[Academic Script]

Linear Regression Model

Subject:	Business Economics
Course:	B.A., 3rd Semester, Undergraduate
Paper No. & Title:	Paper – 304 Basic Econometrics
Unit No. & Title:	Unit - 2 Classical Two Variable Linear Regression Model
Lecture No. & Title:	Lecture – 1 Linear Regression Model

Academic Script

1. Introduction

In this module we talk about two topics.

A. Types of data: Time series data, Cross-section data and Penal Data

B. Construction of two variables linear Regression Model.

2. Types of Data

What is Data?

Data are information about a variable. It can be numerical or categorical.

1. Time series Data:

If we add "time" element in the data it becomes Time series Data. Time series is a set of observations recorded at different point of time. The Chronological order of observations provides important information about that variable.

If observations are made on discrete set of time (i.e. fixed point of time) the series is called discrete time series. The data set of daily stock price, monthly sales and yearly production are collected at regular interval of time and thus generate a discrete time series.

In discrete time series the data sets are collected for single variable over different point of time. e.g. Production of Milk Powder during March 2011 to Sep 2011.

Month & Year	Mar-11	Apr-11	May-11	Jun- 11	Jul-11	Aug-11	Sep-11
Prod. Of Milk Powder	11,252	10,159	8,817	7,953	6,791	8,336	7,862

If the data are recorded continuously over an interval of time i.e. $[t_0, t_1]$ it is called continuous time series. e.g. Temperature of a place on a particular day between 2 p.m. to 3 p.m.

2. Cross-section Data:

In time series we take one variable data over different point of time. Now Instead of this, if we take data of more variables at single point of time it becomes cross section data. i.e. In Cross section data, Information of many variables are collected at a single point of time.

e.g. Production of Milk Powder, Biscuits, Chocolate and Sugar, Confectionary and Malted food in March 2011.

Month & Year	Milk Powder	Biscuits	Chocolate & Sugar Confectionary	Malted Food
Mar-11	11252	50522	11510	4172

3. Panel Data:

If both the time series and cross section data are combined we get panel data. i.e. In panel data we collect information on different variables over different period of time.

e.g. If the information of the same household is collected over different period of time (same survey is repeatedly conducted over different time) it becomes panel data. Data set of various productions over last 10 Months also generates panel data.

Month & Year	Milk Powder	Biscuits	Chocolate and Sugar Confectionary	Malted Food
Mar-2011	11252	50522	11510	4172
Apr-2011	10159	53199	11208	3886
May-2011	8817	57649	10172	4726
Jun-2011	7953	60954	9267	4404
Jul-2011	6791	63477	11484	3962
Aug-2011	8336	59018	9320	4216
Sep-2011	7862	57884	9750	4556

Oct-2011	8356	46243	11061	4137
Nov-2011	13434	52057	11518	6056
Dec-2011	18570	51774	11096	5770

3. Construction of two variable Linear Regression Equation.

1. Concept of Population Regression Function (PRF)

Before understanding the regression let's take a look at Independent variable and Dependent Variable. These concepts are very important in econometrics. The variable which changes independently is called Independent variable. The variable whose change depends on Independent variable or other factors is called dependent variable. e.g. Income and Expense, Income is independent variable and Expense depends on level of Income. Similarly in Rainfall and Production of crop, Rainfall is independent variable and production of crop depends on amount of Rainfall.

Let's assume that there is a population of 100 families living in a village. We have the details of Independent variable X (Income) and dependent variable Y (Expenditure) of these families. We want to predict the average level of family expenditure(Y) depending on their income(X) level.

$$E(Y) = f(X) \quad \dots \quad (1)$$

Equation (1) is functional relationship between two variable and so it is called two variables Population Regression Function (PRF). The objective is to establish relationship between income (X) and expenditure (Y), but we do not know the kind of relationship (linear or non linear) between these two variables. Suppose we assume expenditure and income are linearly related than the population regression function (PRF) of Y is a linear function of X.

i.e. $E(Y_i) = \beta_1+\beta_2X_i$ … (2)

Here β_1 and β_2 are regression co-efficient, where β_1 is known as intercept and β_2 is known as slope of a regression line. It is also known as Population regression Model or regression equation.

