THE EFFECTS OF MULTICOLLINEARITY IN ORDINARY LEAST SQUARES (OLS) ESTIMATION

Weeraratne N.C.
Department of Economics & Statistics
SUSL, BelihulOya, Sri Lanka

ABSTRACT

The explanatory variables are not perfectly linearly correlated is one of the crucial assumption of Ordinary Least Squares (OLS) Estimation. So this paper explains the stability of OLS estimation when the explanatory variables are perfectly linearly correlated. This paper shows, If the correlation between the particular explanatory variables is perfect $(r_{x_i x_j} = 1)$, *then the estimates of the regression coefficients are undesirable and the standard deviation of the regression coefficients are infinitely large.*

Key Words: *OLS Estimation, Multicollinearity, Regression Coefficients*

1. INTRODUCTION

Statistical Inference refers to the process of selecting and using a sample to draw conclusion about the parameter of a population from which the sample is drawn [3]. It deals with two types of problems. Problem of Estimation: This problem arises when no information is available about the parameters of the population from which the sample is drawn. Statistics obtained from the sample are used to estimate the unknown parameter of the population from which the sample is drawn; Problem of Test of Hypothesis or Test of Significance: This problem arises when some information is available about the parameters of the population from which the sample is drawn and it is required to test how far this information about the population parameter is tenable in the light of the information provided by the sample. In the context of statistics, estimation is a statistical technique of estimating unknown population parameters from the corresponding sample statistic. A population parameter can be estimated in two ways. Point Estimation: It provides a single value of a statistic that is used to estimate an unknown population parameter; Interval Estimation: Interval estimation provides an interval of finite width centered at the point estimate of the parameter, within which unknown parameter is expected to lie with a specified probability, such an interval called a confidence interval for population parameter. The statistic which is used to obtain a point estimate is called estimator. The value of statistic is the estimate. For example, If sample regression coefficient$(\hat{\beta})$ is used for estimating the population regression coefficient(β), then $\hat{\beta}$ is called as estimator and the value of $\hat{\beta}$ is called an estimate [5]. According to R.A. Fisher [1,2], the criteria for a good estimator are; Unbiasedness: A statistic is said to be an unbiased estimator of a parameter if its expected value is equal to the value of the parameter; Consistency: A statistic is said to be a consistent estimator of a parameter if it come closer to the value of parameter as the sample size(n) tends to infinity; Efficiency: A consistent statistic is aid to be most efficient estimator of a parameter if its sampling variance is less than that of any other consistent estimator; Sufficiency: A statistic is said to be sufficient estimator of a parameter if it contains all information in the sample about the population parameter. Generally there two types of estimation used to obtain the point estimate. 1. Methods of Ordinary Least Squares (OLS) Estimation 2. Methods of Maximum Likelihood Estimation. In regression these two methods give similar results. Under certain assumptions of OLS has statistical properties that have made it one of the most powerful and popular method of regression analysis.

Ordinary Least Squared (OLS) Estimation

The method of minimizing the error sum of squared function is called Ordinary Least Squares (OLS) estimation. Error sum of squared is to be minimized with respect to estimators. The process of differentiation yields the estimators. Given the data the regression parameters can be estimated using the principle of least squares as before the estimates will be obtained by minimizing the sum of squared errors [4].

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2$$

A necessary condition for this expression to assume a minimum value is that its partial derivatives with respect to $\hat{\beta}_j; j = 0,1,2,\dots,p$ be equal to zero.

Assumptions of Ordinary Least Squares (OLS) Estimation:

1) e_i is a random real variable.
2) The mean value of e in any particular period is zero.
3) The variance of e_i is constant in each period.
4) The variable e_i has a normal distribution.
5) The random terms of different observations (e_i, e_j) are independent.
6) e is independent of the explanatory variable(s).
7) The explanatory variable(s) are measured without error.
8) The explanatory variables are not perfectly linearly correlated.

