Module: The Classical Ordinary Least Square Method

Regression Analysis

The term regression was introduced by Francis Galton. The average height of children born of parents of a given height tended to move or regress toward the average height in population as a whole. This Galton's law of universal regression was confirmed by his friend Karl Pearson.

Modern interpretation of regression is quite different. Regression analysis is concerned with the study of dependence of one variable, the dependent variable, on one or more variables, the explanatory variables.

The objective of the regression analysis is to estimate and/or predicting the (population) mean or average value of the dependent variable in terms of the known or fixed (in repeated sampling) values of the independent or explanatory variables.

Simple Linear Regression

We study the estimation of a linear relationship between two variables, Y_i and X_i of the form:

$$Y_i = \alpha + \beta X_i + u_i, \quad i = 1, 2, \dots\dots, n$$

where Y_i denotes the i-th observation on the dependent variable Y which could be consumption, investment or output, and X_i denotes the

i-th observation on the independent variable X which could be income, the interest rate and an input.

Cross Section Data – If collected on firms or households at a given point in time.

Time Series Data – If collected over time for a specific industry or country.

n: no. of firms or households in case of cross section data.

n: no. of years if the observations are collected annually.

α and β are the unknown parameters to be estimated from the data. A plot of data i.e. Y versus X shows the relationship exists empirically between X and Y. Let us assume that Y is consumption and X is disposable income. Therefore, we would expect a positive relationship between these two variables and the data may look like figure 1 below when we plotted for a random sample of households.