We are interested in estimating the values of β_1 and β_2 on the basis of observed value of X and Y. In equation (2) we predict expenditure on the basis of only one variable, Income. But in real life there are so many other variables like number of family members and number of luxuries used by that family, which also affect the expenditure whose effect are not considered in this equation. If we take u_i as the effect of all the other such variables that affect the expenditure then the equation (2) becomes

$$Y_i = \beta_1+\beta_2X_i + u_i \quad \text{… (3)}$$

Here u_i is known as disturbance term. In other words we can say that if Y_i is the observed value and $E(Y_i)$ is the expected value of expenditure than error term is given by

$$u_i = Y_i - E(Y_i) \quad \text{…(4)}$$

$$\text{or} \quad Y_i = E(Y_i) + u_i$$

$$\text{or} \quad Y_i = \beta_1+\beta_2X_i + u_i$$

Our objective is not only to estimate $\hat{\beta}_1$ and $\hat{\beta}_2$ and obtain functional form but also make inference about β_1 and β_2. For that we must make certain assumptions about Independent variable X_i and error u_i. Without satisfying the assumption, we cannot make inference about Y_i and valid interpretation of regression estimates.

Following are the assumption of Gaussian classical linear regression model.

(i) Mean value of u_i is Zero. E (u_i) = 0

i.e. from the factors that are not considered, some have positive effect and some have negative effect and the average effect is Zero.

(ii) Disturbance terms u_i and u_j are uncorrelated. $Cov(u_i,u_j) = 0$

(No autocorrelation between u_i's.)

(iii) Homoskedasticity (constant Variance σ^2 for each u_i)

$$V(u_i) = E[u_i - E(u_i)]^2 = \sigma^2$$

Variance of u_i for each X_i is positive constant equals to σ^2.

(iv) Disturbance term u_i and explanatory variable (X_i) are uncorrelated.

$$Cov(u_i,x_i) = E\{u_i - E(u_i)\}\{X_i - E(X_i)\}$$
$$= E[u_i \{X_i - E(X_i)\}] \text{ , Since } E(u_i) = 0.$$
$$= E(u_i,x_i) = 0.$$

(V) The Regression model is correctly specified. In our case there is a linear relationship between two variables.

2. Sample Regression Function (SRF):

We have discussed how population regression function (PRF) is constructed. But in most of the practical cases we have sample values of the expenditure (Y) and Income (X) rather than the population values (all the values). So the question is, can we predict the values of expenditure (Y_i) from the given sample values of Income (X_i)?

Yes, like PRF we can also write the sample regression function (SRF) which represents the sample regression line.

$$\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i \qquad \ldots(5)$$

Where, $\hat{Y}_i$ = estimate of E(Y_i) for given X_i

$\hat{\beta}_1$ = estimate of β_1, $\hat{\beta}_2$ = estimate of β_2

If we add unobserved factor, the SRF can be written as

$$\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i + e_i \quad \ldots(6)$$

Where, e_i is the sample residual term.

We can have N number of such samples from the population and we can build N such SRFs. Now the question is which SRF accurately predict Population Regression Function (PRF). Is there any method, which identifies the best SRF from the N SRFs? i.e. $\hat{\beta}_1$ and $\hat{\beta}_2$ are as close as β_1 and β_2. This question can be answered by ordinary least square Method.

3. Estimation of SRF using Ordinary Least Square (OLS) Method:

We have seen that there are N SRFs and the question is to identify the best-fitted SRF. i.e. SRF having error sum of square ($\sum e_i^2$) as small as possible (Minimum). In other words actual and estimated values are very close.

$$\sum e_i^2 = \sum (Y_i - \hat{Y}_i)^2 \quad \ldots(7)$$

From equation (6) we have,

$$\sum e_i^2 = \sum (Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i)^2 \quad \ldots(8)$$

