LECTURE 5 Linear Modelling

Objectives:

- Introduction of Linear Modelling
- Regression Models and assumptions of the model.
- Fitting a Regression Model
- Point estimator of least square parameter.
- Residuals and properties of Regression Models.

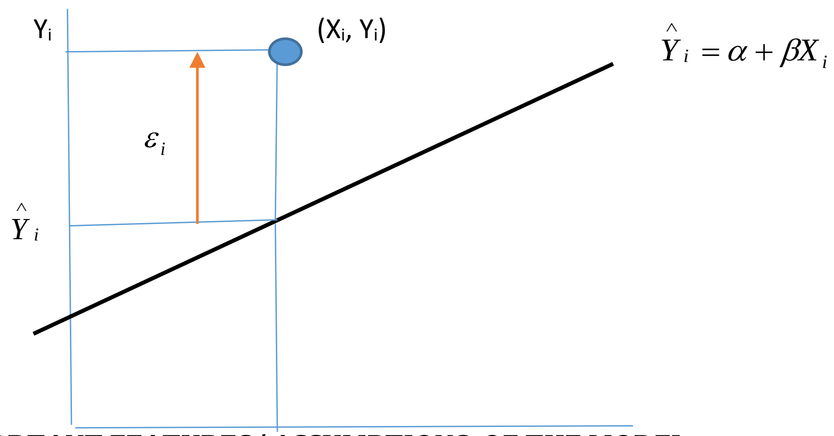
SIMPLE LINEAR LEAST SQUARE REGRESSION:

- It is an important techniques because in real life two variables can have a linear relationship.
- Linear model which derives its foundational concepts from Linear Least Square Regression encompasses and extends include the relationships for qualitative explanatory variable, polynomial and non linear functions.
- Linear regression provides mathematical foundations for newer techniques like weighted least square regressions, robust regressions, nonparametric regressions and generalized linear models.
- Simple Linear Regression is a techniques that explores the relationship between one explanatory variable (independent variable) and response variable (dependent variable).
- Simple Linear Regression can be used for prediction.
- For example Cholesterol and age. Here age is the independent variable (regressor variable) and cholesterol is the dependent variable. The age is the variable that can be controlled wheras cholesterol is not a controlled variable.. The age is a non random variable wheras cholesterol is a random variable.
- The scatter plot evaluates the relationships between two variables as it will show if the relationship is linear, nonlinear specifically polynomial or quadratic etc. The least square line helps determine if the relationship is linear or not.
- Taking a case study of the cholesterol increases with age and be able to predict the cholesterol of a person given the age.
- Let Y denote the Cholesterol of a person and X denote the Age of a person. The equation that can represent this relationship is $Y = A + BX$
- The line cannot pass through all the points even if the relationship is strong. To represent this inaccuracy Residual E has to be included for each data point.
- $Y = A + BX + E$ is the equation for the sample data where Y is the response variable, X regressor variable ,A the y intercept, B is the slope and E is the random error (Residual).
- If we construct the model for the population which is our final motive the regression model is as follows: $Y = \alpha + \beta X + \varepsilon$ where alpha is the y intercept for the population data and beta is the slope for the population data.
- Practically we never have the population data therefore we try to estimate alpha by A and beta by B and ε is estimated by E.

REGRESSION MODEL

Assumptions on the Model 1

$Y = \alpha + \beta X + \varepsilon$ This the model equation for the population



IMPORTANT FEATURES/ ASSUMPTIONS OF THE MODEL

1) Linearity and Constant Variance: ε_i is a random variable with zero mean and variance σ^2 (unknown) ie $E(\varepsilon_i) = 0$ and $Var(\varepsilon_i) = \sigma^2$. Normality and Independence: ε_i is a normally distributed random variable with mean zero and variance σ^2 ie $\varepsilon_i \overset{ind}{\sim} N(0, \sigma^2)$ under the normality conditions they are uncorrelated and also independent.

2) Uncorrelated: ε_i and ε_j are uncorrelated for i≠j. Therefore $Cov(\varepsilon_i, \varepsilon_j) = 0$
3) X is not invariant All Xs will not all be the same.
4) X is deterministic or is measured without error and Independent of the error

Consequence of the assumptions 1 on the model are as follows:

$Y_i = \alpha + \beta X_i + \varepsilon_i \ldots\ldots\ldots\ldots\ldots\ldots\ldots 1$ Here Y and ε_i are random

$E(Y_i) = E(\alpha + \beta X_i + \varepsilon_i) = \alpha + \beta X_i$

$V(Y_i) = V(\alpha + \beta X_i + \varepsilon_i) = V(\varepsilon_i) = \sigma^2$ since $\alpha + \beta X_i$ are not random ie Is deterministic.

Consequence of assumption 3 is as follows:

$\varepsilon_i \overset{ind}{\sim} N(0,\sigma^2)$ is as follows:

Y_i is a component of 2 terms: ε_i (random term which is normally distributed with a constant variance σ^2) and $\alpha+\beta X_i$ (constant term) therefore Y is also normally distributed

$Y_i \overset{ind}{\sim} N(\alpha+\beta X_i,\sigma^2)$ the ith observation is normally distributed with given mean and constant variance.

