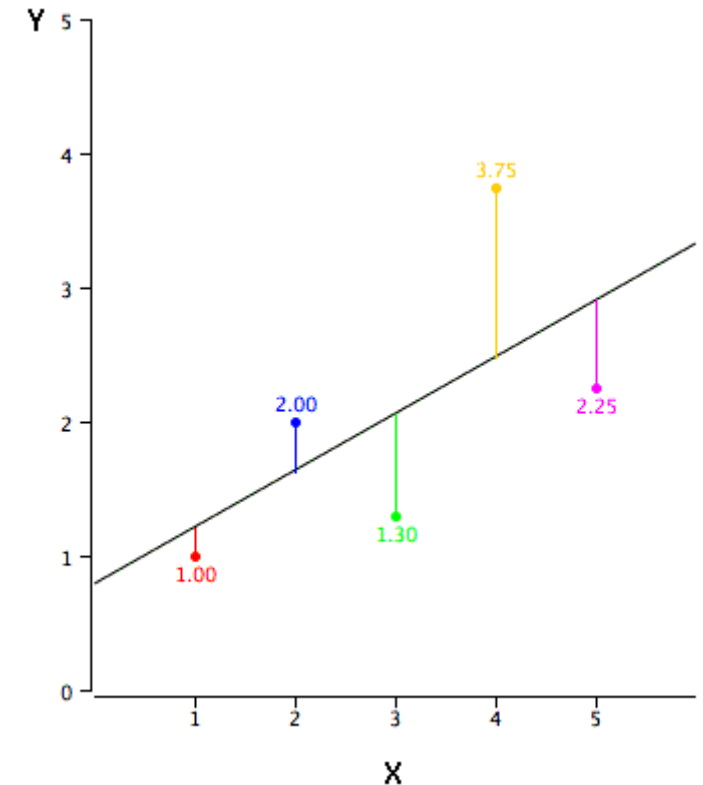
- Consider a Scatter plot in the above slide, if we assume a straight line relationship is present between the variables, then draw a straight line that best captures the relationship between Promotion and Revenue.
- From scatter plot we can clearly say that there are many lines that will cover some of the points
- To know the best fit line refer the below explanation.

## How many lines are possible to capture the relationship?

- ✓ Many lines are possible to capture the relationship.

## Which is the best possible straight line?

- ✓ The black diagonal line is the regression line and consists of the predicted score on Y for each possible value of X. The vertical lines from the points to the regression line represent the errors of prediction.
- ✓ As you can see, the red point is very near the regression line; its error of prediction is small. By contrast, the yellow point is much higher than the regression line and therefore its error of prediction is large.
- ✓ The best-fitting line is the line that minimizes the sum of the squared errors of prediction.



# Ordinary least square

- The ordinary least square regression used for estimating line that minimizes the sum of the squares of the errors
- Why Square of error?
  - ✓ It magnifies or penalizes, the larger errors.
  - ✓ It cancels the effect of the positive and negative values
- Mathematically, minimize

$$Q = \sum_{i=1}^N (Y_i - b_0 - b_1 X_i)^2$$

# Least Square Regression

- Using differential calculus, we will get

$$b_0 = \frac{\sum X_i^2 \sum Y_i - \sum X_i \sum X_i Y_i}{n \sum X_i^2 - (\sum X_i)^2}$$

$$b_1 = \frac{n \sum X_i Y_i - \sum X_i \sum Y_i}{n \sum X_i^2 - (\sum X_i)^2}$$

- These estimates are called the **Least Squares** estimates
- The Least Squares estimate line minimizes errors more than any other line.

# Output of Simple Linear Regression

SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0.568							
R Square	0.323							
Adjusted R Square	0.304							
Standard Error	113822.683							
Observations	38							
ANOVA								
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>			
Regression	1	2.23281E+11	2.23281E+11	17.234	0.000193336			
Residual	36	4.66402E+11	12955603231					
Total	37	6.89682E+11						
	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>	<i>Lower 95.0%</i>	<i>Upper 95.0%</i>
Intercept	881382.223	79117.404	11.14	3.21401E-13	720924.689	1041839.756	720924.689	1041839.756
Promotion	2.423	0.583	4.151	0.00019	1.239	3.607	1.239	3.607

# Output Interpretation

- You can see that there are three distinct tables in the output
  - ✓ Regression Statistics Table
  - ✓ ANOVA Table
  - ✓ Coefficient Table



# Coefficients Table

- In the below coefficients table, the model that has been generated on the data that best possible straight line is actually this line.
- Advertising Revenue =  $881382.223 + 2.423 * \text{Promotion}$
- In our straight line equation, the slope is 2.423 and intercept is 881382.223.
- The intercept of 881382.223 explains that with zero promotion cost we expect the advertising revenue is positive.
- With the slope of 2.423 explains that 1 unit increase in promotion cost increases advertising cost by 2.423 but the reverse is not true means the 2.423 increase in advertising cost doesn't explain that promotion cost is increased by 1 unit due to cause and effect relationship.
- The straight line consist with above intercept and slope is best fitted line.

$$\text{Advertising Revenue} = 881382.223 + 2.423 * \text{Promotion}$$

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	881382.223	79117.404	11.14	3.21401E-13	720924.689	1041839.756	720924.689	1041839.756
Promotion	2.423	0.583	4.151	0.00019	1.239	3.607	1.239	3.607

