Business Research Methods

Course Code – MGMT4013

By

Prof. S.K Shau
Department of Management Sciences
Mahatma Gandhi Central University, Motihari

Content

- **Correlation and Regression**

❑ Introduction About Correlation And Regression

- Correlation quantifies the degree and direction to which two variables are related.
- Correlation does not fit a line through the data points. But simply is computing a correlation coefficient that tells how much one variable tends to change when the other one does. When r is 0.0, there is no relationship.
- When r is positive, there is a trend that one variable goes up as the other one goes up.
- When r is negative, there is a trend that one variable goes up as the other one goes down.
- With correlation, it doesn't have to think about cause and effect. It doesn't matter which of the two variables is call dependent and which is call independent, if the two variables swapped the degree of correlation coefficient will be the same.

- Linear regression finds the best line that predicts dependent variable from independent variable.
- The decision of which variable calls dependent and which calls independent is an important matter in regression, as it'll get a different best-fit line if you swap the two.
- The line that best predicts independent variable from dependent variable is not the same as the line that predicts dependent variable from independent variable in spite of both those lines have the same value for R2. Linear regression quantifies goodness of fit with R2, if the same data put into correlation matrix the square of r degree from correlation will equal R2 degree from regression.
- The sign (+, -) of the regression coefficient indicates the direction of the effect of independent variable(s) into dependent variable, where the degree of the regression coefficient indicates the effect of the each independent variable into dependent variable.

❑ Assumptions of parametric and non parametric Statistics

- Parametric statistics are the most common type of inferential statistics, which are calculated with the purpose of generalizing the findings of a sample to the population it represents.

- Parametric tests make assumptions about the parameters of a population, whereas nonparametric tests do not include such assumptions or include fewer. parametric tests assume that the sample has been randomly selected from the population it represents and that the distribution of data in the population has a known underlying distribution.

- Other distributions include the binomial distribution (logistic regression) and the Poisson distribution (Poisson regression), and non-parametric tests are sometimes called "distribution-free" tests.

Parametric statistics require that the data are measured using an interval or ratio scale, whereas nonparametric statistics use data that are measured with a nominal or ordinal scale.
There are three types of commonly used nonparametric correlation coefficients (Spearman R, Kendall Tau, and Gamma coefficients), where parametric correlation coefficients (Pearson).

❑ Test of Significance level

"significant" means important, while in Statistics "significant" means probably true (not due to chance). A research finding may be true without being important. When statisticians say a result is "highly significant" they mean it is very probably true. They do not (necessarily) mean it is highly important.
Significance levels show you how likely a pattern in your data is due to chance. The most common level, used to mean something is good enough to be believed, is "0.95".
This means that the finding has a 95% chance of being true which also means that the finding has a confidence degree 95% of being true. No statistical package will show you "95%" or ".95" to indicate this level. Instead it will show you ".05," meaning that the finding has a five percent (.05) chance of not being true "error", which is the converse of a 95% chance of being true.

❑ Correlation Analysis

- Correlation is a statistical measure that indicates the extent to which two or more variables fluctuate together.

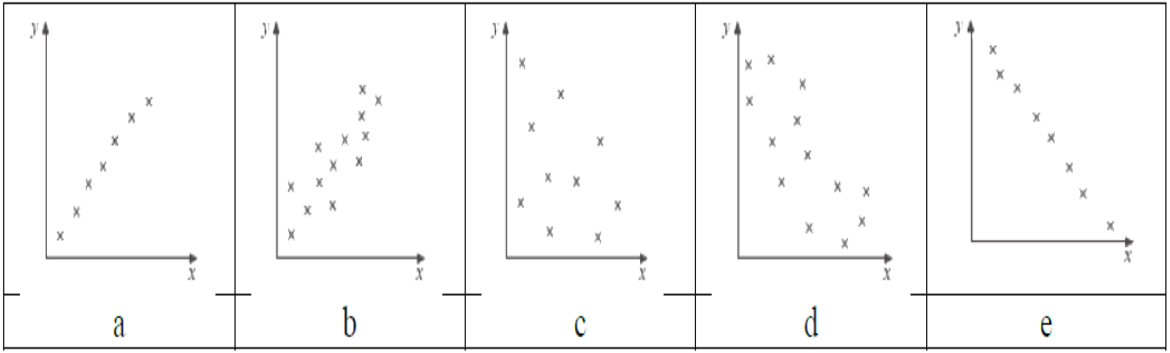
- A positive correlation indicates the extent to which those variables increase or decrease in parallel; a negative correlation indicates the extent to which one variable increases as the other decreases.

