Analysing Relations between data set -I

2012

Author: Naseha Sameen

About: Correlation, covariance and regression between variable, their interpretation and when to use what….

Analysing Relation

LAB NOTEBOOK

NASEHA SAMEEN

Table of Contents

Covariance

Covariance Meaning and Perspective:

Covariance are a measure of the "spread" of a set of points around their centre of mass (mean). It is a measure of how much each of the dimensions vary from the mean with respect to each other.

Covariance is measured between 2 dimensions to see if there is a relationship between the 2 dimensions e.g. number of hours studied & marks obtained. How much two random variables change together. For example, if X is changing, is it inducing similar or opposite effect on Y?

It is also a measurement of strength or weakness of correlation between two or more sets of random variables. To simplify, a covariance tries to look into and measure how much variables change together.

How to Measure degree of Relationship:

Covariance captures the degree to which pairs of points systematically vary around their respective means.

If paired X and Y values tend to both be above or below their means at the same time, this will lead to a high positive covariance. If the paired X and Y values tend to be on opposite sides of their respective means, this will lead to a high negative covariance.

A positive covariance means the variables are positively related, while a negative covariance means the variables are inversely related. However covariance of 0, may mean that random variables are not related or there is a non-linear relationship.

The strength of the linear relationship however cannot be easily interpreted by the magnitude of the calculated value. In order to interpret the strength a related measure called correlation is used.

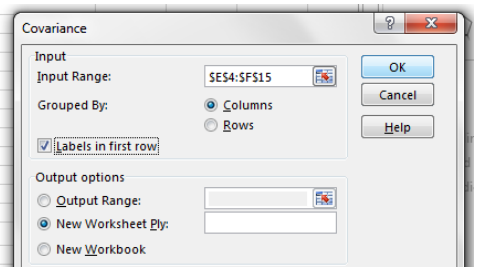
The numerical value of covariance is not very meaningful as it is presented in terms of units squared, and can range from negative infinity to positive infinity

The coefficient of covariance does not have a standard symbol.

How to Calculate:

a. There are two ways to calculate r in Excel
    a. Using the formula
        i. Syntax – COVARIANCE.S(List of variables x, List of variable y)
        ii. Result will be generated in decimal format
        iii. Using Formulae, relation between only two variables can be derived
        iv. PN: COVAR and COVARIANC.P (population) uses n in denominator instead of n-1
    b. Using Statistical Tool in Excel
        i. Using Statistical Tool, relation between two or more variables can be derived
        ii. Under the Tab "DATA", click on the Option, "Data Analysis"
        iii. Click on Covariance
        iv. Enter the range

Sl.No	x (Height)	y (Weight)	xi-x'	yi-y'	xi*yi
1	69	108	3.6	-17.8	7452
2	61	130	-4.4	4.2	7930
3	68	135	2.6	9.2	9180
4	66	135	0.6	9.2	8910
5	66	120	0.6	-5.8	7920
6	63	115	-2.4	-10.8	7245
7	72	150	6.6	24.2	10800
8	62	105	-3.4	-20.8	6510
9	62	115	-3.4	-10.8	7130
10	67	145	1.6	19.2	9715
11	66	132	0.6	6.2	8712
12	63	120	-2.4	-5.8	7560



v. The output would be in a grid. The covariance between x and y would be in the corresponding grid. Here B, which is 23.73

	x (Height)	*y (Weight)*
x (Height)	10.07639	
y (Weight)	23.73611	185.8056

c. Calculate Mathematically

i. Covariance of a sample set: $COV(x,y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$

x = the independent variable
y = the dependent variable
n = number of data points in the sample
$\bar{x}$ = the mean of the independent variable *x*
$\bar{y}$ = the mean of the dependent variable *y*

Sl.No	x (Height)	y (Weight)	xi-x'	yi-y'	(xi-x')*(yi-y')
1	69	108	3.6	-17.8	-63.9
2	61	130	-4.4	4.2	-18.4
3	68	135	2.6	9.2	23.7
4	66	135	0.6	9.2	5.3
5	66	120	0.6	-5.8	-3.4
6	63	115	-2.4	-10.8	26.2
7	72	150	6.6	24.2	159.1
8	62	105	-3.4	-20.8	71.2
9	62	115	-3.4	-10.8	37.0
10	67	145	1.6	19.2	30.3
11	66	132	0.6	6.2	3.6
12	63	120	-2.4	-5.8	14.1
Mean'	**65**	**126**			
Sum	**785**	**1,510**			**285**
Count	**12**	**Count-1**	11.0	sum[(xi-x')*(yi-y')]/count-1 =26	

i. The above formulae can also be expressed as

$$\text{cov } xy = \frac{1}{N}\Sigma\, x_i y_i - \overline{xy}$$

x = the independent variable
y = the dependent variable

n = number of data points in the sample
$\bar{x}$ = the mean of the independent variable *x*
$\bar{y}$ = the mean of the dependent variable *y*