# What is the Advertising Revenue?

- What is the advertising revenue if the promotion cost is 1,40,000?
- Advertising Revenue =  $881382.223 + 2.423 * \text{Promotion} = 881382.223 + 2.423 * 1,40,000 = 12,20,602.223$

# Hypothesis test & P value

- We conduct a hypothesis test to determine whether there is a significant linear relationship between an independent variable  $X$  and a dependent variable  $Y$ .
- $Y = a + bx$
- If we find that the slope of the regression line is significantly different from zero, we will conclude that there is a significant relationship between the independent and dependent variables. ( $b \neq 0$ )
- **State the Hypotheses**
- If there is a significant linear relationship between the independent variable  $X$  and the dependent variable  $Y$ , the slope will *not* equal zero

**H<sub>0</sub>:** A statistically significant relationship doesn't exist between the two variables. ( $B = 0$ )

**H<sub>a</sub>:** A statistically significant relationship exists between the two variables. ( $B \neq 0$ )

In our Example, P value is 0.00019 which is less than 0.05

- Hence, We can reject the null hypothesis and say that slope ( $b$ ) is not equal to 0 means "a statistically significant relationship exists between promotion and revenue".

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	881382.223	79117.404	11.14	3.21401E-13	720924.689	1041839.756	720924.689	1041839.756
Promotion	2.423	0.583	4.151	0.00019	1.239	3.607	1.239	3.607

# Standard Error

- "Standard error" gives the standard errors (i.e. the estimated standard deviation) of the least squares estimates  $b_j$  of  $\beta$
- Y variable is a random variable which is influenced by factors that are outside of anyone's control.
- There will always be some variation in Y that cannot be explained. It's because of random variation.
- In our example standard error is 0.583

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	881382.223	79117.404	11.14	3.21401E-13	720924.689	1041839.756	720924.689	1041839.756
Promotion	2.423	0.583	4.151	0.00019	1.239	3.607	1.239	3.607

# Confidence Interval

- In statistics, a **confidence interval (CI)** is a type of interval estimate(of a population parameter) that is computed from the observed data.
- The **confidence level** is the frequency (i.e., the proportion) of possible confidence intervals that contain the true value of their corresponding parameter.
- Smaller the standard error narrower the confidence interval
- With 95% confidence interval, when x increases by 1 unit doesn't mean that y will always increase by 2.423 it may increase more or less than 2.423.
- In our example 1, We can say that when x increases by 1 unit 95% times y will decrease to 1.239 and increase to 3.607.
- So we know 95% of the time this values show how much Y will increase by.

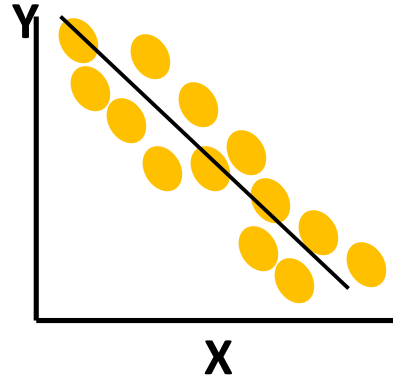
	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>	<i>Lower 95.0%</i>	<i>Upper 95.0%</i>
Intercept	881382.223	79117.404	11.14	3.21401E-13	720924.689	1041839.756	720924.689	1041839.756
Promotion	2.423	0.583	4.151	0.00019	1.239	3.607	1.239	3.607

# Coefficient of Determination – R2

- The coefficient of determination (denoted by R2) is a key output of regression analysis. It is interpreted as the proportion of the variance in the dependent variable that is predictable from the independent variable. (Goodness of fit)
- Coefficient of determination is used in trend analysis & computed as a value between 0 and 1
- Our aim is to find out the best possible straight line that explains or captures the relationship between X and Y.
- The higher the value of R Square the more variation in Y and therefore the model is good.
- It is a good measure of variation but it is only one measure of model fit.

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

# Coefficient of Determination – R<sup>2</sup>

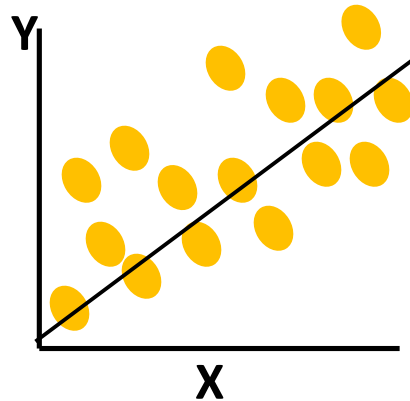


$$R^2 = 1$$

$$R^2 = 1$$

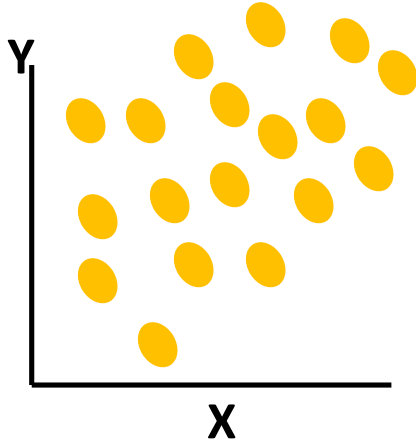
Perfect linear relationship  
between x and y:

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



$$R^2 = +1$$

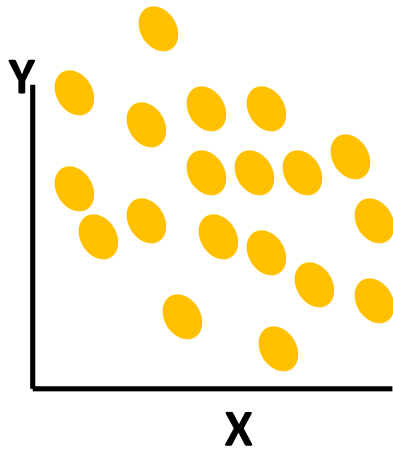
# Coefficient of Determination – R<sup>2</sup>



$$0 < R^2 < 1$$

Weaker linear relationship between x and y:

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





# R-Square: Television Production House Example 2

32.37% of the variation in advertising revenue is explained by variation in promotion in money

SUMMARY OUTPUT									
Regression Statistics									
Multiple R	0.568								
R Square	0.323								
Adjusted R Square	0.304								
Standard Error	113822.683								
Observations	38								
ANOVA									
	df	SS	MS	F	Significance F				
Regression	1	2.23281E+11	2.23281E+11	17.234	0.000193336				
Residual	36	4.66402E+11	12955603231						
Total	37	6.89682E+11							
	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%	
Intercept	881382.223	79117.404	11.14	3.21401E-13	720924.689	1041839.756	720924.689	1041839.756	
Promotion	2.423	0.583	4.151	0.00019	1.239	3.607	1.239	3.607	

$$R^2 = \frac{SSR}{SST} = \frac{2.23281E + 11}{6.89682E + 11} = 0.3237 = 32.37\%$$

# R-Square: Television Production House Example 2

- In our example, The R Square is 32.37%
- It means that 32.37% of the variation in advertising revenue is explained by variation in promotion money

SUMMARY OUTPUT	
<i>Regression Statistics</i>	
Multiple R	0.568
R Square	0.323
Adjusted R Square	0.304
Standard Error	113822.683
Observations	38

# Adjusted $R^2$

- The adjusted R-squared is a modified version of R-squared that has been adjusted for the number of predictors in the model.
- The adjusted R-squared increases only if the new term improves the model more than would be expected by chance.
- The adjusted R-squared compares the explanatory power of regression models that contain different numbers of predictors.
- It decreases when a predictor improves the model by less than expected by chance.
- The adjusted R-squared can be negative, but it's usually not.
- It is always lower than the R-squared.
- In our Example, the adjusted  $R^2$  is 30.4%

SUMMARY OUTPUT	
<i>Regression Statistics</i>	
Multiple R	0.568
R Square	0.323
Adjusted R Square	0.304
Standard Error	113822.683
Observations	38

# Standard Error

- The standard error of the mean is the standard deviation of the sampling distribution of the mean.
- The magnitude of the standard error of the mean depends on both the variability of the observation(s) and no of observations (n).
- It is the standard deviation of a large number of sample means of the same sample size drawn from sample population.
- Standard error is used to estimate the variation in the set of means.

Where,

$$SE = \frac{s}{\sqrt{n}}$$

s = Sample standard deviation

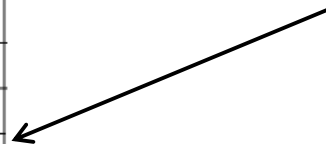
n= No. of observations in the sample

# Standard Error

- Standard error is a statistical term that measures the accuracy with which a sample represents a population.
- In statistics, a sample mean deviates from the actual mean of a population; this deviation is the standard error.

SUMMARY OUTPUT	
<i>Regression Statistics</i>	
Multiple R	0.568
R Square	0.323
Adjusted R Square	0.304
Standard Error	113822.683
Observations	38

$$S_e = 113822.683$$



# ANOVA Table

- ANOVA and Regression have similar kind of model called as generalized linear model.
- F test is used to test the significance in regression
- With only one independent variable f test will provide same conclusion as t test
- If t test indicates  $b \neq 0$  and hence a significance relationship, the f test will also indicate a significant relationship.
- With more number of independent variable only F test is used for overall significant relationship.

ANOVA					
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>
Regression	1	2.23281E+11	2.23281E+11	17.234	0.000193336
Residual	36	4.66402E+11	12955603231		
Total	37	6.89682E+11			



# Multiple Linear Regression

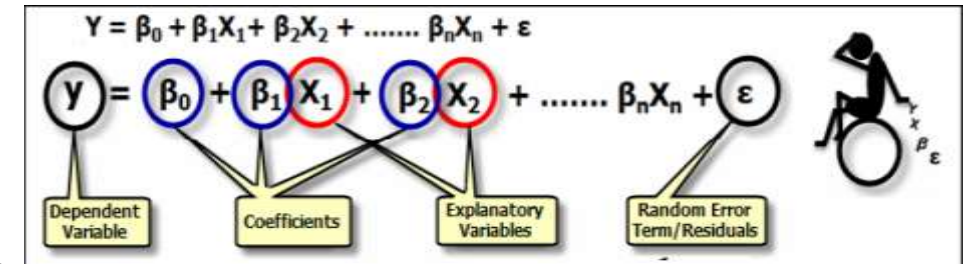
# Multiple Linear Regression

- Let's continue our example 1 (Simple linear regression), the television production house decided to predict revenue by considering only one factor i.e. promotion to predict revenue by considering only one factor i.e. promotion to find out the advertising revenue.
- Now they decide to add other factors i.e. television rating point (TRP), Promotion, Language of Program to predict the revenue.
- Here, the dependent and independent variables are
  - X= Television rating point(TRP), Promotion, Language of program.
  - Y= Advertising Revenue
- So, this is the form of Multiple linear regression



# Multiple Regression

- Multiple regression is the most common form of linear regression analysis.
- The multiple linear regression is used to explain the relationship between one dependent and two or more independent variables.
- The independent variables can be continuous or categorical.
- The multiple linear regression fits a line through a multi-dimensional space of data points.
- There are 3 major uses for multiple linear regression analysis are as follows.
  1. To identify the strength of the effect that the independent variables have on a dependent variable.
  2. To forecast effects or impacts of changes.
  3. It predicts trends and future values.
  4. It is used to get point estimates.
- An important consideration for selecting a model is model fit.
- Addition of an independent variables to a model will increase the amount of explained variance in the dependent variable (i.e.  $R^2$ ).
- Addition of many independent variables without any theoretical justification may result in an over-fit model.