- Correlation is a statistical technique that can show whether and how strongly pairs of variables are related. Although this correlation is fairly obvious your data may contain unsuspected correlations. You may also suspect there are correlations, but don't know which are the strongest. An intelligent correlation analysis can lead to a greater understanding of your data.

Correlation Analysis cont…

- Correlation is **Positive** or direct when the values **increase** together, and

- Correlation is **Negative** when one value **decreases** as the other increases, and so called inverse or contrary correlation.



❑ Correlation Analysis cont…

If the points plotted were all on a straight line we would have perfect correlation, but it could be positive or negative as shown in the diagrams above,

a. Strong positive correlation between x and y. The points lie close to a straight line with y increasing as x increases.

b. Weak, positive correlation between x and y. The trend shown is that y increases as x increases but the points are not close to a straight line .

c. No correlation between x and y; the points are distributed randomly on the graph.

d. Weak, negative correlation between x and y. The trend shown is that y decreases as x increases but the points do not lie close to a straight line

e. Strong, negative correlation. The points lie close to a straight line, with y decreasing as x increases

Correlation can have a value:

1. 1 is a perfect positive correlation.
2. 0 is no correlation (the values don't seem linked at all).
3. -1 is a perfect negative correlation

✓ **Assumption of Correlation**

- The variables are assumed to be independent, assume that they have been randomly selected from the population; the two variables are normal distribution; association of data is homoscedastic (homogeneous), homoscedastic data have the same standard deviation in different groups where data are heteroscedastic have different standard deviations in different groups and assumes that the relationship between the two variables is linear. The correlation coefficient is not satisfactory and difficult to interpret the associations between the variables in case if data have outliers.

- Descriptive statistics that express the degree of relation between two variables are called correlation coefficients.

- Correlation used to examine the presence of a linear relationship between two variables providing certain assumptions about the data are satisfied. The results of the analysis, however, need to be interpreted with care, particularly when looking for a causal relationship.

- ✓ **Bivariate Correlation**

Bivariate correlation is a measure of the relationship between the two variables; it measures the strength and direction of their relationship, the strength can range from absolute value 1 to 0.
The stronger the relationship, the closer the value is to 1. Direction of The relationship can be positive (direct) or negative (inverse or contrary); correlation generally describes the effect that two or more phenomena occur together and therefore they are linked.

The Pearson correlation coefficient is given by the following equation:

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

Where , $\bar{x}$ is the mean of variable x values, and
$\bar{y}$ is the mean of variable y values.

✓ Partial Correlation

The Partial Correlations procedure computes partial correlation coefficients that describe the linear relationship between two variables while controlling for the effects of one or more additional variables. Correlations are measures of linear association.

Two variables can be perfectly related, but if the relationship is not linear, a correlation coefficient is not an appropriate statistic for measuring their association.

Partial correlation is the correlation between two variables after removing the effect of one or more additional variables. Suppose we want to find the correlation between and controlling by W

This is called the partial correlation and its symbol is $r_{YX.W}$

$$r_{YX.W} = \frac{r_{XY} - r_{XW} r_{YW}}{\sqrt{(1 - r_{XW}^2)(1 - r_{YW}^2)}}$$

$r_{YX.W}$ or the correlation between *x* and *y* controlling by *W*.

- ✓ **Correlation Coefficients Pearson and Spearman**

Correlation is a Bivariate analysis that measures the strengths of association between two variables. In statistics, the value of the correlation coefficient varies between +1 and -1. When the value of the correlation coefficient lies around ± 1, then it is said to be a perfect degree of association between the two variables. As the correlation coefficient value goes towards 0, the relationship between the two variables will be weaker. Usually, in statistics, we measure three types of correlations: Pearson correlation, Kendall rank correlation and Spearman correlation.