- The data set that we use should follow these three assumptions. If the data set does not follow the assumptions then we cannot use the Least square model.
- Diagram from notes

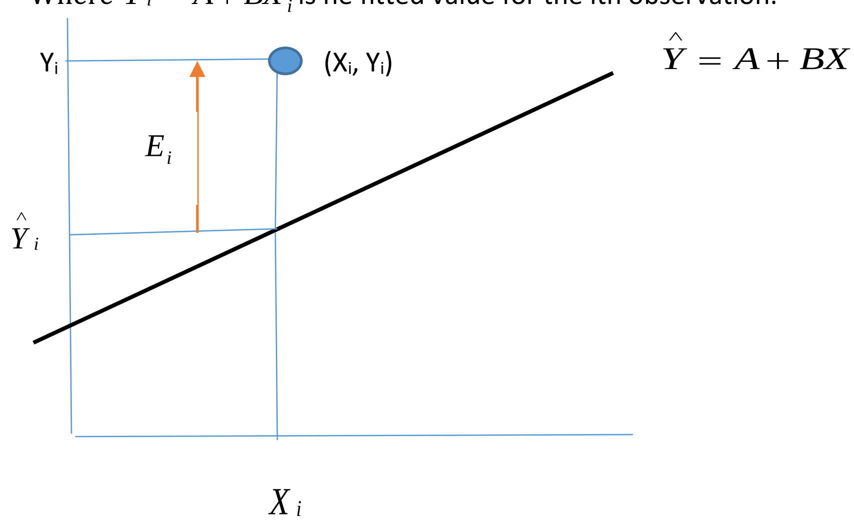
FITTING A REGRESSION MODEL:

- Usually we do not have the entire information about the population therefore we have to use the sample data instead of the population data. We will fit the regression model by the equation
  $Y = A + BX + E$ instead of $Y_i = \alpha + \beta X_i + \varepsilon_i$
- The regression equation for the sample and i-th data observation of the sample:
  $Y = A + BX + E$

  $Y_i = A + BX_i + E_i$....................1

  $Y_i = \hat{Y}_i + E_i$

  Where $\hat{Y}_i = A + BX_i$ is he fitted value for the ith observation.



RESIDUAL FOR THE iTH OBSERVATIONS :

$E_i = Y_i - \hat{Y}_i = Y_i - (A + BX_i)$ i=1.....n

- If the data observation is above the line the residual is positive and if the data observation is below the line then the residual is negative.
- The best fitted line is a line that minimizes the sum of residuals.
- If we add up the residuals the negatives will cancel out the positives and therefore the residuals cancel each other. To capture the residuals we can either add the absolute values or add the squares of the residuals.

Least square estimation of parameters A and B. The simple linear regression model is $Y = A + BX + E$.

Least square estimation of parameters is the estimation of parameter s A and B. A is the y intercept and B is the slope and these estimate α and β for the population.

Fitting a regression model implies the determination of A and B which are the regression coefficients

POINT ESTIMATE OF LEAST SQUARE ESTIMATORS A and B

- The parameters A and B are unknown and have to be determined by the sample data $(x_1,y_1),(x_2,y_2)$..........(x_n,y_n) If the scatterplot shows a linear relationship then the linear model can be fitted to the data.
- There can be more than one fitted line then we need to determine which line is the best fit.
- The line that is fitted by the least square methodology is the one that minimizes the residuals (vertical errors) .
  Graph
- The least square estimation fits the model (determines A and B) such that it minimizes the vertical errors(residuals)
  $\sum_{i=1}^{n} E_i^2$ is minimized This called Sum of squares of residuals .
  $$RSS = S(A,B) = \sum_{i=1}^{n} E_i^2 = \sum (Y_i - \hat{Y}_i)^2 = \sum (Y_i - A - BX_i)^2$$ This equation shows the dependence of Sum of Squares on the parameter A and B.
- If we have all the data points (population) we can find α and β. If the data that we have is a sample then we can estimate them by A and B.
- To find the least squares coefficient the partial derivative with respect to the variables has to be computed and then set to zero.
- $$\frac{\partial S(A,B)}{\partial (A)} = \sum (-1)(2)(Y_i - A - BX_i) = 0$$
  $$\frac{\partial S(A,B)}{\partial (B)} = \sum (-1)(2)X_i(Y_i - A - BX_i) = 0$$

These equations are called Normal equations. These are independent equations. A and B are the solution of these equations.
From the first equation A can be computed as follows:

$$\sum_{i=1}^{n}(Y_i - A - BX_i) = 0$$

$$\sum_{i=1}^{n}Y_i - nA - B\sum_{i=1}^{n}X_i = 0$$

$$nA = \sum_{i=1}^{n}Y_i - B\sum_{i=1}^{n}X_i$$

$$A = \overline{Y} - B\overline{X}$$

Also $\overline{Y} = A + B\overline{X}$ This implies that the Least Square Regression line passes through the mean of the X and Y series.