Thus e_i is the function of $\hat{\beta}_1$ and $\hat{\beta}_2$. If we select different set of values of Y_i and X_i and obtained different $\hat{\beta}_1$ and $\hat{\beta}_2$ then we have different values of e_i and we can find minimum $\sum e_i^2$. Do we have to apply this trial and error method every time to find minimum $\sum e_i^2$? No, we apply method of least square for the selection of best SRF. The principal of least square gives us $\hat{\beta}_1$ and $\hat{\beta}_2$ in such a way that $\sum e_i^2$ will be minimum. Differentiating above equation with respect to $\hat{\beta}_1$ and $\hat{\beta}_2$ and equating it with zero we get

$$\frac{\partial \sum e_i^2}{\partial \hat{\beta}_1} = 2(-1)\sum (Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i) = 0$$

$$\frac{\partial \sum e_i^2}{\partial \hat{\beta}_2} = 2(-1)\sum(Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i)X_i = 0$$

The normal equations of line are

$$\sum Y_i = n\,\hat{\beta}_1 + \hat{\beta}_2 \sum X_i \quad \ldots(9)$$

$$\sum Y_i X_i = \hat{\beta}_1 \sum X_i + \hat{\beta}_2 \sum X_i^2 \quad \ldots(10)$$

Where n is sample size. The estimates are known as least square estimate as they are derived from least squared principle.

Solving these equations we get

$$\hat{\beta}_2 = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2} \quad \ldots(11)$$

and $\hat{\beta}_1 = \bar{Y} - \hat{\beta}_2 \bar{X}$ …(12)

which is also useful in finding Regression coefficients.

Let's take one example of fitting of regression line.

Example: Fitting of Regression Line

From the following sample data construct linear Regression line and estimate the regression coefficients. Also estimate Expense Y from given value of Income X and find Residue/error.

X (Income)	4	7	3	9	17
Y(Expense)	3	5	2	6	9

Solution:

The normal equation of line is given by

$$\sum Y = n\hat{\beta}_1 + \hat{\beta}_2 \sum X$$

$$\sum XY = \hat{\beta}_1 \sum X + \hat{\beta}_2 \sum X^2$$

First let's find all this values n, $\sum X$, $\sum Y$, $\sum XY$ and $\sum X^2$

Income (X)	Expense(Y)	XY	X^2	$\hat{Y}$	$e = (Y - \hat{Y})$
4	3	12	16	3.06	-0.06
7	5	35	49	4.52	0.48

3	2	6	9	2.58	-0.58
9	6	54	81	5.48	0.52
17	9	153	289	9.35	-0.35
$\sum X$ =40	$\sum Y$ =25	$\sum XY$ =260	$\sum X^2$ = 444	$\sum \hat{Y}$ =25	$\sum e_i$ =00

$\bar{X} = \frac{\sum X}{n} = \frac{40}{5} = 8$ $\bar{Y} = \frac{\sum Y}{n} = \frac{25}{5} = 5$

By putting all the values in above equation we get,

$25 = \hat{\beta}_1(5) + \hat{\beta}_2(40)$ $\therefore$ $5\hat{\beta}_1 + 40\hat{\beta}_2 = 25$ ...(1.1)

and

$260 = \hat{\beta}_1 (40) + \hat{\beta}_2(444)$ $\therefore$ $40\hat{\beta}_1 + 444\hat{\beta}_2 = 260$ …(1.2)

By solving them we get $\hat{\beta}_2 = 0.4838$ and

$\hat{\beta}_1 = 1.1296$

The regression equation is written as

$$\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i$$

$$\hat{Y}_i = 1.1296 + 0.4838\, X_i$$

Now by putting different values of X we will have different estimated values of Y which is shown in second last column of the table. The last column indicates error term which is obtained by subtracting Predicted value from original value. i.e. $e_i = (Y_i - \hat{Y}_i)$

By Putting X =10 we get predicted value of Y for X = 10.

