General linear Model

General linear Model (GLM) is given by

$$y_i = \beta_1 x_{11} + \beta_2 x_{i2} + \cdots + \beta_n x_{in} + \varepsilon_i$$

where

y_i = independent Variable

x_i = Regressor or independent variables

β_i = Weights or parameters

ε_i = Residual or error, such that

$E(\varepsilon_i) = 0$ and $V(\varepsilon_i) = \sigma^2$.

we can write

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_n x_{in} + \varepsilon_i$$

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nn} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

or

$$Y = Xb + \varepsilon$$

where Y, is $n \times 1$ matrix, X is $n \times n$ and b and ε are $n \times 1$ matrices.

Ordinary least Squares

The sum of squared residuals is given by

$$\varepsilon^2 = (Y - Xb)^2$$
$$= (Y - Xb)'(Y - Xb)$$
$$= (Y' - b'X')(Y - Xb)$$
$$= Y'Y - Y'Xb - b'X'Y + b'X'Xb$$

differentiating with respect to parameter b and equating with zero we have our normal equation given by

$$-Y'X + b'X'X = 0$$

$$b'X'X = Y'X$$

$$\text{but} \left\{ \begin{array}{l} b'X'X = X'Xb \\ Y'X = X'Y \end{array} \right\}$$

$$X'Xb = X'Y$$

$$b = (X'X)^{-1} X'Y$$

or

$$\hat{\beta} = (X'X)^{-1} X'Y$$

Generalized least Squares

We have the model

$$Y = X\beta + \varepsilon$$

s.t. $E(\varepsilon) = 0$ and $V(\varepsilon) = \sigma^2 V$

where V is known $n \times n$ matrix. If V is diagonal but with unequal diagonal elements, the observations y are uncorrelated but have unequal variance, while if V has non zero off diagonal elements, the observations are correlated.

If we estimate β by ordinary least squares, $\hat{\beta} = (X'X)^{-1}X'Y$, the estimator is not optimum. The solution is to transform the model to a new set of observations that satisfy the constant variance assumption.

Since $\sigma^2 V$ is a covariance matrix, V is a symmetric non singular matrix, therefore

$$V = K'K = KK \quad \text{also} \quad V^{-1} = K^{-1}K^{-1}$$

K is called the square root of V.

Let us define

$$z = K^{-1}y \quad , \quad B = K^{-1}X$$

$$g = K^{-1}\varepsilon$$

$$\Rightarrow \qquad z = B\beta + g$$

Here the least square function is

$$S(\beta) = (z - B\beta)^2$$

$$= (z - B\beta)'(z - B\beta)$$

$$= (K^{-1}Y - K^{-1}X\beta)'(K^{-1}Y - K^{-1}X\beta)$$

$$= (Y - X\beta)' K^{-1} K^{-1} (Y - X\beta)$$

$$= (Y - X\beta)' V^{-1} (Y - X\beta)$$

$$= (Y' - \beta' X') V^{-1} (Y - X\beta)$$

$$= Y'V^{-1}Y - Y'V^{-1}X\beta - \beta'X'V^{-1}Y + \beta'X'V^{-1}X\beta$$

differentiating with respect to β and setting it to zero, we get

$$\Rightarrow \quad -Y'V^{-1}X + \beta'X'V^{-1}X = 0$$

$$\beta'X'V^{-1}X = Y'V^{-1}X$$

$$\begin{cases} \text{but} \quad \beta' X' V^{-1} X = X' V^{-1} X \beta \\ \text{and} \quad Y' V^{-1} X = X' V^{-1} Y \end{cases}$$

Therefore

$$X' V^{-1} X \beta = X' V^{-1} Y$$

$$\hat{\beta} = (X' V^{-1} X)^{-1} X' V^{-1} Y$$

This is generalized least square of β.

Assumptions of OLS

1. Model is linear in parameters.
2. The data are a random sample of the population.
3. The expected value of the errors is always zero. i.e. $E(\varepsilon) = 0$.
4. The independent variables are not too strongly collinear.
5. The independent variables are measured precisely.
6. The residuals have constant variance. $V(\varepsilon) = \sigma^2$.
7. The errors are normally distributed.

Multicollinearity :

A basic assumption is multiple linear regression model is that the rank of the matrix of observations on explanatory variables is the same as the number of explanatory variables. In other words, such a matrix is of full column rank. This, in turn, implies that all the explanatory variables are independent, i.e. there is no linear relationship among the explanatory variables. It is termed that the explanatory variables are orthogonal.

In many situations, in practice, the explanatory variables may be not remain independent due to various reasons. The situations where the explanatory variables are higher ~~inco~~ intercorrelated is referred to as multicollinearity.

Consider the multiple regression model

$$y = X\beta + \varepsilon \qquad \varepsilon \sim N(0, \sigma^2 I)$$

with K explanatory variables $X_1, X_2, \ldots, X_k$ with usual assumptions including $\text{Rank}(X) = k$.

Assume the observations on all X_i's and Y_i's are centered and scaled to unit length. So.

- $X'X$ becomes a $k \times k$ matrix of correlation coefficients between the explanatory variables and
- $X'Y$ becomes a $k \times 1$ vector of correlation coeffient between explanatory and study variables.

Let $X = [X_1, X_2, \ldots, X_k]$ where X_j is the j^{th} column of X denoting the n observations on X_j. The column vectors $X_1, X_2, \ldots, X_k$ are linearly dependent if there exists a set of constants $l_1, l_2, \ldots, l_k$ not all zero, such that

$$\sum_{j=1}^{K} l_j K_j = 0$$

If this holds exactly for a subset of the $X_1, X_2, \cdots, X_K$ then Rank $(X'X) < K$. Consequently $(X'X)^{-1}$ does not exists. If the condition $\sum l_j X_j = 0$ is approximately true for some subset of $X_1, X_2, \cdots, X_K$, then there will be a near linear dependency in $X'X$. In such a case, the multicollinearity problem exists. It is also said that $X'X$ becomes ill-conditioned.