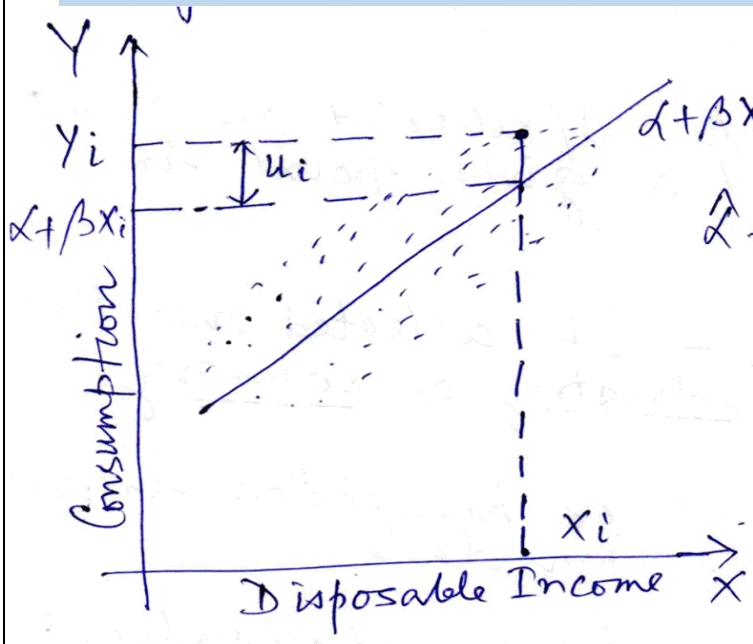


Figure 1: True Consumption function

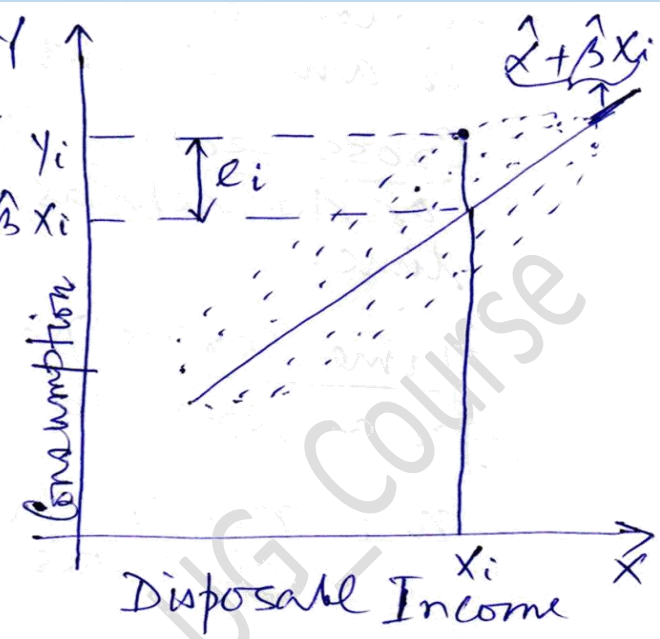


Figure 2: Estimated Consumption Fn

If α and β are known, one could draw a straight line $(\alpha + \beta X_i)$ as shown in figure 1. It is clear that not all the observations (X_i, Y_i) lie on the straight line $(\alpha + \beta X_i)$ is due to random error u_i. This error may be due to:

(i) Omission of relevant factors that could influence consumption other than income, like real wealth, varying taste or unforeseen events, age, sex, religion, no. of family members that induce households to consume more or less.

(ii) Measurement errors, which could be the result of households not reporting their consumption or income accurately.

(iii) Wrong choice of a linear relationship between consumption and income, when the true relationship may be non-linear.

These different causes of error term will have different effects on the distribution of this error.

In real life, α and β are not known, and have to be estimated from the observed data $\{(X_i, Y_i) for\ i = 1,2, \ldots\ldots, n\}$. Thus, true line $(\alpha + \beta X_i)$ as well as true disturbances (the u_is) are unobservable. Here, α and β could be estimated by the best fitting line through the data. Different researchers may draw different lines through the same data. What makes the line better than the other? One measure of misfit is the amount of error from the observed Y_i to the guessed line/estimated line. Therefore, $\widehat{Y}_i = \hat{\alpha} + \hat{\beta} X_i$ is the estimated line. $\wedge$ denotes the estimate or guess on the appropriate parameter or variable.

In other word, we obtain the predicted or guessed $Y_i(\widehat{Y}_i)$ corresponding to each X_i, from the predicted line, $\hat{\alpha} + \hat{\beta} X_i$. Next, we find the error in guessing that Y_i, by subtracting the actual Y_i from the guessed $\widehat{Y}_i$.

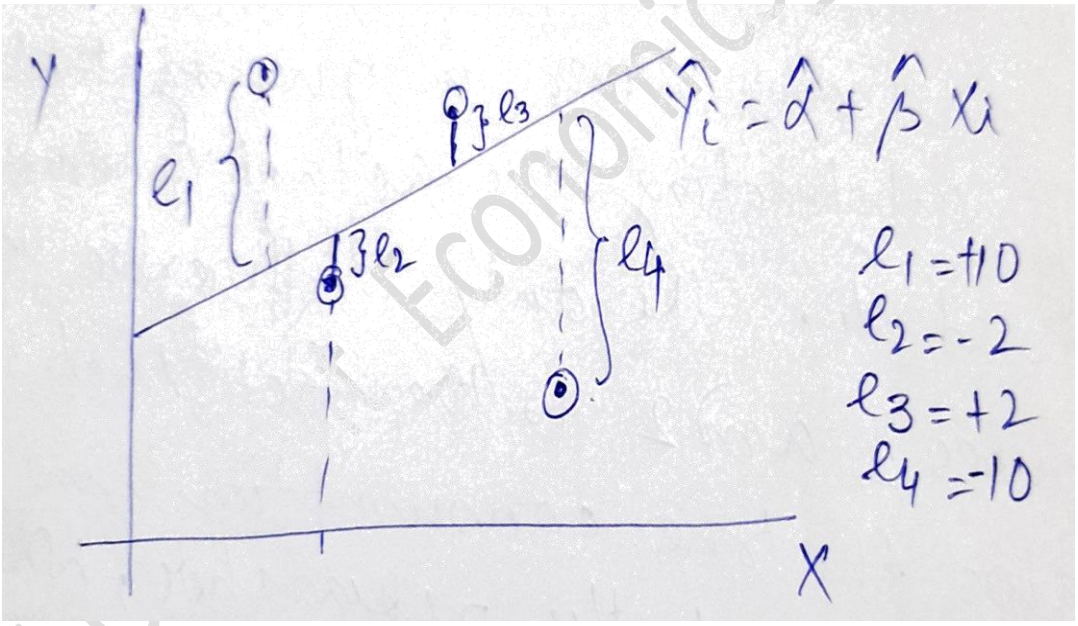
$\therefore\ e_i = Y_i - \widehat{Y}_i$ is the error