Pearson r correlation: Pearson correlation is widely used in statistics to measure the degree of the relationship between linear related variables. For example, in the stock market, if we want to measure how two commodities are related to each other, Pearson correlation is used to measure the degree of relationship between the two commodities. The following formula is used to calculate the Pearson correlation coefficient $\boldsymbol{r}$.

✓ Correlation Coefficients Pearson and Spearman cont…

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

Spearman rank correlation: Spearman rank correlation is a non-parametric test that is used to measure the degree of association between two variables. It was developed by Spearman, thus it is called the Spearman rank correlation. Spearman rank correlation test does not assume any assumptions about the distribution of the data and is the appropriate correlation analysis when the variables are measured on a scale that is at least ordinal.

The following formula is used to calculate the Spearman rank correlation coefficient:

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2 - 1)}$$

Where:

ρ = Spearman rank correlation coefficient
di= the difference between the ranks of corresponding values Xi and Yi
n= number of value in each data set.

- ✓ **Correlation Coefficients Pearson and Spearman cont…**

The Spearman correlation coefficient,ρ, can take values from +1 to -1.
A ρ of +1 indicates a perfect association of ranks, a ρ of zero indicates no association between ranks and a ρ of -1 indicates a perfect negative association of ranks. The closer ρ to zero, the weaker the association between the ranks.

❑ Regression Analysis

- Regression analysis is one of the most commonly used statistical techniques in social and behavioral sciences as well as in physical sciences which involves identifying and evaluating the relationship between a dependent variable and one or more independent variables, which are also called predictor or explanatory variables.

- It is particularly useful for assess and adjusting for confounding. Model of the relationship is hypothesized and estimates of the parameter values are used to develop an estimated regression equation.

- Linear regression explores relationships that can be readily described by straight lines or their generalization to many dimensions.

- A surprisingly large number of problems can be solved by linear regression, and even more by means of transformation of the original variables that result in linear relationships among the transformed variables.

❑ Regression Analysis cont…

When there is a single continuous dependent variable and a single independent variable, the analysis is called a **simple linear regression analysis**. This analysis assumes that there is a linear association between the two variables. **Multiple regression** is to learn more about the relationship between several independent or predictor variables and a dependent or criterion variable.

Independent variables are characteristics that can be measured directly; these variables are also called predictor or explanatory variables used to predict or to explain the behavior of the dependent variable.

Dependent variable is a characteristic whose value depends on the values of independent variables.

Reliability and Validity:

- Does the model make intuitive sense? Is the model easy to understand and interpret?
- Are all coefficients statistically significant? (p-values less than .05)
- Are the signs associated with the coefficients as expected?
- Does the model predict values that are reasonably close to the actual values?
- Is the model sufficiently sound? (High R-square, low standard error, etc.)

✓ Objectives of Regression Analysis

- Regression analysis used to explain variability in dependent variable by means of one or more of independent or control variables and to analyze relationships among variables to answer; the question of how much dependent variable changes with changes in each of the independent's variables, and to forecast or predict the value of dependent variable based on the values of the independent's variables.

- The primary objective of regression is to develop a linear relationship between a response variable and explanatory variables for the purposes of prediction, assumes that a functional linear relationship exists, and alternative approaches (functional regression) are superior.

- ✓ **Assumption of Regression Analysis**

The regression model is based on the following assumptions.

- The relationship between independent variable and dependent is linear.
- The expected value of the error term is zero .
- The variance of the error term is constant for all the values of the independent variable, the assumption of homoscedasticity.
- There is no autocorrelation.
- The independent variable is uncorrelated with the error term.
- The error term is normally distributed.
- On an average difference between the observed value (yi) and the predicted value (yi) is zero.
- On an average the estimated values of errors and values of independent variables are not related to each other.
- The squared differences between the observed value and the predicted value are similar.
- There is some variation in independent variable. If there are more than one variable in the equation, then two variables should not be perfectly correlated.

