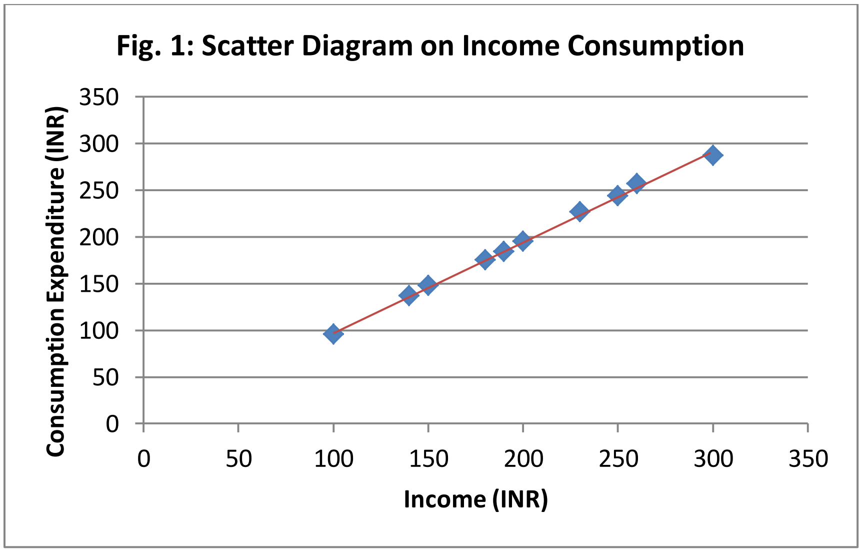
Determining Simple Linear Regression

Fitting a straight line

Very often, it may be found that a relationship exists between two variables say X and Y. To study the nature of relationship between X and Y, we first collect data on the variables. Suppose we are able to collect data on income and consumption for a sample size of say 10 observations. Suppose we denote income by X and consumption by Y. We all know that consumption is dependent on income. Hence for each observed income level (X_i), there must be a corresponding level of consumption (Y_i), where (i=1, 2, 3,..,10). The data obtained for say ten years is represented in Table 1 as follows:

Year	Number of observations (n)	Income (X) In INR	Consumption (Y) In INR	Coordinates on the XY plane
2001	1	100	96	100,96
2002	2	140	137	140,137
2003	3	150	148	150,148
2004	4	180	175	180,175
2005	5	190	184	190,184
2006	6	200	195	200,195
2007	7	230	227	230,227
2008	8	250	244	250,244
2009	9	260	257	260,257
2010	10	300	287	300,287

The next step in determining the relationship between X and Y that is in our example income and consumption, we need to plot the points determined as coordinates (X_i,Y_i), in the Table 1,on XY plane or on a rectangular coordinate plane.



From the scatter diagram (Fig.1), it is almost possible to observe a smooth straight line curve which touches most of the points of the scatter diagram. This approximately predicts the nature of relationship between income and consumption to be a **linear** relationship. Such a curve drawn to fit a set of data is called approximating curve and the method is called free hand method of curve fitting. The biggest demerit of fitting a curve by this method is that the result is dependent on individual judgement which may be biased. To determine the best fitting curve between two variables we usually take the help of a method called ‘Method of Ordinary Least Squares’.

Method of Ordinary Least Squares

The Method of Ordinary Least Squares or OLS is a statistical technique to determine the best fitting curve for a set of points like ((X_i,Y_i), **plotted on a curve, based on a given data, by minimizing the sum of the squared residuals. It is used to predict the behaviour of the dependent variable** Y_i **in relation to the independent variable** X_i**.**

Here our first task is to clarify the meaning of the concept ‘residuals’. Let us take the earlier example on income and consumption. All the combination points on income and consumption **(**(X_i,Y_i) seem to lie on a straight line. However we might not be so lucky each time and there might be data points on the scatter diagram which lie outside the plotted curve. This is shown in Fig.2.

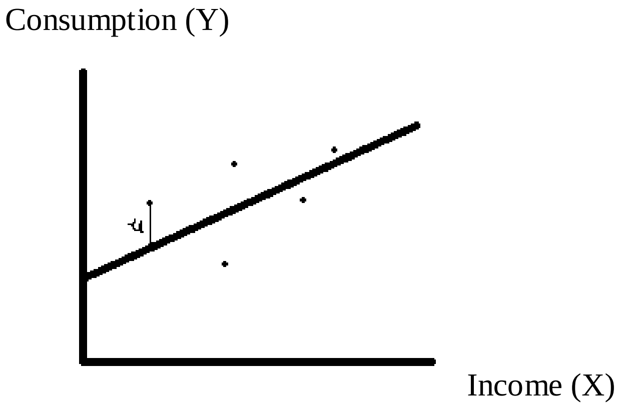


Fig 2. Diagram showing residual ‘U’

The vertical distance or deviation of a point from the fitted straight line is called the residual shown as U in fig. 2. The OLS method aims to find the best fitting curve by minimizing the sum of the squared residuals.