# Multiple Linear Regression Example

- In our current example, the dependent and independent variables are,
  - X = Television rating point (TRP), Promotion, Language of program.
  - Y = Advertising Revenue
- Let's see how to calculate the advertising revenue based on other factors using Excel in Multiple linear regression
- The Multiple linear regression equation is

$$Y = a + b_1 * \text{Television Rating Point} + b_2 * \text{Promotion} + b_3 * \text{Language}$$

# How to Analyse a Linear Model

## **Step 1: Exploratory Data Analysis(EDA)**

- Univariate Analysis.
  - Analysing each variable for missing value treatment, Outlier treatment.
- Bi-variate Analysis
  - Analysing each independent variable with the dependent variable to check the relationship between the variables.

## **Step 2:Model Building.**

- Model Building involves checking the multi-collinearity and forming an linear regression equation.

## **Step 3: Checking the Assumptions.**

- There are principal assumptions which justify the use of linear regression models for purposes of inference or prediction which is one of the most important step before validation

## **Step 4: Model Validation**

- The validation process can involve analyzing the goodness of fit of the regression, analyzing whether the regression residuals are random, and checking whether the model's predictive performance deteriorates substantially when applied to data that were not used in model estimation.

# Assumptions for Linear Regression

- **Assumption 1 :** There must be a linear relationship between the outcome variable and the independent variables  
How to Check it: Scatterplots can show whether there is a linear or curvilinear relationship or no relationship between the variables.
- **Assumption 2:** There must be no Linear relationship between the Residuals and the independent variables or between the residuals and the fitted value(Predicted variable)  
How to Check it: Scatterplot can be drawn considering Residuals as Y variable and independent variables as X variables or Residuals as Y variables and Fitted values as X variables respectively.
- **Assumption 3 :** Multicollinearity means that some of the Independent variables are highly correlated with each other.  
How to check it: Calculating the Variance inflation factor.
- **Assumption 4:** Homoscedasticity must be present as it describes a situation in which the error term in the relationship between the independent and the dependent variable is same across all values of the independent variables.

How to Check: Using a scatter plot for residuals

**We will learn about assumptions in detail in the up coming slides.**

# Dataset 3

Serial	TRP	Promotion	Language	Revenue
1	133	111600	1	1197576
2	111	104400	0	1053648
3	129	97200	1	1124172
4	117	79200	1	987144
5	130	126000	1	1283616
6	154	108000	1	1295100
7	149	147600	0	1407444
8	90	104400	0	922416
9	118	169200	0	1272012
10	131	75600	0	1064856
11	141	133200	0	1269960
12	119	133200	1	1064760
13	115	176400	0	1207488
14	102	180000	0	1186284
15	129	133200	0	1231464
16	144	147600	1	1296708
17	153	122400	1	1320648
18	96	158400	0	1102704
19	104	165600	1	1184316
20	156	104400	1	1326360
21	119	136800	1	1162596
22	125	115200	1	1195116

# Multiple Linear Regression Output

SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0.932725089							
R Square	0.869976091							
Adjusted R Square	0.858503393							
Standard Error	51356.65693							
Observations	38							
ANOVA								
	df	SS	MS	F	Significance F			
Regression	3	6.00007E+11	2.00002E+11	75.83012324	3.86382E-15			
Residual	34	89675211173	2637506211					
Total	37	6.89682E+11						
	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	13409.52169	81642.40721	0.164247015	0.870509245	-152507.8121	179326.8555	-152507.8121	179326.8555
TRP	5841.424103	515.3816965	11.33417066	4.29687E-13	4794.04248	6888.805726	4794.04248	6888.805726
Promotion	3.240654727	0.272586637	11.88853114	1.16478E-13	2.68669203	3.794617425	2.68669203	3.794617425
Language	53211.02783	16871.23458	3.153949854	0.003359968	18924.554	87497.50167	18924.554	87497.50167

# Coefficient Table

- In coefficient table one thing we need to check is the sign of the coefficient and p value.
- $Y = 41060.006 + 5481.424 * TRP + 3.240 * Promotion + 53211.027 * Language$
- The positive coefficients indicates that it have positive effect
- The null and alternate hypothesis of our example is mentioned below:
  - ✓ **Ho:** A statistically significant relationship doesn't exists between the two variable.
  - ✓ **Ha:** A statistically significant relationship exists between the two variable.
- The p-value of all independent variables are less than significance level. Means we reject null hypothesis. Hence all variables are significant.