✓ **Assumption of Regression Analysis cont…**

Intercept or Constant

- Intercept is the point at which the regression intercepts y-axis.
- Intercept provides a measure about the mean of dependent variable when slope(s) are zero.
- If slope(s) are not zero then intercept is equal to the mean of dependent variable minus slope × mean of independent variable.

Slope

- Change is dependent variable as we change independent variable.
- Zero Slope means that independent variable does not have any influence on dependent variable.
- For a linear model, slope is not equal to elasticity. That is because; elasticity is percent change in dependent variable, as a result one percent change in independent variable.

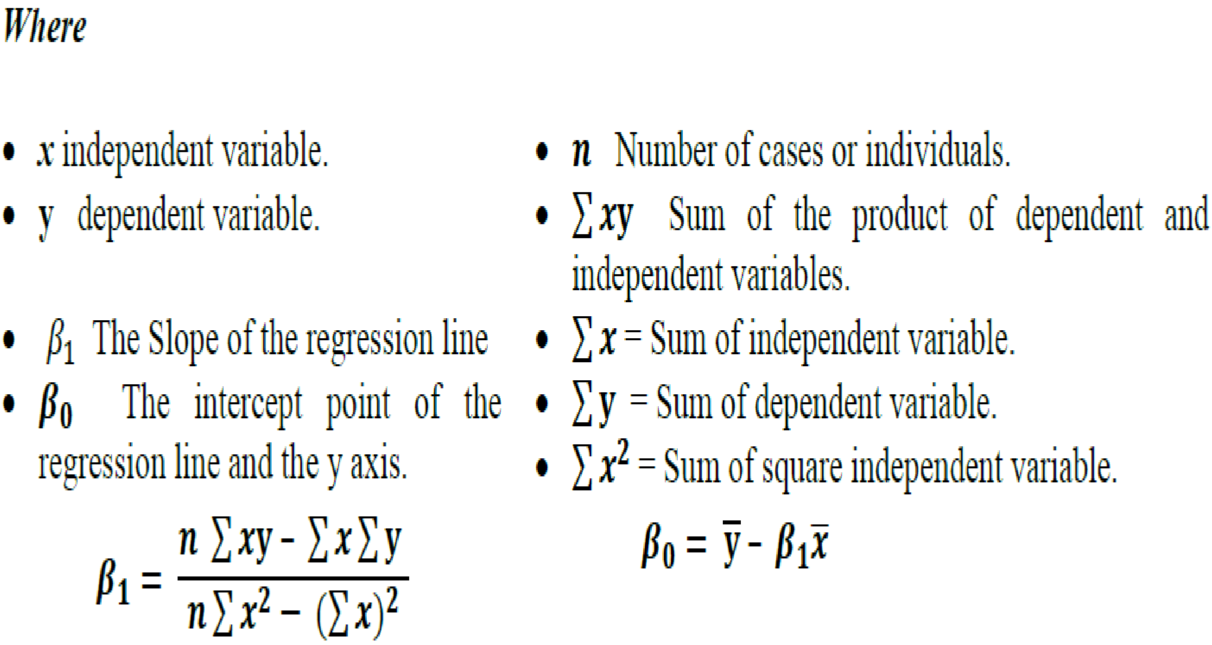
✓ Simple Regression Model

Simple linear regression is a statistical method that allows us to summarize and study relationships between two continuous (quantitative) variables. In a cause and effect relationship, the **independent variable** is the cause, and the **dependent variable** is the effect.
Least squares linear regression is a method for predicting the value of a dependent variable **y**, based on the value of an independent variable ***x***.

- One variable, denoted (***x***), is regarded as the **predictor**, **explanatory**, or **independent** variable.
- The other variable, denoted (***y***), is regarded as the **response**, **outcome**, or **dependent** variable.

Mathematically, the regression model is represented by the following equation: $\mathbf{y} = \beta 0 \pm \beta 1 \, \boldsymbol{x} 1 \pm \varepsilon 1$

✓ Simple Regression Model cont…



Where

- x independent variable.
- y dependent variable.
- β_1 The Slope of the regression line
- β_0 The intercept point of the regression line and the y axis.

$$\beta_1 = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - (\sum x)^2}$$

- n Number of cases or individuals.
- $\sum xy$ Sum of the product of dependent and independent variables.
- $\sum x$ = Sum of independent variable.
- $\sum y$ = Sum of dependent variable.
- $\sum x^2$ = Sum of square independent variable.

$$\beta_0 = \bar{y} - \beta_1\bar{x}$$

- ✓ **Multiple Regressions Model**

Multiple regression is an extension of simple linear regression. It is used when we want to predict the value of a dependent variable (target or criterion variable) based on the value of two or more independent variables (predictor or explanatory variables).