Model with two Explanatory Variables

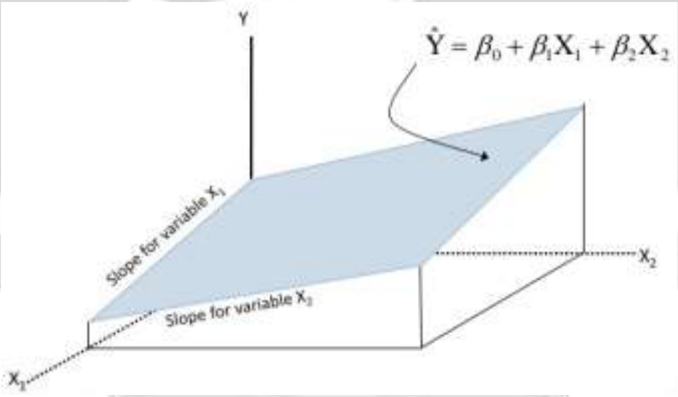

Figure 1: Two Variable Model

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2$$

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left(Y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i}\right)\right)^2$$

Performing the partial differentiations, we get the following system of 3 normal equations in the three unknown parameters $\hat{\beta}_0, \hat{\beta}_1$ and $\hat{\beta}_2$.

$$\frac{\partial \sum_{i=1}^{n} e_i^2}{\partial \hat{\beta}_0} = \frac{\partial \sum_{i=1}^{n}\left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{1i} - \hat{\beta}_2 X_{2i}\right)^2}{\partial \hat{\beta}_0} = 0$$

$$2\sum_{i=1}^{n}\left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{1i} - \hat{\beta}_2 X_{2i}\right)(-1) = 0$$

$$\sum_{i=1}^{n}\left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{1i} - \hat{\beta}_2 X_{2i}\right) = 0$$

$$\sum_{i=1}^{n} Y_i - n\hat{\beta}_0 - \sum_{i=1}^{n} \hat{\beta}_1 X_{1i} - \sum_{i=1}^{n} \hat{\beta}_2 X_{2i} = 0$$

$$\sum_{i=1}^{n} Y_i = n\hat{\beta}_0 + \sum_{i=1}^{n} \hat{\beta}_1 X_{1i} + \sum_{i=1}^{n} \hat{\beta}_2 X_{2i} \quad - - - - - - - - - - -(1)$$

$$\frac{\partial \sum_{i=1}^{n} e_i^2}{\partial \hat{\beta}_1} = \frac{\partial \sum_{i=1}^{n}\left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{1i} - \hat{\beta}_2 X_{2i}\right)^2}{\partial \hat{\beta}_1} = 0$$

$$2\sum_{i=1}^{n}\left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{1i} - \hat{\beta}_2 X_{2i}\right)(-X_{1i}) = 0$$

$$\sum_{i=1}^{n} X_{1i}\left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{1i} - \hat{\beta}_2 X_{2i}\right) = 0$$

$$\sum_{i=1}^{n} X_{1i} Y_i - \hat{\beta}_0 \sum_{i=1}^{n} X_{1i} - \hat{\beta}_1 \sum_{i=1}^{n} X_{1i}^2 - \hat{\beta}_2 \sum_{i=1}^{n} X_{1i} X_{2i} = 0$$

$$\sum_{i=1}^{n} X_{1i} Y_i = \hat{\beta}_0 \sum_{i=1}^{n} X_{1i} + \hat{\beta}_1 \sum_{i=1}^{n} X_{1i}^2 + \hat{\beta}_2 \sum_{i=1}^{n} X_{1i} X_{2i} - - - - - - - - - - - -(2)$$

$$\frac{\partial \sum_{i=1}^{n} e_i^2}{\partial \hat{\beta}_2} = \frac{\partial \sum_{i=1}^{n}\left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{1i} - \hat{\beta}_2 X_{2i}\right)^2}{\partial \hat{\beta}_2} = 0$$

$$2\sum_{i=1}^{n}\left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{1i} - \hat{\beta}_2 X_{2i}\right)(-X_{2i}) = 0$$

$$\sum_{i=1}^{n} X_{2i}\left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{1i} - \hat{\beta}_2 X_{2i}\right) = 0$$

$$\sum_{i=1}^{n} X_{2i} Y_i - \hat{\beta}_0 \sum_{i=1}^{n} X_{2i} - \hat{\beta}_1 \sum_{i=1}^{n} X_{1i} X_{2i} - \hat{\beta}_2 \sum_{i=1}^{n} X_{2i}^2 = 0$$

$$\sum_{i=1}^{n} X_{2i} Y_i = \hat{\beta}_0 \sum_{i=1}^{n} X_{2i} + \hat{\beta}_1 \sum_{i=1}^{n} X_{1i} X_{2i} + \hat{\beta}_2 \sum_{i=1}^{n} X_{2i}^2 - - - - - - - - - - - -(3)$$