Sl.No	x (Height)	y (Weight)	xi-x'	yi-y'	xi*yi
1	69	108	3.6	-17.8	7452
2	61	130	-4.4	4.2	7930
3	68	135	2.6	9.2	9180
4	66	135	0.6	9.2	8910
5	66	120	0.6	-5.8	7920
6	63	115	-2.4	-10.8	7245
7	72	150	6.6	24.2	10800
8	62	105	-3.4	-20.8	6510
9	62	115	-3.4	-10.8	7130
10	67	145	1.6	19.2	9715
11	66	132	0.6	6.2	8712
12	63	120	-2.4	-5.8	7560
Mean'	**65**	**126**			
Sum	**785**	**1,510**			**99,064**
Count	**12**	**Count-1**	11.0	=Sum(xi*yi)-[(Sum x * Sum y)/N] =99064-(785*1510)/12 =26	

Correlation

Correlation Meaning and Perspective:

Correlation by definition indicates the relationship between the variables.

This term is used when both the variables are random and the end result of the activity is to find out the relation between the variables. In other words, are the variables related, if yes, how? And to what degree the variables are related?

Mathematically, they provide a measure of degree variables vary together. In other words it tells how much one variable tends to change when the other one does.

How to Measure degree of Relationship:

To measure this we use the term correlation coefficient (Also known as Pearson's Correlation Coefficient) (r). It measures the linear association between two random variables. Values of the correlation coefficient varies between -1 and +1.

The sign indicates positive or negative relation

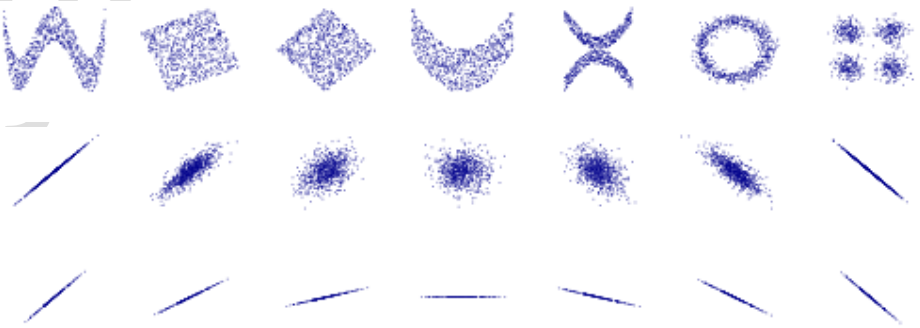
When r is 0, it indicates that there is no linear relationship between the two variables

r of +1 indicates that two variables are perfectly related in a positive linear sense. There is a trend. If one variable goes up as the other one goes up.

When r is of -1, there is a trend that one variable goes up as the other one goes down. The variables are perfectly related in a negative linear sense.

Correlation is used to test independence of the variables.

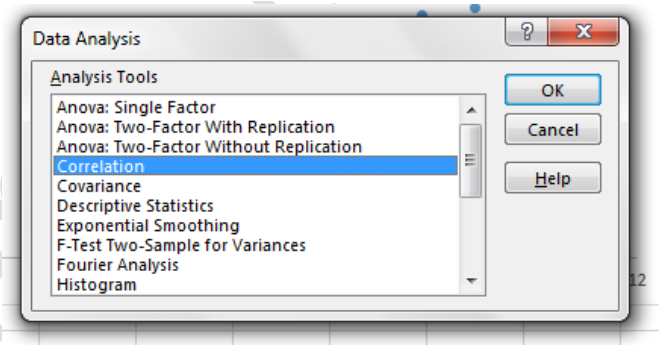
How to Calculate:



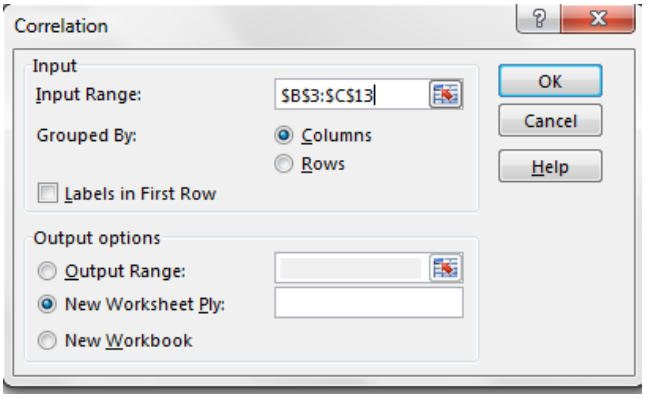
b. Create a Scatter Plot. Scatter plots can be of any of these shapes
c. Eye balling the shape roughly gives us the positive or negative trend.
d. To interpret its value, see where the value of r falls in the grid given below:

• **Exactly –1.** A perfect downhill			
Values between -.90 and -1.00 Variables are very strongly related Downhill	Values between -0.5 and -0.9 Variables are strongly related Downhill	Values between -0.1 and 0.5 Variables are weakly related Downhill	Values between -0.0 and -0.1 Little or no relationship between the variables Downhill
0. No linear relationship			
Values between 0.0 and 0.1 Little or no relationship between the variables	Values between 0.1 and 0.5 Variables are weakly related	Values between 0.5 and 0.9 Variables are strongly related	Values between 0.90 and 1.00 Variables are very strongly related
Exactly +1. A perfect uphill (positive) linear relationship			

e. There are two ways to calculate r in Excel
   a. Using the formula
      i. Syntax - CORREL(List of variables x, List of variable y)
      ii. Result will be generated in decimal format
      iii. Using Formulae, relation between only two variables can be derived
   b. Using Statistical Tool in Excel
      i. Using Statistical Tool, relation between two or more variables can be derived
      ii. Under the Tab "DATA", click on the Option, "Data Analysis"
      iii. From the popup select Correlation → Click on OK



      iv. Select the column in Input Range (Select Column if the data is in column format)



      v. In Output option select, either new workbook or same sheet
      vi. A table will be generated

	Experience	*Salary*	*Education in Yrs*
Experience	1		
Salary	0.963354	1	
Education in Yrs	-0.13593	0.055568	1

vii. Reading this table is easy, Correlation between Experience and Salary is 0.96, which is very strongly related. Correlation between Education and Experience and Education and Salary are weakly related. However, Correlation between Education and Experience is negatively related, while Education and Salary is positively related though very weakly.

c. Calculate Mathematically

i. Correlation Co-efficient: Correlation(r) =[$N\Sigma XY - (\Sigma X)(\Sigma Y) / Sqrt([N\Sigma X^2 - (\Sigma X)^2][N\Sigma Y^2 - (\Sigma Y)^2])$]

ii. Where N = Number of values or elements
X = First Score
Y = Second Score
ΣXY = Sum of the product of first and Second Scores
ΣX = Sum of First Scores
ΣY = Sum of Second Scores
ΣX^2 = Sum of square First Scores
ΣY^2 = Sum of square Second Scores

	X Values	Y Values	X*Y	X*X	Y*Y
	60	3.1	186	3600	9.61
	61	3.6	219.6	3721	12.96
	62	3.8	235.6	3844	14.44
	63	4	252	3969	16
N=5	65	4.1	266.5	4225	16.81
Sum	**311**	**18.6**	**1159.7**	**19359**	**69.82**

iii. Correlation(r) =[$N\Sigma XY - (\Sigma X)(\Sigma Y) / Sqrt([N\Sigma X^2 - (\Sigma X)^2][N\Sigma Y^2 - (\Sigma Y)^2])$]
= $((5)*(1159.7)-(311)*(18.6))/sqrt([(5)*(19359)-(311)^2]*[(5)*(69.82)-(18.6)^2])$
= (5798.5 - 5784.6)/sqrt([96795 - 96721]*[349.1 - 345.96])
= 13.9/sqrt(74*3.14)
= 13.9/sqrt(232.36)
= 13.9/15.24336 = 0.9119

d. If covariance is known then correlation can also be derived. The example is given in footnote as covariance is not yet explained here.

e. Word of caution, correlation does not mean causation. Cause and Effect cannot draw conclusions based on correlation. Because both the variables are interchangeable, we don't know the direction of the cause - Does X cause Y or does Y cause X? and a third variable "Z" may be involved that is responsible for the covariance between X and Y

Regression

Regression Meaning and Perspective:

Regression tells us the exact kind of linear association that exists between those two variables. The term is used when one of the variables is a fixed variable, and the end goal is to use the measure of relation to predict values of the random variable based on values of the fixed variable

In particular, it indicates the extent to which you can predict some variables by knowing others, or the extent to which some are associated with others.