Multiple regression allows you to determine the overall fit (variance explained) of the model and the relative contribution of each of the predictors to the total variance explained. For example, you might want to know how much of the variation in exam performance can be explained by revision time and lecture attendance "as a whole", but also the "relative contribution" of each independent variable in explaining the variance.

Mathematically, the multiple regression model is represented by the following equation:

$$Y = \beta_0 \pm \beta_i X_i \dots\dots\dots\dots \pm \beta_n X_n \pm u$$

✓ Multiple Regressions Model cont…

Where:

- X_i *to* X_n Represent independent variables.
- $\mathbf{Y}$ Dependent variable.
- β_1 The regression coefficient of variable x_1
- β_2 The regression coefficient of variable x_2
- β_0 The intercept point of the regression line and the y axis.

By using method of deviation

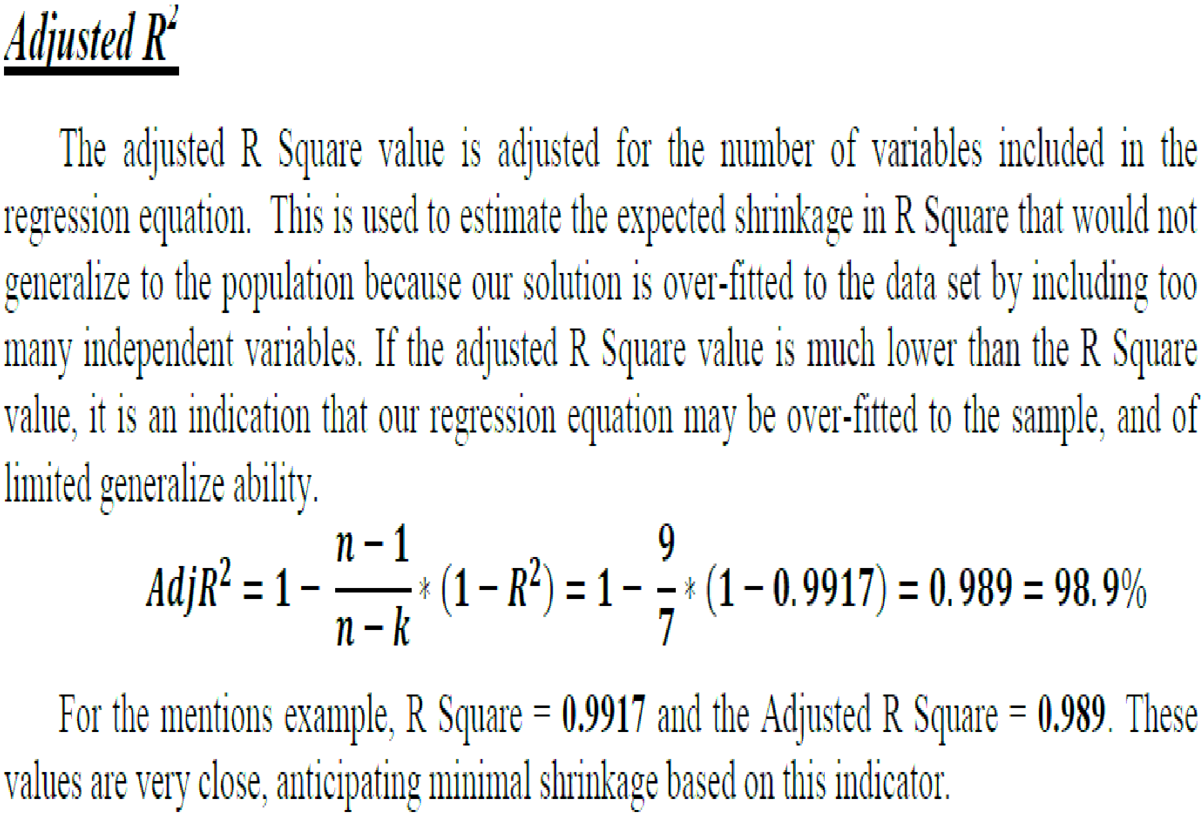
- $\overline{\mathbf{y}}$ The mean of dependent variable values.
- $\overline{X_1}$ The mean of X_1 independent variable values.
- $\overline{X_2}$ The mean of X_2 independent variable values.
- $\sum x_1\mathbf{y} = \sum(x_1 * y)$
- $\sum x_2\mathbf{y} = \sum(x_2 * y)$
- $\sum x_1x_2 = \sum(x_1 * x_2)$
- $(\sum x_2^2) =$ *Sum of square of* x_2