B can be computed from the second equation and the value of A :

$\sum_{i=1}^{n}X_i(Y_i - A - BX_i) = 0$ Substituting $A = \overline{Y} - B\overline{X}$ we get

$$\sum_{i=1}^{n}X_i(Y_i - \overline{Y} + B\overline{X} - BX_i) = 0$$

$$\sum_{i=1}^{n}X_i(Y_i - \overline{Y}) = B\sum_{i=1}^{n}X_i(X_i - \overline{X})$$

$$B = \frac{\sum_{i=1}^{n}X_i(Y_i - \overline{Y})}{\sum_{i=1}^{n}X_i(X_i - \overline{X})} = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})^2}$$

Since $\sum_{i=1}^{n}\overline{X}(Y_i - \overline{Y}) = \overline{X}\sum_{i=1}^{n}Y_i - \sum_{i=1}^{n}\overline{X}\,\overline{Y}$

$$\sum_{i=1}^{n}\overline{X}(Y_i - \overline{Y}) = \overline{X}n\overline{Y} - n\overline{X}\,\overline{Y} = 0$$

$$\sum_{i=1}^{n}(-\overline{X})(X_i - \overline{X}) = \sum_{i=1}^{n}(-\overline{X}X_i + \overline{X}\,\overline{X}) = -n\overline{X}^2 - n\overline{X}^2 = 0$$

And $\sum_{i=1}^{n}X_i(X_i - \overline{X}) = \sum_{i=1}^{n}X_i^2 - X_i\overline{X} = n^2\overline{X}^2 - n\overline{X}^2$ The denominator can also be written this way

The value of B can be alternatively written as

$$B = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{n^2\overline{X^2} - n\overline{X}^2}$$

In some of the textbook A and B are represented by other variable names like $\hat{\beta}_0$ and $\hat{\beta}_1$

B is the equation $\hat{Y} = A + BX$ Informs us regarding the increase/decrease in response variable with every one unit increase of the explanatory variable.

A is the equation $\hat{Y} = A + BX$ Informs us regarding the value of response variable when the explanatory variable is zero. Sometimes this value is not interpretative.

PROPERTIES OF FITTED REGRESSION MODEL:

1) Sum of the residuals in any regression model that contains intercept A is always zero. $\sum E_i = \sum (Y_i - \hat{Y}) = 0$ Rounding errors will sometimes not allow the value to be exactly zero
2) As a consequence of 1) $\sum Y_i = \sum \hat{Y}_i$ Sum of the observed response values is equal to the sum of the fitted response variable.
3) $\sum X_i E_i = 0$ The sum of weighted residuals weighted by the value of independent regressor variable is zero.
4) $\sum \hat{Y}_i E_i = 0$ The sum of weighted residuals weighted by the value of fitted response variable is zero .The mandatory condition here is that $\sum E_i Y_i \neq 0$

Proof of 1) $\sum E_i = \sum (Y_i - \hat{Y}) = 0$

Using the least square method minimizes the sum of squared residuals (also called SS Residuals or RSS)

$$RSS = S(A,B) = \sum_{i=1}^{n} E_i^2 = \sum (Y_i - \hat{Y}_i)^2 = \sum (Y_i - A - BX_i)^2$$

Gives us the first normal equation when it is partially differentiated wrt A

$$(-2)\sum_{i=1}^{n}(Y_i - A - BX_i) = 0 \sum_{i=1}^{n}(Y_i - \hat{Y}_i) \Rightarrow \sum E_i = 0$$ The first property is proved.

Proof of 2) $\sum Y_i = \sum \hat{Y}_i$

Using the first property $\sum E_i = 0$

$$\sum E_i = 0 \Rightarrow \sum (Y_i - \hat{Y}_i) = 0$$

$$\Rightarrow \sum Y_i = \sum \hat{Y}_i$$

Sum of observed values= sum of the fitted values

Proof of 3) $\sum X_i E_i = 0$ the sum of weighted residuals weighted by the value of independent regressor variable is zero

Using the least square method minimizes the sum of squared residuals (also called SS Residuals or RSS)

$$RSS = S(A,B) = \sum_{i=1}^{n} E_i^{\ 2} = \sum (Y_i - \hat{Y}_i)^2 = \sum (Y_i - A - BX_i)^2$$

Gives us the first normal equation when it is partially differentiated wrt B

$$(-2)\sum_{i=1}^{n} X_i(Y_i - A - BX_i) = 0 \sum_{i=1}^{n} X_i(Y_i - \hat{Y}_i) \Rightarrow \sum X_i E_i = 0$$ The third property is proved.

Proof of 4)

$\sum \hat{Y}_i E_i = 0$ The sum of weighted residuals weighted by the value of fitted response variable is zero .The mandatory condition here is that $\sum E_i Y_i \neq 0$

This is a consequence of the first and third property.

$$\sum \hat{Y}_i E_i \Rightarrow \sum (A + BX_i)E_i \ \sum AE_i + \sum BX_i E_i \Rightarrow A\sum E_i + B\sum X_i E_i = 0$$

The two terms are zero by property 1 & 3