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	13409.52169	81642.40721	0.164247015	0.870509245	-152507.8121	179326.8555	-152507.8121	179326.8555
TRP	5841.424103	515.3816965	11.33417066	4.29687E-13	4794.04248	6888.805726	4794.04248	6888.805726
Promotion	3.240654727	0.272586637	11.88853114	1.16478E-13	2.68669203	3.794617425	2.68669203	3.794617425
Language	53211.02783	16871.23458	3.153949854	0.003359968	18924.554	87497.50167	18924.554	87497.50167

# Assumptions

- Before we start on assumptions, let's understand the Residuals.
- Residuals are the difference between Predicted Values of Y and the Actual Values of Y

Serial Number	Actual	Predicted	Residuals
1	1197576	1205187.023	-7611.02
2	1053648	1053342.979	305.0215
3	1124172	1081944.871	42227.13
4	987144	953515.9962	33628
5	1283616	1234328.179	49287.82
6	1295100	1316190.572	-21090.6
7	1407444	1362102.351	45341.65
8	922416	930673.0724	-8257.07
9	1272012	1251016.346	20995.65
10	1064856	1023629.577	41226.42
11	1269960	1268705.53	1254.47
12	1064760	1140194.2	-75434.2
13	1207488	1256824.787	-49336.8
14	1186284	1192552.631	-6268.63
15	1231464	1251819.469	-20355.5
16	1296708	1386106.258	-89398.3
17	1320648	1357014.576	-36366.6
18	1102704	1087505.944	15198.06
19	1184316	1210781.079	-26465.1
20	1326360	1316207.063	10152.94

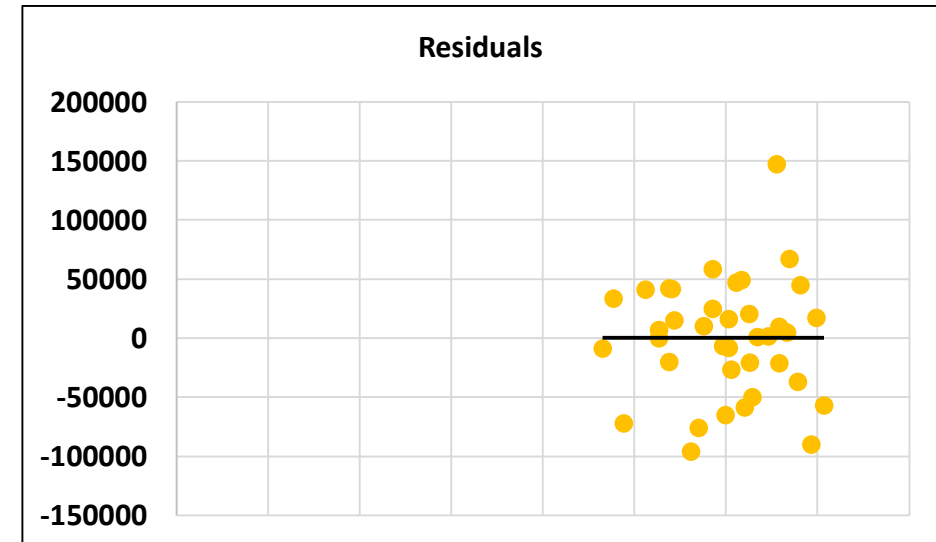


## CHECKING IF ASSUMPTIONS ARE VALID

### 1. Plot the residuals against Predicted values

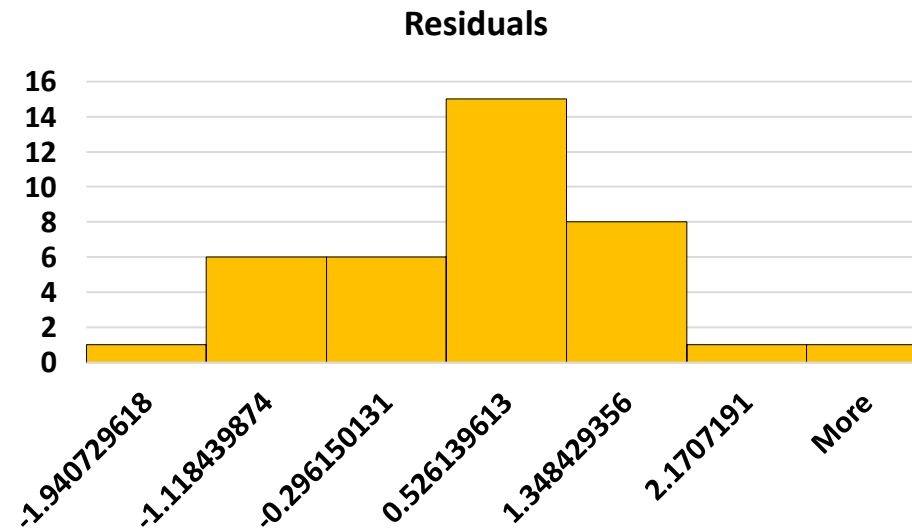
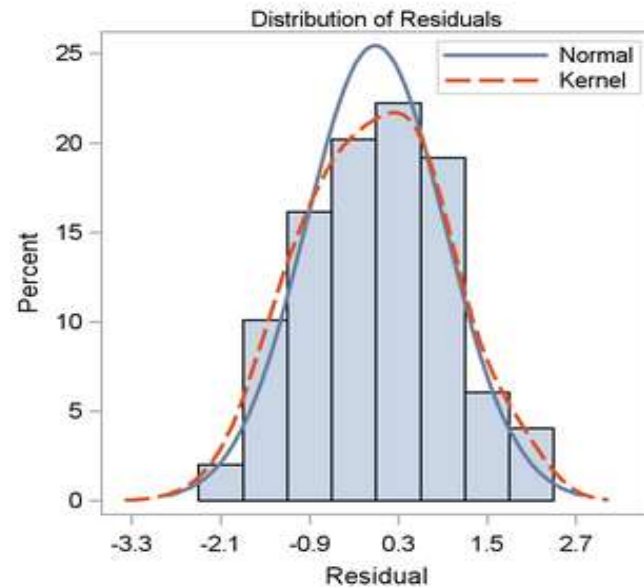
- Residuals = Observed value – Predicted value (Fitted Value)
- For linear regression, the linear relationship should not be present between residuals and predicted values.
- If a relationship is present, it means that the predict values do not fit in the model.
- It is important to check for outliers since linear regression is sensitive to outlier effect
- This assumption is tested with scatter plot