How to Measure degree of Relationship:

Regression equation allows us to express the relationship between two (or more) variables algebraically. It indicates the nature of the relationship between two (or more) variables. It implicitly assumes that there is one-way cause and effect.

Linear regression finds the best line that predicts Y from X. It is usually used when X is a variable we manipulate (time, concentration, etc.) and result is indicated by Y.

Careful selection is required to choose X and Y. The line that best predicts Y from X is not the same as the line that predicts X from Y (however both those lines have the same value for R^2)

A linear regression equation is usually written
$Y = mX + b + e$
where
Y is the dependent variable
b is the intercept
m is the slope or regression coefficient
X is the independent variable (or covariate)
e is the error term
The equation will specify the average magnitude of the expected change in Y given a change in X.
The regression equation is often represented on a scatterplot by a regression line.

Linear regression quantifies goodness of fit with r^2, sometimes shown in uppercase as R^2. If you put the same data into correlation (which is rarely appropriate; see above), the square of r from correlation will equal r^2 from regression.

How to Calculate:

There are three ways to calculate r in Excel

a. Using the formulae
   i. LINEST – Syntax(List of variables for known y, List of variable for known x, optional value for stats and optional value for b - const)
      1. Select two cells say B3 & C3.
      2. Type LINEST and Select the ranges and then press Ctrl+Shift+Enter
      3. The first value is the slope m and the second value is the intercept b
      4. Replace them in the formulae mX+b to find out value of Y
   ii. Using SLOPE, INTERCEPT
      1. Slope, m: =SLOPE(known_y's, known_x's)
      2. y-intercept, b: =INTERCEPT(known_y's, known_x's)

3. Replace them in the formulae mX+b to find out value of Y

iii. To find R-square, use Syntax – RSQ (List of variables for known y, List of variable for known x) or just multiply result of CORREL with itself.

iv. The closer to 1, the better the regression line fits the data.

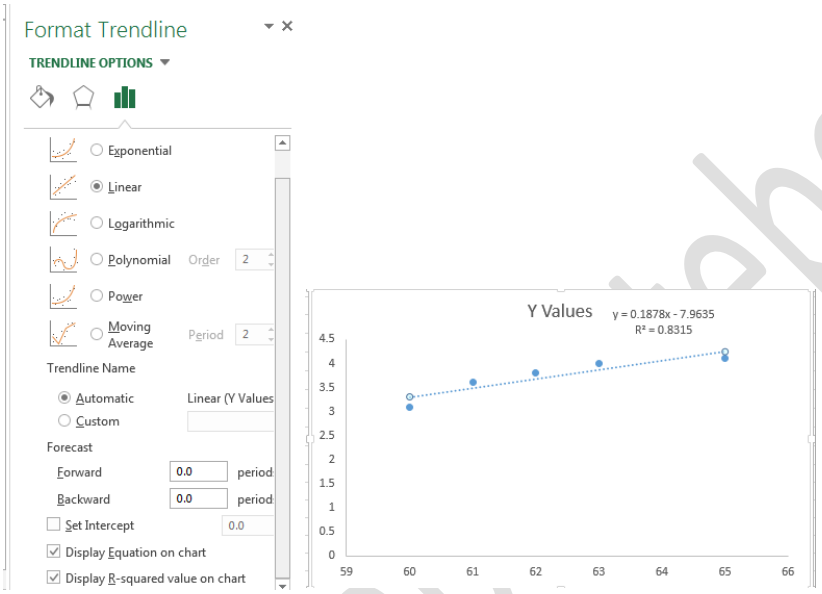
b. Using Graphs

i. Create a Scatter Plot from the values

ii. Draw a "best-fit" straight line through the data by right clicking on data points → click on Add trend line