$\widehat{Y}_i = \hat{\alpha} + \hat{\beta} X_i$: Estimated Relationship

The only difference between figures 1 and 2 is the fact that figure 1 draws the true consumption line which is unknown to the researcher. Whereas figure 2 is a predicted or guessed consumption line drawn from the sample data. Therefore, $u_i s$ are unobservable, the $e_i s$ are

observable. Note that there will be n errors for each line, one error corresponding to every observation.

Now given the n pairs of observations on Y and X, we would like to determine the estimated consumption line in such a manner it is as close as possible to the actual/true consumption line. To do this we choose the consumption line (estimated regression line) in such a way the sum of the errors $\sum_{i=1}^{n} e_i = \sum_{i=1}^{n}(Y_I - \hat{Y}_i)$ is as small as possible.



If we adopt the criterion of minimizing $\sum_{i=1}^{n} e_i$, the above figure shows that the residuals e_2 and e_3 as well as the errors e_1 and e_4 receive the same weight in the sum ($e_1 + e_2 + e_3 + e_4$). However, the errors e_2 and e_3 are much closer to the regression line than the latter two errors e_1 and e_4.

As a consequence of giving equal weight to the errors, the sum of these errors is zero ($\sum_{i=1}^{n} e_i = 10 - 2 + 2 - 10 = 0$) although e_1 and e_4 are more scattered around the regression line than e_2 and e_3. We can avoid this problem if we adopt the Least Square Criterion, which states that the estimated regression line can be chosen in such a way that $\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2$ is as small as possible i.e., sum of the squares of the errors is minimum. By squaring e_i, this method of least squares gives more weight to the residuals such as e_1 and e_4 than e_2 and e_3. Under the minimum $\sum_{i=1}^{n} e_i$ criterion, the sum can be small (even zero) even though the e_i's are widely spread about the estimated regression line. But this is not possible under the least squares criterion, for the larger e_i (in absolute value), the larger the $\sum_{i=1}^{n} e_i^2$.

Ordinary Least Squares (OLS) method gives the estimated α and β by minimizing $\sum_{i=1}^{n} e_i^2$.

Least Squares Estimation/Least Squares Criterion

Least squares estimation minimizes the residual sum of squares where the residuals are given by

$e_i = Y_i - \widehat{Y}_i = Y_i - \widehat{\alpha} - \hat{\beta}X_i, i = 1,2, \ldots\ldots, n$ and $\widehat{\alpha}$ and $\hat{\beta}$ denote the estimated values of the regression parameters α and β. The residual sum of squares (RSS) is written as

$$RSS = \sum_{i=1}^{n} e_i^2 = (Y_i - \widehat{\alpha} - \hat{\beta}X_i)^2$$

This sum of squared residuals is minimized by two first order conditions

$$\frac{\partial \sum_{i=1}^{n} e_i^2}{\partial \hat{\propto}} = 0 \text{ ---------- (1)}$$

$$\frac{\partial \sum_{i=1}^{n} e_i^2}{\partial \hat{\beta}} = 0 \text{ ---------- (2)}$$

From (1) we get

$$\frac{\partial \sum_{i=1}^{n} e_i^2}{\partial \hat{\propto}} = 0 \quad \Rightarrow \quad \frac{\partial \sum_{i=1}^{n} \left(Y_i - \hat{\propto} - \hat{\beta} X_i\right)^2}{\partial \hat{\propto}} = 0$$

$$\Rightarrow 2\sum_{i=1}^{n}\left(Y_i - \hat{\propto} - \hat{\beta} X_i\right) = 0 \quad \Rightarrow \sum_{i=1}^{n} Y_i = \sum_{i=1}^{n} \hat{\alpha} + \hat{\beta} \sum_{i=1}^{n} X_i$$

$$\Rightarrow \sum_{i=1}^{n} Y_i = n\hat{\alpha} + \hat{\beta} \sum_{i=1}^{n} X_i \text{ ---------- (1a)}$$