The Scatter plot clearly shows that there is no relationship between the residuals and the predicted variables.



# Assumption - Normality

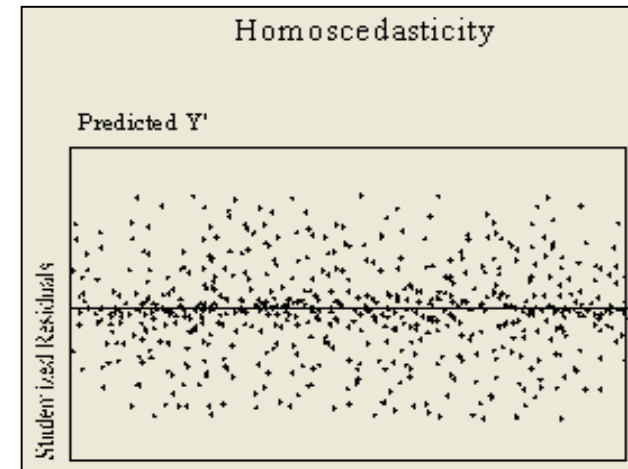
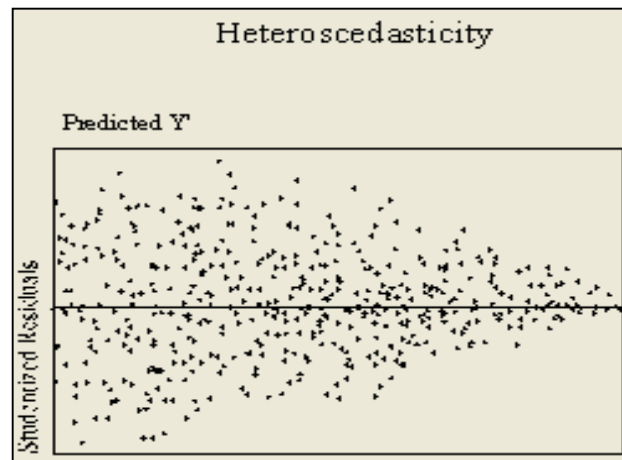
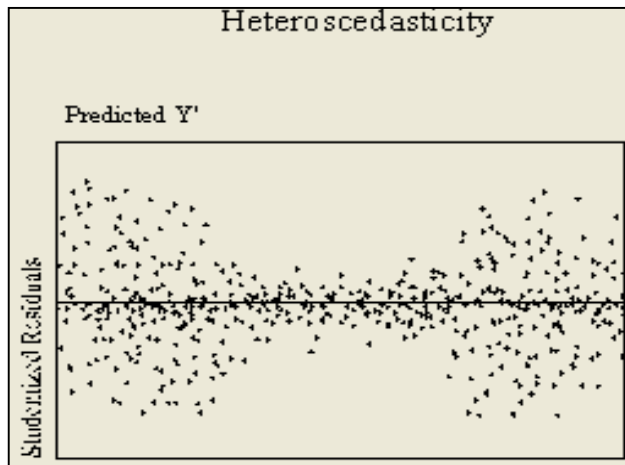
- MLR requires that the error between observed and predicted values (i.e., the residuals of the regression) should be normally distributed. (ref the below distribution of residuals diagram)
- This assumption can best be checked by plotting residual values on a histogram with a fitted normal curve or by reviewing a Q-Q-Plot.
- Normality can also be checked with a goodness of fit test e.g., the Kolmogorov-Smirnov test
- When the data is not normally distributed, a non-linear transformation might correct this issue if one or more of the individual predictor variables are to blame, though this does not directly respond to the normality of the residuals.



Our histogram shows that residuals in our model normally are distributed

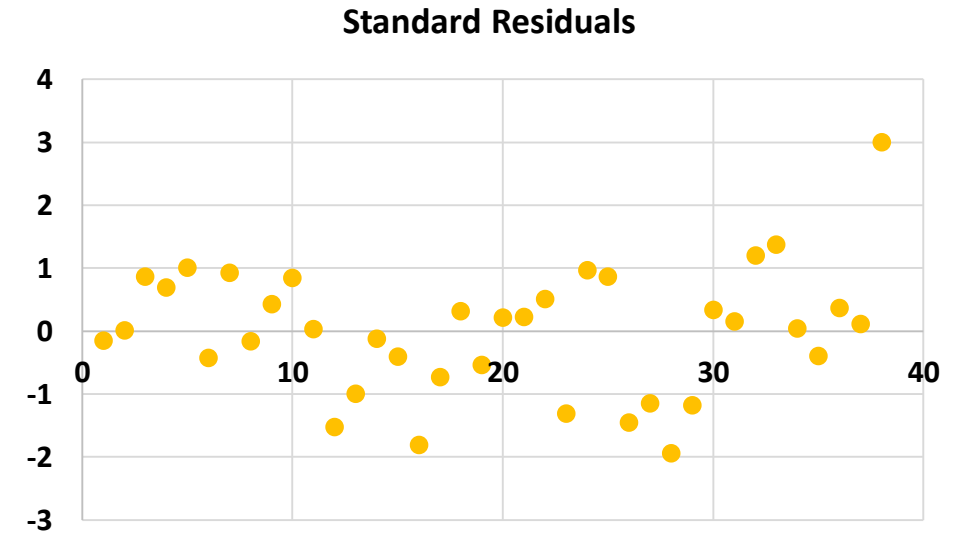
# Assumption – Homoscedasticity and Heteroscedasticity