iii. Select Linear in Trend line Options

iv. Select the options to display Equation and R-Square value on chart



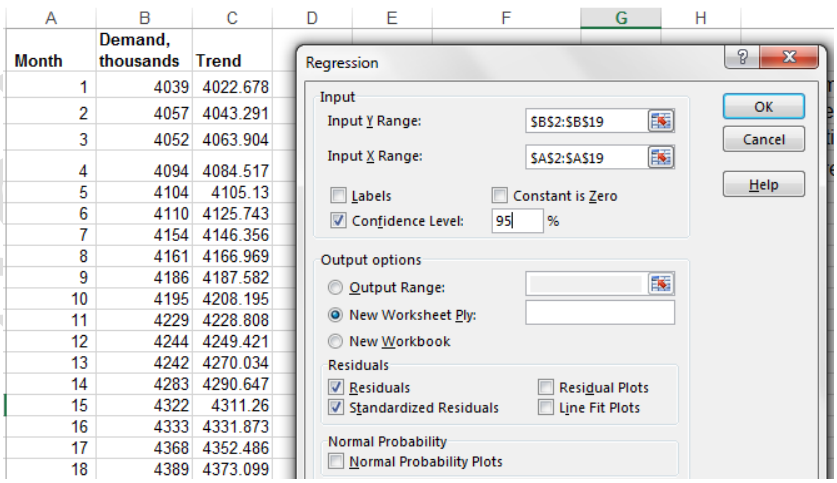
v. Copy the equation and replace the value of x for predictive value of Y

c. Using Statistical Tool in Excel

i. Under the Tab "DATA", click on the Option, "Data Analysis"

ii. Click on Regression

iii. Select the range in Input for X and Y, Output and confident level



Month	Demand, thousands	Trend
1	4039	4022.678
2	4057	4043.291
3	4052	4063.904
4	4094	4084.517
5	4104	4105.13
6	4110	4125.743
7	4154	4146.356
8	4161	4166.969
9	4186	4187.582
10	4195	4208.195
11	4229	4228.808
12	4244	4249.421
13	4242	4270.034
14	4283	4290.647
15	4322	4311.26
16	4333	4331.873
17	4368	4352.486
18	4389	4373.099

SUMMARY OUTPUT	
Regression Statistics	
Multiple R	0.993531
R Square	0.987103
Adjusted R Square	0.986297
Standard Error	12.96562
Observations	18

ANOVA								
	df	*SS*	*MS*	*F*	*Significance F*			
Regression	1	205862.0619	205862.0619	1224.587713	0.000000000			
Residual	16	2689.715858	168.1072411					
Total	17	208551.7778						
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	4002.065359	6.376000715	627.6764289	1.45262E-36	3988.548842	4015.581877	3988.548842	4015.581877
X Variable 1	20.6130031	0.589042078	34.99410969	1.51197E-16	19.36428967	21.86171652	19.36428967	21.86171652

iv. M (slope is indicated in Coefficient X variable 1, and intercept b is 4002)

v. So our regression equation is Y = 20.61(X)+4002.006

vi. Adjusted R square gives idea of the goodness of fit measure, here it says 98.62% of Y is determined by X

vii. The residuals show you how far away the actual data points are from the predicted data points. This is obtained from plotting x, y and y obtained from the formulae in a table and subtracting the y from y obtained from the table. Both the sum and the mean of the residuals are equal to zero.

viii. Standardized Residual It is the residual divided by the standard deviation of the residual; that is, it is a residual standardized to have standard deviation 1. If standardized residuals is larger than about ±2.5, should be investigated as a potential outlier. For very large samples, many observations could have standardized residuals outside ±2.5 while not being outliers.

ix. Random pattern of residuals indicates a good fit for a linear model. The other plot patterns are non-random (U-shaped and inverted U), suggesting a better fit for a non-linear model.

x. The sum of the residuals is always zero, whether the data set is linear or nonlinear.

xi. If Significance F should be lesser than 0.05. If is greater than 0.05, it's probably better to stop using this set of independent variables. Delete a variable with a high P-value (greater than 0.05) and rerun the regression until Significance F drops below 0.05.

xii. Most or all P-values should be below 0.05.

d. Calculate Mathematically

i. Y = mx + b, where

Student	Aptitude test xi	Statistics grades yi	$(x_i - x')$	$(x_i - x')^2$	$(y_i - y')$	$(y_i - y')^2$	$(x_i - x')(y_i - y')$
1	95	85	17	289	8	64	136
2	85	95	7	49	18	324	126
3	80	70	2	4	-7	49	-14
4	70	65	-8	64	-12	144	96
5	60	70	-18	324	-7	49	126
Sum	390	385		730		630	470
Mean	78	77					

ii. $m = \Sigma [(x_i - x')(y_i - y')] / \Sigma [(x_i - x')^2]$
$= 470/730$
$=0.64$

iii. $b = y' - m * x'$
$=77 - 0.64*78$
$=26.78$

iv. Therefore, the regression equation is: Y = 26.768 + 0.644x

e. Word of caution, when you use a regression equation, do not use values for the independent variable that are outside the range of values used to create the equation. That is called **extrapolation**, and it can produce unreasonable estimates.