Suppose the relationship between income and consumption can be expressed in the form of a linear mathematical model of the form $Y = \beta_1 + \beta_2 X$. Here Y = consumption and is called the dependent variable. X =income and is called the independent or explanatory variable, β_1 and

β_2 are known as parameters, β_1 indicates the intercept of the linear equation on Y axis and the coefficient β_2 is the slope of the linear equation.

In the mathematical model, we state that the consumption of the consumer is exactly dependent on income or exactly determined by income (and no other variable), i.e., it is a **deterministic model**. In reality, very often, there might be other factors like the habits and preferences of a customer which influence the consumption pattern of the consumer. Hence to capture such behaviour or unexplained variation, the economist adds a disturbance term or error term. This is also referred to as residuals. These residuals are random (stochastic) variables with a probability distribution.

$$Y = \beta_1 + \beta_2 X + U, \quad 0 < \beta_2 < 1 \text{ and U is the disturbance term.}$$

This is known as an **econometric model**.

In the econometric model, $Y = \beta_1 + \beta_2 X + U$, for each i, $Y_i = \beta_1 + \beta_2 X_i + U_i$, where i= 1, 2, 3,...,n. [assuming the total number of observed data set is n].

This means for each X_i, a corresponding value of Y_i can be observed in the data obtained. If the values of β_1 and β_2 were known, it would become easy to calculate the error term U_i for each i, where (i= 1, 2, 3,..., n).

If the values of the parameters β_1 and β_2 and the error term U_i were already determined, then using the values, for each corresponding X_i, a new value of Y would be obtained which we call the estimated Y or $\hat{Y}_i$ for a given i, (i= 1, 2, 3,...,n). The difference $(Y_i – \hat{Y}_i) = U_i$ for all the values of i, (i= 1, 2, 3,...,n).

Thus the first task is to determine the values of β_1 and β_2 on the basis of data shown in Table 1. To do so, we take the help of Ordinary Least Squares Method, i.e. OLS method. **The term Least Squares means we are trying to minimize the sum of squares or we are trying to minimize the squared error terms.**

Generalized Working

For each i, (i= 1, 2, 3,...,n) ,we can now write from $Y_i = \beta_1 + \beta_2 X_i + U_i$, the error term as the difference between observed Y_i and estimated $\hat{Y}_i$, where $\hat{Y}_i = \beta_1 + \beta_2 X_i$

This means $U_i = Y_i – \hat{Y}_i$, or, $U_i = Y_i – (\beta_1 + \beta_2 X_i)$, or, $U_i = Y_i – \beta_1 – \beta_2 X_i$

$U_1 = Y_1 – \beta_1 – \beta_2 X_1$	$U_1^2 = (Y_1 – \beta_1 – \beta_2 X_1)^2$
$U_2 = Y_2 – \beta_1 – \beta_2 X_2$	$U_2^2 = (Y_2 – \beta_1 – \beta_2 X_2)^2$
$U_3 = Y_3 – \beta_1 – \beta_2 X_3$	$U_3^2 = (Y_3 – \beta_1 – \beta_2 X_3)^2$
$U_4 = Y_4 – \beta_1 – \beta_2 X_4$	$U_4^2 = (Y_4 – \beta_1 – \beta_2 X_4)^2$

$U_5 = Y_5 – \beta_1 – \beta_2 X_5$	$U_5^2 = (Y_5 – \beta_1 – \beta_2 X_5)^2$
$U_6 = Y_6 – \beta_1 – \beta_2 X_6$	$U_6^2 = (Y_6 – \beta_1 – \beta_2 X_6)^2$
$U_7 = Y_7 – \beta_1 – \beta_2 X_7$	$U_7^2 = (Y_7 – \beta_1 – \beta_2 X_7)^2$
$U_8 = Y_i – \beta_1 – \beta_2 X_8$	$U_8^2 = (Y_i – \beta_1 – \beta_2 X_8)^2$
$U_9 = Y_9 – \beta_1 – \beta_2 X_9$	$U_9^2 = (Y_9 – \beta_1 – \beta_2 X_9)^2$
$U_{10} = Y_{10} – \beta_1 – \beta_2 X_{10}$	$U_{10}^2 = (Y_{10} – \beta_1 – \beta_2 X_{10})^2$
Continued to nth data set $U_n = Y_n – \beta_1 – \beta_2 X_n$	Continued to nth data set $U_n^2 = (Y_n – \beta_1 – \beta_2 X_n)^2$
$\Sigma U_i = \Sigma (Y_i – \beta_1 - \beta_2 X_i) = 0$, **Where i =** 1, 2, 3,...,n	**$\Sigma U_i^2 = \Sigma (Y_i – \beta_1 - \beta_2 X_i)^2$** **Where i =** 1, 2, 3,...,n

The summation of error terms is given by $\Sigma U_i = \Sigma (Y_i – \beta_1 - \beta_2 X_i) = 0$ and the summation of the square of error terms is given by $\Sigma U_i^2 = \Sigma (Y_i – \beta_1 - \beta_2 X_i)^2$.