From 2 we get

$$\frac{\partial \sum_{i=1}^{n} e_i^2}{\partial \hat{\beta}} = 0$$

$$\Rightarrow \frac{\partial \sum_{i=1}^{n} \left(Y_i - \hat{\propto} - \hat{\beta} X_i\right)^2}{\partial \hat{\beta}} = 0 \quad \Rightarrow 2\left[\sum_{i=1}^{n}\left(Y_i - \hat{\propto} - \hat{\beta} X_i\right)\right](-X_i) = 0$$

$$\Rightarrow \sum_{i=1}^{n} X_i Y_i = \hat{\propto} \sum_{i=1}^{n} X_i + \hat{\beta} \sum_{i=1}^{n} X_i^2 \text{ ---------- (1b)}$$

(1a) and (1b) equations are called normal equations. Solving those normal equations, we get values of $\hat{\propto}$ and $\hat{\beta}$. Multiplying (1a) by $\sum_{i=1}^{n} X_i$ and (1b) by n and subtract them we get **(1a*$\sum_{i=1}^{n} X_i$ - 1b*n)**

$$\sum_{i=1}^{n} X_i \sum_{i=1}^{n} Y_i = n\hat{\alpha} \sum_{i=1}^{n} X_i + \hat{\beta}(\sum_{i=1}^{n} X_i)^2$$

$$n\sum_{i=1}^{n} X_i Y_i = n \hat{\propto} \sum_{i=1}^{n} X_i + n\hat{\beta} \sum_{i=1}^{n} X_i^2$$

\- \- \-

$$\sum_{i=1}^{n} X_i \sum_{i=1}^{n} Y_i - n\sum_{i=1}^{n} X_i Y_i = \hat{\beta}(\sum_{i=1}^{n} X_i)^2 - n\hat{\beta} \sum_{i=1}^{n} X_i^2$$

$$\Rightarrow \hat{\beta}[(\sum_{i=1}^{n} X_i)^2 - n\sum_{i=1}^{n} X_i^2] = \sum_{i=1}^{n} X_i \sum_{i=1}^{n} Y_i - n\sum_{i=1}^{n} X_i Y_i$$

$$\Rightarrow \hat{\beta}[n\sum_{i=1}^{n} X_i^2 - (\sum_{i=1}^{n} X_i)^2] = n\sum_{i=1}^{n} X_i Y_i - \sum_{i=1}^{n} X_i \sum_{i=1}^{n} Y_i$$

$$\Rightarrow \hat{\beta} = \frac{n\sum_{i=1}^{n} X_i Y_i - \sum_{i=1}^{n} X_i \sum_{i=1}^{n} Y_i}{n\sum_{i=1}^{n} X_i^2 - (\sum_{i=1}^{n} X_i)^2}$$

$$\Rightarrow \hat{\beta} = \frac{\sum_{i=1}^{n} X_i Y_i - {}^1/_n \sum_{i=1}^{n} X_i \sum_{i=1}^{n} Y_i}{(\sum_{i=1}^{n} X_i)^2 - {}^1/_n \sum_{i=1}^{n} X_i^2}$$

$$\Rightarrow \hat{\beta} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

$$\Rightarrow \hat{\boldsymbol{\beta}} = \frac{\sum_{i=1}^{n} \boldsymbol{x_i y_i}}{\sum_{i=1}^{n} \boldsymbol{x_i}^2}$$

Where $x_i = \sum_{i=1}^{n}(X_i - \bar{X})$ and $y_i = \sum_{i=1}^{n}(Y_i - \bar{Y})$

Dividing (1a) by n we get

$$\hat{\boldsymbol{\alpha}} = \bar{\boldsymbol{Y}} - \hat{\boldsymbol{\beta}}\bar{\boldsymbol{X}}$$