- $\sum \mathbf{y} = \sum(Y - \overline{\mathbf{Y}})$
- $\sum x_1 = \sum(X_1 - \overline{X_1})$
- $\sum x_2 = \sum(X_2 - \overline{X_2})$
- $(\sum x_1^2) =$ *Sum of square of* x_1
- $(\sum x_2^2) =$ *Sum of square of* x_2
- $(\sum x_1^2) =$ *Sum of square of* x_1

$$\beta_1 = \frac{(\sum x_1 y)(\sum x_2^2) - (\sum x_2\, y)(\sum x_1x_2)}{(\sum x_1^2)(\sum x_2^2) - (\sum x_1x_2)^2}$$

$$\beta_2 = \frac{(\sum x_2 y)(\sum x_1^2) - (\sum x_1\, y)(\sum x_1x_2)}{(\sum x_1^2)(\sum x_2^2) - (\sum x_1x_2)^2}$$

$$\beta_0 = \overline{\mathbf{y}} - \beta_1\overline{x_1} - \beta_2\overline{x_2}$$



Adjusted R^2

The adjusted R Square value is adjusted for the number of variables included in the regression equation. This is used to estimate the expected shrinkage in R Square that would not generalize to the population because our solution is over-fitted to the data set by including too many independent variables. If the adjusted R Square value is much lower than the R Square value, it is an indication that our regression equation may be over-fitted to the sample, and of limited generalize ability.

$$AdjR^2 = 1 - \frac{n-1}{n-k} * (1 - R^2) = 1 - \frac{9}{7} * (1 - 0.9917) = 0.989 = 98.9\%$$

For the mentions example, R Square = **0.9917** and the Adjusted R Square = **0.989**. These values are very close, anticipating minimal shrinkage based on this indicator.

❑ Correlation Vs. Regression

Basis for Comparison	Correlation	Regression
Meaning	Correlation is a statistical measure which determines co-relationship or association of two variables.	Regression describes how an independent variable is numerically related to the dependent variable.
Usage	To represent linear relationship between two variables.	To fit a best line and estimate one variable on the basis of another variable.
Dependent and Independent	No difference	Both variables are different.
Variables Indicates	Correlation coefficient indicates the extent to which two variables move together.	Regression indicates the impact of a unit change in the known variable (x) on the estimated variable (y).
Objective	To find a numerical value expressing the relationship between variables.	To estimate values of random variable on the basis of the values of fixed variable.

❑ Sources

- https://stats.libretexts.org/Bookshelves/Introductory_Statistics/Book%3A_Introductory_Statistics_(Shafer_and_Zhang)/10%3A_Correlation_and_Regression

- Whitley E, Ball J. Statistics review 1: Presenting and summarising data. Crit Care. 2002;6:66–71. doi: 10.1186/cc1455.

- Kirkwood BR, Sterne JAC. Essential Medical Statistics. 2. Oxford: Blackwell Science; 2003.

- Whitley E, Ball J. Statistics review 2: Samples and populations. Crit Care. 2002;6:143–148. doi: 10.1186/cc1473.

- Bland M. An Introduction to Medical Statistics. 3. Oxford: Oxford University Press; 2001.

- Bland M, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet. 1986;i:307–310.

THANK
YOU