Suppose $f = \Sigma U_i^2 = \Sigma (Y_i – \beta_1 - \beta_2 X_i)^2$.

Now using partial derivative our objective under OLs method is to minimize f with respect to β_1 and β_2

$df/d\beta_1 = -2 \Sigma (Y_i – \beta_1 - \beta_2 X_i) = 0$

i.e. $\Sigma Y_i - n \beta_1 - \beta_2 \Sigma X_i = 0$ [for n observations]

i.e. $\Sigma Y_i = n \beta_1 + \beta_2 \Sigma X_i$ ………………1.

and

$df/d\beta_2 = -2 \Sigma (Y_i – \beta_1 - \beta_2 X_i). X_i = 0$

i.e. $\Sigma X_i Y_i - \beta_1 \Sigma X_i - \beta_2 \Sigma X_i^2 = 0$

i.e. $\Sigma X_i Y_i = \beta_1 \Sigma X_i + \beta_2 \Sigma X_i^2$ …………2.

The equations 1. and 2. are called normal equations. Solving the equations simultaneously, we get solutions to β_1 and β_2. Suppose the solutions give $\beta_1 = \beta_1^\wedge$ (pronounced beta 1 hat) and $\beta_2 = \beta_2^\wedge$ (pronounced beta 2 hat), then

$$\beta_2^\wedge = \frac{n \Sigma X_i Y_i - \Sigma X_i \Sigma Y_i}{n \Sigma X_i^2 - (\Sigma X_i)^2} \text{ or, } \beta_2^\wedge = \frac{\text{cov}(X,Y)}{\sigma_x^2} = \frac{\Sigma x_i y_i}{\Sigma x_i^2} \quad ……………………………4.$$

where $x_i = (X_i – \overline{X})$ and $y_i = (Y_i – \overline{Y})$

$\beta_1^\wedge = \overline{Y} - \beta_2^\wedge \overline{X}$ ………………………………….5.

Here for n number of observations, $\beta_1^\wedge$ and $\beta_2^\wedge$ are called the estimators for the true parameters β_1 and β_2 respectively. From data collected in Table 1, let us determine the estimators.

Table 2: Statistical Workings using Table1

Number of observations (n)	Income (X) In INR	Consumption (Y) In INR	x_i	y_i	x_iy_i	x_i^2
1	100	96	-100	-99	9900	10000
2	140	137	-60	-58	3480	3600
3	150	148	-50	-47	2350	2500
4	180	175	-20	-20	400	400
5	190	184	-10	-11	110	100
6	200	195	0	0	0	0
7	230	227	30	32	960	900
8	250	244	50	49	2450	2500
9	260	257	60	62	3720	3600
10	300	287	100	92	9200	10000
n =10	$\Sigma X_i =$ 2000 $\overline{X} = 200$	$\Sigma Y_i = 1950$ $\overline{Y} = 195$	$\Sigma x_i = 0$	$\Sigma y_i = 0$	$\Sigma\, x_iy_i = 32570$	$\Sigma\, x_i^2 = 33600$

Inserting the required values in 4, we obtain

$$\beta_2^\wedge = \frac{\Sigma x_iy_i}{\Sigma x_i^2} = \frac{32570}{33600} = 0.969 \ldots\ldots\ldots\ldots\ldots\ldots 6.$$

Inserting the value of $\beta_2^\wedge$, $\overline{X}$ and $\overline{Y}$ in equation 5, we get

$\beta_1^\wedge = 195 - 0.969 \times 200 = 195 - 193.8 = 1.2$ ………………7.

Thus solutions in 6 and 7 give the values of $\beta_2^\wedge$ and $\beta_1^\wedge$

Thus the estimated regression equation between income and consumption applicable to the data obtained by us is $\hat{Y}_i = 1.2 + 0.969X_i$

As $\beta_2^\wedge$ denotes slope, then slope = 0.969. The intercept of the consumption function is $\beta_1^\wedge = 1.2$. The estimated regression equation $\hat{Y}_i = 1.2 + 0.969X_i$ is known as a simple linear regression.

Definition of Simple Linear Regression

A Simple Linear Regression is a statistical tool used in prediction. It exhibits a functional relationship between two continuous variables X and Y, where X is called the independent variable or the explanatory variable or the predictor and Y is called the dependent variable or the response variable. In a population of N size, the general format for a simple linear regression can be denoted as

$$Y_i = \beta_1 + \beta_2 X_i + U_i, \quad \text{for all values of i, (i=1, 2, 3, ….., N)}$$

U_i is the disturbance or error term or residual term. β_1 and β_2 are known as parameters of the linear regression.

References:

Spiegel, M. R., & Stephens, L. J. (1961). *Schaum's outline of theory and problems of statistics* (Third ed.).McGraw Hill International.

Gujarati, D. N. Porter, D.C., Gunasekar, S. (2009), *Basic econometrics*. (Fifth ed.) McGraw-Hill Education (India).

Salvatore, D., & Reagle, D. (2011). *Schaum's outline of statistics and econometrics* (2nd ed.). McGraw-Hill Education.