What to use When:

Covariance	Correlation	Regression
Covariance tries to look into and measure how much variables change together Covariance is to quantify the dependence between two random variables *X* and *Y*	Correlation is used to test independence of the variables. Correlation tells us if the variables are related, and if so, how strong is the relation	Linear regression finds the best line that predicts Y from X.
Covariance calculations are used to find relationships between dimensions in high dimensional data sets (usually greater than 3) where visualization is difficult.	Correlation is almost always used when you measure both variables. It rarely is appropriate when one variable is something you experimentally manipulate.	Linear regression is usually used when X is a variable you manipulate (time, concentration, etc.). Y is a function of X, that is, it changes with X
The covariance itself gives us little info about the relation we are interested in, because it is sensitive to the standard deviation of X and Y. It must be transformed (standardized) before it is useful.	With correlation, you don't have to think about cause and effect. The strength of an association between two variables, and is completely symmetrical. It doesn't matter which of the two variables you call "X" and which you call "Y". You'll get the same correlation coefficient if you swap the two.	The decision of which variable you call "X" and which you call "Y" matters in regression, as you'll get a different best-fit line if you swap the two. The line that best predicts Y from X is not the same as the line that predicts X from Y (however both those lines have the same value for R2)
Exact value is not as important as it's sign. The value ranges from - ∞ through 0 to + ∞	Correlation computes the value of the Pearson correlation coefficient, r. Its value ranges from -1 to +1.	Linear regression quantifies goodness of fit with r^2, sometimes shown in uppercase as R^2.
Covariance is unstandardized version of correlation.	A correlation is covariance measured in the units of the original variables.	The square of r from correlation will equal r^2 from regression.

Foot Note:

- Derivation of Correlation Using covariance

  The correlation, r_{xv}, is defined as

  $$\text{Correlation} = r_{xy} = \frac{\text{cov } xy}{s_x s_y}$$

  In the formula for correlation, the products of the deviations are divided by the product of the standard deviations of *x* and *y* (s_x and s_y).

Student	Aptitude test xi	Statistics grades yi	$(x_i - x')$	$(x_i - x')^2$	$(y_i - y')^2$	$(x_i - x')(y_i - y')$
1	95	85	17	289	64	136
2	85	95	7	49	324	126
3	80	70	2	4	49	-14
4	70	65	-8	64	144	96
5	60	70	-18	324	49	126
Sum	390	385		730	630	470
Mean	78	77			=Covariance/(std x * std y) =117.5/(13.5*12.5) =0.69	
Std Dev	13.5	12.5	**Covariance**	117.5		

- Random Variable: A random variable, usually written X, is a variable whose possible values are numerical outcomes of a random phenomenon. There are two types of random variables, discrete and continuous.

- **Spurious correlation** arises when two variables are closely related but bear no causal relationship because they are both caused by a third, unexamined variable.

- ANCOVA: Analysis of covariance. The method of ANCOVA allows the analyst to make comparisons between groups that are not comparable with respect to some important variable, often referred to as a covariate. This is done by making an adjustment based on fitting a particular kind of regression line. In addition to allowing for imbalances, the method removes variation due to the covariate and therefore provides a more precise analysis. A geometrical interpretation is that the 'unexplained variation' with respect to which the significances of group differences are ultimately assessed

- Variance: measure of the deviation from the mean for points in one dimension. Covariance as a measure of how much each of the dimensions vary from the mean with respect to each other.

  The covariance between one dimension and itself is the variance. It is a measure of the variability or spread in a set of data. We use the following formula to compute variance.
  $\mathrm{Var}(X) = \Sigma (X_i - X)^2 / N = \Sigma x_i^2 / N$
  where
  N is the number of scores in a set of scores
  X is the mean of the N scores.
  X_i is the i^{th} raw score in the set of scores
  x_i is the i^{th} deviation score in the set of scores
  $\mathrm{Var}(X)$ is the variance of all the scores in the set

- Outliers: Data points that diverge from the overall pattern and skews the graphs. To identify any outlier.
  - It could have an extreme X &/or Y value compared to other data points.
  - It might be distant from the rest of the data, even without extreme X or Y values.
  - All outliers may not be an influencer – (An influential point is an outlier that greatly affects the slope of the regression line). One way to test the influence of an outlier is to compute the regression equation with and without the outlier.

- Details of Regression and Terms is explained in another document. Please refer to the document

Naseha Sameen | +91 95999 641 69| naseha@sevensolutions.in