Two Variable Regression Analysis

Regression analysis is largely concerned with estimating and/or predicting the (population) mean value of the dependent variable on the basis of the known or fixed values of the explanatory variables.

![Graph showing consumption expenditure vs income](image)

There is considerable variation in weekly consumption expenditure in each income group. On average, weekly consumption expenditure increases as income increases.

E (Y/X) conditional expected values, as they depend on the given values on the (conditioning) variable X. (expected values of Y given one value of X)

Geometrically, then, a population regression curve is simply the locus of the conditional means of the dependent variable for the fixed values of the explanatory variables. For simplicity, we are assuming that these Y values are distributed symmetrically around their prospective (conditional) mean values. And the regression line passes through these conditional mean values.

The Concepts of Population Regression Function

Each conditional mean E (Y/X_i) is a function of X_i, where X_i is a given value of X. 

E (Y/X_i) = f(X_i)

The expected value of the distribution of Y given X_i is functionally related to X_i. It tells how the mean or average response of Y varies with X.

PRF ---> E(Y / X_i) = \( \beta_1 + \beta_2 X_i \) if we assume linear relationship.

Where \( \beta_1 \) and \( \beta_2 \) are unknown but fixed parameters known as the regression coefficients or as intercept and slope coefficients, respectively.
Linear Regression Models

The expenditure of an individual family, given its income level, can be expressed as the sum of two components:

1.) $E(Y/X_i)$, which is simply the mean consumption expenditure of all the families with the same level of income.

2.) $U_i$, which is random component.

$Y_i = E(Y/X_i) + U_i = \beta_1 + \beta_2 X_i + U_i$

The Sample Regression Function

Now we want to estimate the PRF on the basis of the sample information.

We may not be able to estimate the PRF “accurately” because of sampling fluctuations. Every time we draw a sample from population, each sample will be somewhat different. There is no way to tell which sample drawn would represent population unless we have a way to know population, and if we do have a way to know population, then we do not need a sample.
Sample regression function can be expressed as:

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

where \( \hat{Y}_i \) is estimator of \( E(Y/X_i) \)

\( \hat{\beta}_1 \) is estimator of \( \beta_1 \)

\( \hat{\beta}_2 \) is estimator of \( \beta_2 \)

and in terms of the PRF, it can be expressed as \( Y_i = E(Y/X_i) + U_i \)

How should SRF be constructed so that \( \hat{\beta}_1 \) is “close” as possible to the true \( \beta_1 \) and \( \hat{\beta}_2 \) is as “close” as possible to the true \( \beta_2 \) even though we will never know the true \( \beta_1 \) and \( \beta_2 \)?