- Homoscedasticity describes a situation in which the error term is normally distributed, there is no such pattern in the residuals as shown in the diagram 3.
- Heteroscedasticity describes a situation when the residuals follow a specific pattern as shown in the diagram 1 and diagram 2.
- For a model to be good fit, Homoscedasticity should be present (No specific pattern) and there should be no Heteroscedasticity (a specific data Pattern)



# Assumption- Homoscedasticity and Heteroscedasticity

- According to the assumptions, there must be no pattern in the standard residuals scatter plot but we can see a pattern in the scatter plot in our scatter plot. Which means, heteroscedasticity is present.
- The solution to this problem is log transformation and we will learn more about the log transformation in the next case study done on R.



## Assumption – Multi-collinearity

- Multiple linear regression analysis requires that there is little or no multi-collinearity in the independent variables.
- Multicollinearity generally occurs when there are high correlations between two or more predicted variables.
- In other words, one predictor variable can be used to predict the other. This creates a false information, skewing the results in a regression model.
- Examples of correlated predictor variables (also called multicollinear predictors) are: a person's height and weight, age and sales price of a car, or years of education and annual income.
- Cook's distance is a commonly used estimate of the influence of a data point when performing a least-squares regression analysis

# Model Validation

## Validate the model

- $R^2$
- Adjusted  $R^2$
- Fit Chart - Actual vs Fitted Values
- MAPE – Mean Absolute Percentage Error
- RMSE – Root Mean Square Value

# R Square

- **R-squared** supposes that every independent variable in the model explains the variation in the dependent variable.
- It gives the percentage of explained variation as if all independent variables in the model affect the dependent variable
- The drawback of using Coefficient of determination is that the value of the coefficient of determination always increases as the number of independent variables are increased even if the marginal contribution of the incoming variable is statistically insignificant.
- To take care of the above drawback, coefficient of determination is adjusted for the number of independent variables taken. This adjusted measure of coefficient of determination is called adjusted  $R^2$
- Lets continue example 1, The value of R square is 86.99%

Regression Statistics	
Multiple R	0.932725089
R Square	0.869976091
Adjusted R Square	0.858503393
Standard Error	51356.65693
Observations	38



# Adjusted R Square

- The **adjusted R-squared** gives the percentage of variation explained by only those independent variables that in reality affect the dependent variable.
- The adjusted R-squared compensates for the addition of variables and only increases if the new term enhances the model above what would be obtained by probability and decreases when a predictor enhances the model less than what is predicted by chance.
- In an overfitting condition, an incorrectly high value of R-squared, which leads to a decreased ability to predict, is obtained. This is not the case with the adjusted R-squared.
- Lets continue example 1, The value of adjusted R square is 85.85%
- Adjusted  $R^2$  is given by the following formula:

$$R_a^2 = 1 - \left[ \left( \frac{n-1}{n-k-1} \right) \times (1 - R^2) \right]$$

where

N = Number of Observations

k = Number of Independent Variables

= Adjusted  $R^2$

Regression Statistics	
Multiple R	0.932725089
R Square	0.869976091
Adjusted R Square	0.858503393
Standard Error	51356.65693
Observations	38

# Actual vs Fitted Values

- Fitted Values means we fit a x value in a straight line equation and finding the value of Y.
- Fitted values are values of the Dependent variable (Advertising Revenue) according to the model.
- We can automatically generate the fitted values in Excel, using the actual data values for the X variable values:

$$Y = 13409.52 + 5841.42 * \text{Television Rating Point} + 3.24 * \text{Promotion} + 53211.02 * \text{Language}$$

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
10	9	118	169200	0	1272012											
11	10	131	75600	0	1064856											
12	11	141	133200	0	1269960											
13	12	119	133200	0	1064760											
14	13	115	176400	0	1207488											
15	14	102	180000	0	1186284											
16	15	129	133200	1	1231464											
17	16	144	147600	1	1296708											
18	17	153	122400	1	1320648											
19	18	96	158400	0	1102704											
20	19	104	165600	1	1184316											
21	20	156	104400	1	1326360											
22	21	119	136800	0	1162596											
23	22	125	115200	1	1195116											
24	23	130	115200	1	1134768											
25	24	123	151200	0	1269024											
26	25	128	97200	0	1118688											
27	26	97	122400	0	904776											
28	27	124	208800	0	1357644											
29	28	138	93600	0	1027308											
30	29	137	115200	1	1181976											
31	30	129	118800	1	1221636											

Regression

Input

Input Y Range:

Input X Range:

☒ Labels ☐ Constant is Zero

☐ Confidence Level:  %

Output options

☐ Output Range:

☒ New Worksheet Ply:

☐ New Workbook

Residuals

☒ Residuals ☐ Residual Plots

☐ Standardized Residuals ☐ Line Fit Plots

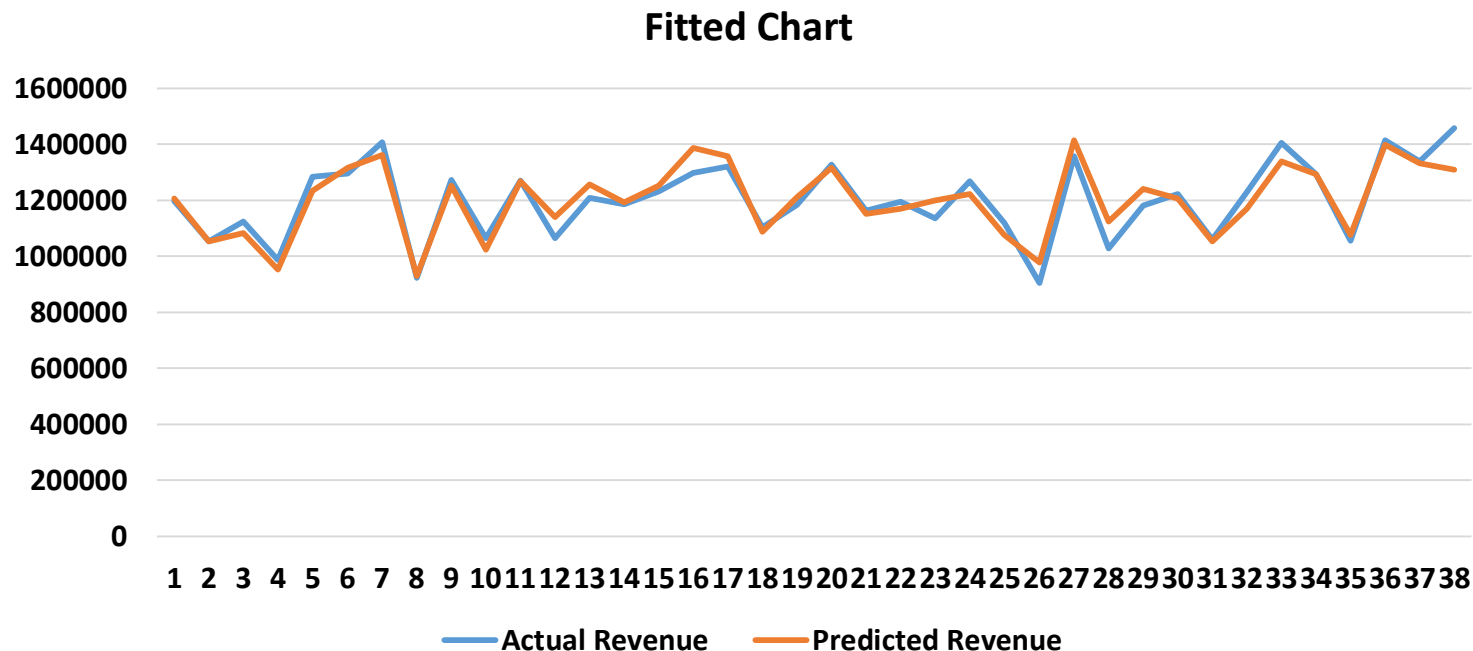
Normal Probability

☐ Normal Probability Plots

OK Cancel Help

# Actual vs Fitted Values

- Fitted value is used to validate a model
- In a good model, the difference between actual and predicted is very less and are close to each other
- A good model have higher number of overlap



Actual Revenue	Predicted Revenue
1197576	1205187.023
1053648	1053342.979
1124172	1081944.871
987144	953515.9962
1283616	1234328.179
1295100	1316190.572
1407444	1362102.351
922416	930673.0724
1272012	1251016.346

# MAPE – Mean Absolute Percentage Error

- The **mean absolute percentage error (MAPE)**, also known as **mean absolute percentage Deviation (MAPD)**.
- MAPE is a measure of prediction accuracy of a forecasting method in statistic.
- It has major Drawback in practical application

$$\text{MAPE} = \frac{1}{n} \sum_{i=1}^n \frac{|Y_i - \hat{Y}_i|}{Y_i}$$

# MAPE Example

- Lets continue example 1 to find out the MAPE.
- In our example mean absolute percentage error is 3%
- Ideally MAPE values are low i.e. 5% or lower.

	Actual Advertising Revenue	Predicted Advertising Revenue	Residuals	Absolute Residuals	Abs value of error/ actual value		
Serial Number	Actual	Predicted	Actual – Predicted	Absolute of Residuals			
1	1197576	1205187	-7611.02	7611.023	0.006355		
2	1053648	1053343	305.0215	305.0215	0.000289	MAPE	3%
3	1124172	1081945	42227.13	42227.13	0.037563		
4	987144	953516	33628	33628	0.034066		
5	1283616	1234328	49287.82	49287.82	0.038398		
6	1295100	1316191	-21090.6	21090.57	0.016285		
7	1407444	1362102	45341.65	45341.65	0.032216		
8	922416	930673.1	-8257.07	8257.072	0.008952		
9	1272012	1251016	20995.65	20995.65	0.016506		
10	1064856	1023630	41226.42	41226.42	0.038715		
.....	.....	.....	.....	.....	.....		
37	1338060	1332763	5296.699	5296.699	0.003958		
38	1457400	1310052	147347.7	147347.7	0.101103		
Total					1.183679		



**Thank You.**