REGRESSION ANALYSIS

Dr. Sabiha Khan

1.1 REGRESSION ANALYSIS

- The statistical method that helps to formulate an algebraic relation between two or more variables in the form of an equation to estimate the value of a continuous random variable, given the value of another variable.

- **Dependent Variable** (response variable): the variable whose value is estimated using the algebraic equation.
- **Independent Variable** (predictor variable): the variable whose value is used as the basis for the estimate.
- **Linear regression equation**: it is used for expressing a dependent variable in terms of independent variable.

The basic differences between correlation and regression analysis are summarized as follows:

1. The process of developing an algebraic equation between two variables from sample data and predicting the value of one variable, given the value of the other variable is referred to as regression analysis, while measuring the strength or degree of the relationship between two variables is referred as correlation analysis. The sign of correlation coefficient indicates the nature (direct or inverse) of relationship between two variables, while the absolute value of correlation coefficient indicates the extent of relationship.
2. Correlation analysis assures the existence of an association between two variables x and y but not that they have a cause-and-effect relationship. Regression analysis, in contrast to correlation, is used to indicate the cause-and-effect relationship between x and y, that is, a change in the value of independent variable x *causes* a corresponding change (*effect*) in the value of dependent variable y if all other factors that affect y remain unchanged.
3. In regression analysis there is only one dependent variable, while in correlation analysis both variables have to be independent.
4. The coefficient of determination r^2 indicates the proportion of variance in the dependent variable that is explained statistically by the independent variable. The value of r^2 is a sample value and is subject to sampling error. Moreover, the value of r^2 may be high, but the assumption of a linear regression may be incorrect because it may represent a portion of the relationship that actually is in the form of a curve.

1.2 Advantages of regression analysis:

14.2 ADVANTAGES OF REGRESSION ANALYSIS

The following are some important advantages of regression analysis:

1. Regression analysis helps in developing a regression equation by which the value of a dependent variable can be estimated given a value of an independent variable.
2. Regression analysis helps to determine standard error of estimate to measure the variability or spread of values of a dependent variable with respect to the regression line. Smaller the variance and error of estimate, the closer the pair of values (x, y) fall about the regression line and better the line fits the data, that is, a good estimate can be made of the value of variable y. When all the points fall on the line, the standard error of estimate equals zero.
3. When the sample size is large ($df \geq 29$), the interval estimation for predicting the value of a dependent variable based on standard error of estimate is considered to be acceptable by changing the values of either x or y. The magnitude of r^2 remains the same regardless of the values of the two variables.

1.3 Types of regression models:

1.3.1.Simple and multiple regression models:

Simple regression model: If a regression model characterizes the relationship between a dependent y and only one independent variable x, then such a regression model is called a Simple regression model.

Multiple regression model: If more than one independent variables are associated with a dependent variable, then such a regression model is called a multiple regression model.

1.3.2 LINEAR AND NONLINEAR REGRESSION MODELS:

LINEAR REGRESSION MODELS: If the value of a dependent variable y in a regression model tends to increase in direct proportion to an increase in the values of independent variables, then such a regression model is called a linear regression model.

NONLINEAR REGRESSION MODELS: Nonlinear regression is a form of regression analysis in which data is fit to a model and then expressed as a mathematical function. Nonlinear regression relates two variables and must generate a line (typically a curve) as if every value of Y was a random variable.

1.4 Assumptions for a simple linear regression model:

To make valid statistical inference using regression analysis, we start with certain assumptions about the bivariate population from which a sample of paired observations is drawn and the manner in which observations are generated. Certain assumptions that form the basis for application of simple linear regression models are as under:

1. The relationship between the dependent variable y and independent variable x exists and is linear. The average relationship between x and y can be described by a simple linear regression equation $y = a + bx + \epsilon$, where ϵ is the deviation of a particular value of y from its expected value for a given value of x.
2. For every value of the independent variable x, there is an expected (or mean) value of the dependent variable y and this value varies with the independent variable according to a linear equation.
3. The dependent variable y is a continuous random variable, whereas values of the independent variable x are fixed values and are not random.
4. The sampling error, e, associated with the expected value of the dependent variable y is assumed to be an independent random variable distributed normally about the regression line.
5. The standard deviation and variance of expected values of the dependent variable about the regression line are constant for all values of the independent variable within the range of the sample data.
6. The value of the dependent variable cannot be estimated for a value of an independent variable lying outside the range of values in the sample data.

1.5 Parameters of simple linear regression model:

- The devise used for estimating the values of one variable from the value of the other consists of a line through the points drawn in such manner as to represent the average relationship between the two variables, such a line is called ***line of regression.***
- The two variables x and y which are correlated can be expressed in terms of each other in the form of straight line equations called ***regression equations.***

Cont…

- Such lines should be able to provide the best fit of sample data to the population data.
- The algebric expression of regression lines is written as:
- <u>The regression equation of y on x:</u>

Y= a + bx, is used for estimating the value of y for given values of x.

- <u>The regression equation of x on y:</u>

x = c + dy, is used for estimating the value of x for given values of y.

Cont…

Remarks:

- When variables x and y are correlated perfectly (eitther positive or negative) these lines coincide, only with one line.
- Higher the degree of correlation, nearer the two regression lines are to the each other.
- Lesser the degree of correlation, more the two regression lines are away from each other. That is when r = 0, the two lines are at right angle to each other.
- Two linear regression lines intersect each other at the point of the average values of variables x and y.

1.5.1 Regression coefficient Method:

- Regression coefficients are estimates of the unknown population parameters and describe the relationship between a predictor variable and the response. In linear regression, coefficients are the values that multiply the predictor values.
- Suppose you have the following regression equation:

$$y = 3X + 5.$$

In this equation, +3 is the coefficient,

X is the predictor, and

+5 is the constant.

Cont…

- The sign of each coefficient indicates the direction of the relationship between a predictor variable and the response variable.
- A **positive** sign indicates that as the predictor variable increases, the response variable also increases.
- A **negative** sign indicates that as the predictor variable increases, the response variable decreases.
- The coefficient value represents the mean change in the response given a one unit change in the predictor.
- For example, if a coefficient is +3, the mean response value increases by 3 for every one unit change in the predictor.

1.5.2. Least Squares Method

- The “least squares" method is a form of mathematical regression analysis used to determine the line to best fit for a set of data, providing a visual demonstration of the relationship between the data points. Each point of data represents the relationship between a known independent variable and an unknown dependent variable.
- It creates a straight line that minimizes the sum of the squares of the errors that are generated by the results of the associated equations, such as the squared residuals resulting from differences in the observed value, and the value anticipated, based on that model.

cont….

- This least square method of regression analysis begins with a set of data points to be plotted on an x- and y-axis graph. An analyst using the least squares method will generate a line of best fit that explains the potential relationship between independent and dependent variables.
- In regression analysis, dependent variables are illustrated on the vertical y-axis, while independent variables are illustrated on the horizontal x-axis. These designations will form the equation for the line of best fit, which is determined from the least squares method.

THANK YOU

Disclaimer: The content displayed in the PPT has been taken from variety of different websites and book sources. This study material has been created for the academic benefits of the students alone and I do not seek any personal advantage out of it.