1. **Simple Linear Regression**

Regression analysis is a tool for building mathematical and statistical models that characterize relationship between a dependent variable which must be a numerical variable and one or more independent variable.

A regression model that involves one independent variable is called simple regression

1. Simple linear regression involves finding a linear relationship between one independent variable

called X, and one dependent variable Y.

The first thing to do is to verify if there is linear relationship between the two variables. We suggest that you create a scatter chart that can show the relationship between variables visually.

Example, in the case of determining home market value. We clearly understand home market is related to its size. We might want to investigate the relationship between the market value and the size of the home. We will represent the number of square feet as an independent variable labeled by X, and the market value as a dependent variable Y.

If there is a linear relationship between the two variables, we will see a higher market values that are associated with larger house sizes.

1. The idea behind simple linear regression is to expression the relationship between dependent variable and independent variables by a simple linear equation such as:

Market value = a + b\* square feet. Where b is the slope and a is the intercept.

**4.**

a. Independent variable may be: numerical, ratio and categorical

Dependent must be numerical

1. Simple linear regression involves only two variables, a dependent variable and an independent variable

c. there is a restriction for categorical variable, for that we need to transform the categorical variables into dummies data.

**5.** In regression analysis, we assume that the values of the dependent variable Y in the sample data are drawn form some un known population for each independent variable X. We are also assuming that a linear relationship exists, the expected value of Y is b0+b1 for each value of X. the coefficients b0 and b1 are population parameters that represent the intercept and slope, respectively, of the population from which a sample of observations is taken. The intercept is the mean value of Y when X=0, and the slope is the change in the mean value of Y as X changes by one unit.

The best-fitting line should minimize some measure of these errors.

**6.**

1. Pandas, NumPy, SciPy, Seaborn, ggplot Scikit-learn
2. Linear Regression, predict, fit, intercept

**7.**

a. when we run regression analysis, there are some key elements to look for.

In excel, the multiple R is equivalent to the coefficient of correlation which help to determine the relationship between the two variables. If the multiple R is greater than 0 indicates positive correlation.

That means, as the independent variable increases, the dependent variable also increases. If the multiple R is less than 0 indicates that as the dependent variable increase, the independent decreases. A Multiple R equal to zero means there is no correlation at all. We can evaluate a regression analysis by looking at R-squared which is called the coefficient of determination. It indicates how well the regression line fits the data. The R-squared varies between 0 and 1. A value of 1 indicates a perfect fit means all the data points lie on the regression line, whereas a value of 0 indicates no relationship exists.

Another way to evaluate the regression analysis, is the adjusted R square which is a statistic that modifies the value of R square by incorporating the sample size and the number of explanatory variables in the model.

The adjusted R square is useful when comparing two different models.

The last thing to look is the standard error of the estimate which is the variability of the observed Y values from the predicted values. if the data are clustered close to the regression line, the standard error will be small.

**Multiple Linear Regression**

Regression analysis is a tool for building mathematical and statistical models that characterize relationship between a dependent variable which must be a numerical variable and one or more independent variable. A regression model that involves two or more independent variable is called multiple regression.

**2.** Multiple linear regression involves finding a linear relationship between two or more independent

variable called X, and one dependent variable Y.

The first thing to do is to verify if there is linear relationship between the variables. We suggest that you create a scatter chart that can show the relationship between the variables visually.

Example, in the case of predicting the student performance as a function of serval characteristics.

Let us assume we want to predict the graduation rate as function of acceptance rate, median SAT, expenditures/student, and the percent in the top 10% in their high school. We might want to investigate the relationship between graduation rate, median SAT, expenditures/student, and top 10%.

We will represent top 10%, median SAT, expenditures/student as an independent variable labeled by X, and graduation rate as a dependent variable Y.

**3.** A linear regression with more than one independent variable is called multiple linear regression model. A multiple regression model has the form:

Y =b0+b1\*x1+b2\*x2+b3\*x3+…+bk\*xk+e where Y is the dependent variable,

X1, x2,…xk are independent or explanatory variables,

b0 is the intercept,

b1, b2,…. bk are regression coefficients for the independent variables.

**4.**

1. List type of data

Independent variable may be: numerical, ratio and categorical

Dependent must be numerical

1. multiple linear regression involves two or more variables independent variable

**5.** In regression analysis, we assume that the values of the dependent variable Y in the sample data are drawn from some unknown population for each independent variable X. We are also assuming that a linear relationship exists, the expected value of Y is b0+b1 for each value of X. the coefficients b0 and b1are population parameters that represent the intercept and slope, respectively, of the population from which a sample of observations is taken. The intercept is the mean value of Y when X=0, and the slope is the change in the mean value of Y as X changes by one unit.

The results from the regression tool are in the same format as we saw form simple linear regression. However, some key differences exist. Multiple R and R square are called the correlation coefficient and the coefficient of multiple determination respectively. In the context of multiple regression, they indicate the strength of association between the dependent variable and independent variables.

**Reference**

reference is the text book use in class

<https://newonlinecourses.science.psu.edu/stat501/node/285/>