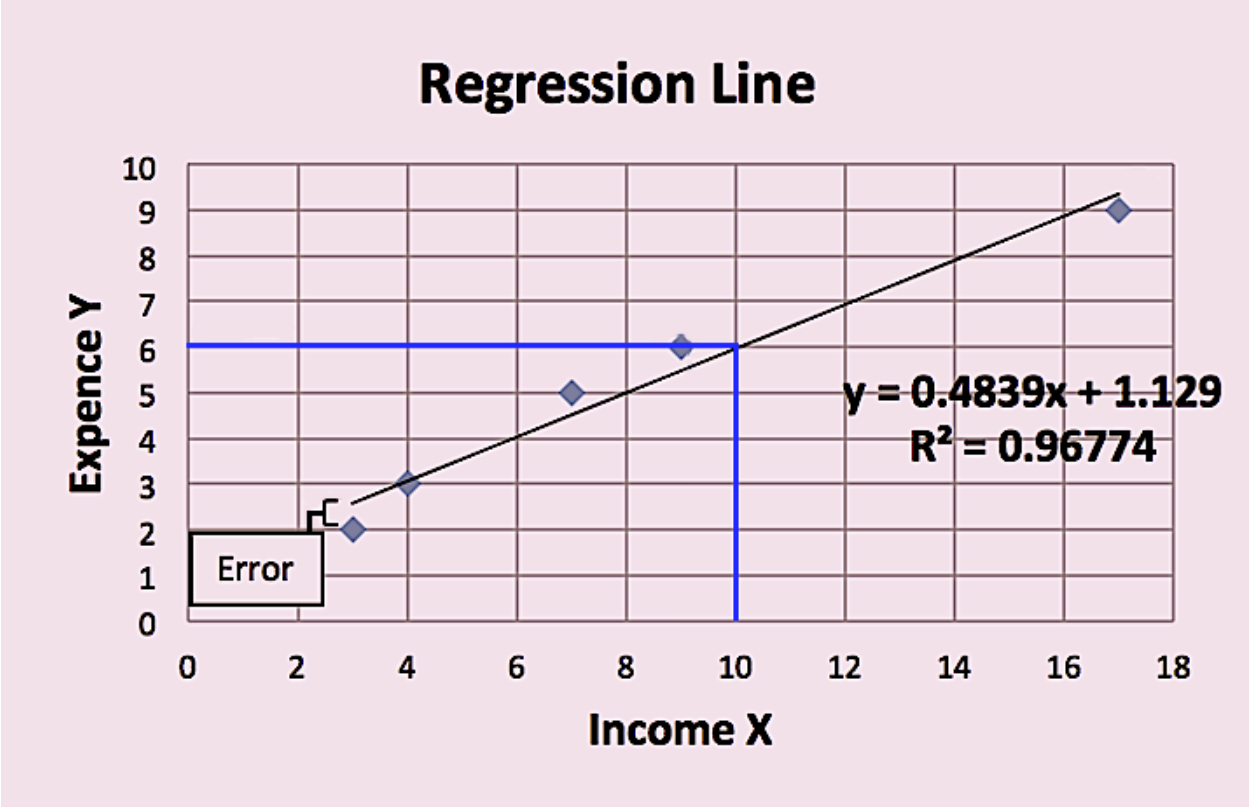
` $\hat{Y} = 1.1296 + 0.4838\,(10)$

$= 5.9676$

Here we have to note that $\sum Y$ and $\sum \hat{Y}$ are equal (In our case it is 25) and $\sum e_i$=0.

We can also understand the above example with following chart. X axis represent Income and Y Axis indicates Expense. Pairs of points of Income and expenses (X,Y) i.e (4,3),(7,5) … is denoted

by blue points. The line passing through the points are regression line obtained using ordinary least square method. Vertical line at X=10 intersect regression line around 6 indicate that expected value of Y at X=10 is 6. The distance between original value and regression line is error.



4. Properties of Regression Line:

1. It passes through the sample mean $\bar{X}$ and $\bar{Y}$.
2. Mean value of estimated Y (= $\hat{Y}_i$) is equal to mean value of actual Y_i. $\bar{\hat{Y}} = \bar{Y}$.
3. The mean value of Residual is zero. $\sum e_i = 0$.
4. The Residuals are uncorrelated with Predicted Y_i. $\sum \hat{Y} e_i = 0$.
5. The Residuals are uncorrelated with X_i. $\sum X_i e_i = 0$.

4. Summary

In this talk we have discussed different types of data namely Time series data, Cross-section data and Penal data with example. Concept of Population Regression Function and Sample regression function based on two Variables are discussed. OLS method for solving regression equation is explained and properties of regression line are shown.