Unit V

- Regression analysis - correlation and regression analysis comparison - multiple regression analysis- reliability of estimates – coefficient of multiple determinations.

Regression analysis

- Regression analysis- is a set of statistical methods used for the estimation of relationships between a dependent variable and one or more independent variables.
- Regression analysis includes several variations, such as linear, multiple linear and nonlinear.
- The most common models are simple linear and multiple linear. Nonlinear regression analysis is commonly used for more complicated data sets in which the dependent and independent variables show a nonlinear relationship.

Regression

Regression

- A statistical tool used to find the nature of relationship
- Estimates the value of a dependent variable with the help of an independent variable
- Types:
  - Regression of y on x is, y = a + bx (a and b are constants)

$$\sum Y = na + b\sum X$$

$$\sum XY = a\sum X + b\sum X^2$$

  - Regression of x on y is, x = a + by (a and b are constants)

$$\sum X = na + b\sum Y$$

$$\sum XY = a\sum Y + b\sum Y^2$$

Correlation and Regression analysis comparison

- Correlation - is when a change to one variable is then followed by a change in another variable, whether it be direct or indirect.
- Variables are considered “uncorrelated” when a change in one does not affect the other. In short, it measures the relationship between two variables.
- Regression analysis - is how one variable affects another, or changes in a variable that trigger changes in another, essentially cause and effect. It implies that the outcome is dependent on one or more variables.

Correlation and Regression analysis comparison

- Correlation quantifies the direction and strength of the relationship between two numeric variables, X and Y, and always lies between -1.0 and 1.0.
- Simple linear regression relates X to Y through an equation of the form Y = a + bX
- **Key similarities**
- Both quantify the direction and strength of the relationship between two numeric variables.
- When the correlation (r) is negative, the regression slope (b) will be negative.
- When the correlation is positive, the regression slope will be positive.
- The correlation squared (r2 or R2) has special meaning in simple linear regression. It represents the proportion of variation in Y explained by X.

Correlation and Regression analysis comparison

- **Key differences**
- Regression attempts to establish how X causes Y to change and the results of the analysis will change if X and Y are swapped. With correlation, the X and Y variables are interchangeable.
- Regression assumes X is fixed with no error, such as a dose amount or temperature setting. With correlation, X and Y are typically both random variables*, such as height and weight or blood pressure and heart rate.
- Correlation is a single statistic, whereas regression produces an entire equation.

Multiple regression analysis

- Multiple regression analysis- generally explains the relationship between multiple independent or predictor variables and one dependent or criterion variable.
- A dependent variable is modeled as a function of several independent variables with corresponding coefficients, along with the constant term.
- Multiple regression requires two or more predictor variables, and this is why it is called multiple regression.

Reliability of estimates-

- Reliability of estimates- The problem of determining the accuracy of estimates from the multiple regression is basically the same as for estimates from a simple regression equation.
- The measure of reliability is an average of deviations of the actual value of non-dependent variable from the estimate from the regression equation or , in other words, the standard error of estimate.

$$S_{1.23} = \sqrt{\frac{\sum (X_1 - Xl_{ast})^2}{N-3}}$$

$S_{1.23}$ represents standard error of estimate of X1 on X2 and X3.
X_{last} indicates the estimated value of x1 as calculated from the regression equations

correlation coefficient

- In terms of the correlation coefficient r_{12},r_{13} and r_{23} , the standard error of estimate can also be computed from the result:

$$S_{1.23} = S1 \frac{\sqrt{1 - r^2_{12} - r^2_{13} - r^2_{23} + 2r_{12}r_{13}r_{23}}}{\sqrt{1 - r^2 23}}$$

The standard error measures the closeness of estimates derived from the regression equation to actual observed values.

Coefficient of multiple determinations

- The coefficient of multiple determination (R^2) measures the proportion of variation in the dependent variable that can be predicted from the set of independent variables in a multiple regression equation.
- When the regression equation fits the data well, R^2 will be large (i.e., close to 1); and vice versa.
- In statistics, the **coefficient of determination**, denoted R^2 or r^2 and pronounced "R squared", is the proportion of the variance in the dependent variable that is predictable from the independent variable(s).

Coefficient of multiple determinations

- The coefficient of determination can also be found with the following formula: $R^2 = MSS/TSS = (TSS - RSS)/TSS$,
- where *MSS* is the model sum of squares (also known as *ESS*, or explained sum of squares), which is the sum of the squares of the prediction from the linear regression minus the mean for that variable; *TSS* is the total sum of squares associated with the outcome variable, which is the sum of the squares of the measurements minus their mean; and *RSS* is the residual sum of squares, which is the sum of the squares of the measurements minus the prediction from the linear regression.

References

- https://www.slideshare.net/21_venkat/multiple-regression-17406485
- https://www.slideshare.net/AvjinderSingh/multiple-linear-regression-61224783