Expressions (1), (2) and (3) are the three normal equations of the least squares method. The following formula in which the variables are expressed in deviations form their mean may also be used for obtaining values for the parameter estimates.

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + e_i - - - - - - - - - - - - - - -(4)$$

$$\bar{Y} = \beta_0 + \beta_1 \bar{X}_1 + \beta_2 \bar{X}_2 + \bar{e} - - - - - - - - - - - - - - - - (5)$$

$$(Y_i - \bar{Y}) = \beta_1 (X_{1i} - \bar{X}_1) + \beta_2 (X_{2i} - \bar{X}_2) + (e_i - \bar{e}) \qquad ;(4) - (5)$$

$$\downarrow \qquad\qquad \downarrow \qquad\qquad \downarrow \qquad\qquad \downarrow$$

$$y_i \quad x_{1i} \quad x_{2i} \quad e_i$$

$$y_i = \beta_1 x_{1i} + \beta_2 x_{2i} + e_i \text{-----------------}(6)$$

$$\hat{e}_i = \hat{y}_i - \hat{\beta}_1 x_{1i} - \hat{\beta}_2 x_{2i}$$

$$\hat{e}_i^2 = \left(\hat{y}_i - \hat{\beta}_1 x_{1i} - \hat{\beta}_2 x_{2i}\right)^2$$

$$\sum_{i=1}^{n} \hat{e}_i^2 = \sum_{i=1}^{n} \left(\hat{y}_i - \hat{\beta}_1 x_{1i} - \hat{\beta}_2 x_{2i}\right)^2 \text{-------------}(7)$$

$$\frac{\partial \sum_{i=1}^{n} \hat{e}_i^2}{\partial \hat{\beta}_1} = 2\sum_{i=1}^{n} \left(\hat{y}_i - \hat{\beta}_1 x_{1i} - \hat{\beta}_2 x_{2i}\right)(-x_{1i}) = 0$$

$$\sum_{i=1}^{n} \hat{y}_i x_{1i} - \hat{\beta}_1 \sum_{i=1}^{n} x_{1i}^2 - \hat{\beta}_2 \sum_{i=1}^{n} x_{1i} x_{2i} = 0$$

$$\sum_{i=1}^{n} \hat{y}_i x_{1i} = \hat{\beta}_1 \sum_{i=1}^{n} x_{1i}^2 + \hat{\beta}_2 \sum_{i=1}^{n} x_{1i} x_{2i} \text{------------}(8)$$

$$\frac{\partial \sum_{i=1}^{n} \hat{e}_i^2}{\partial \hat{\beta}_2} = 2\sum_{i=1}^{n} \left(\hat{y}_i - \hat{\beta}_1 x_{1i} - \hat{\beta}_2 x_{2i}\right)(-x_{2i}) = 0$$

$$\sum_{i=1}^{n} \hat{y}_i x_{2i} - \hat{\beta}_1 \sum_{i=1}^{n} x_{1i} x_{2i} - \hat{\beta}_2 \sum_{i=1}^{n} x_{2i}^2 = 0$$

$$\sum_{i=1}^{n} \hat{y}_i x_{2i} = \hat{\beta}_1 \sum_{i=1}^{n} x_{1i} x_{2i} + \hat{\beta}_2 \sum_{i=1}^{n} x_{2i}^2 \text{------------}(9)$$

$$\begin{bmatrix} \sum_{i=1}^{n} x_{1i}^2 & \sum_{i=1}^{n} x_{1i} x_{2i} \\ \sum_{i=1}^{n} x_{1i} x_{2i} & \sum_{i=1}^{n} x_{2i}^2 \end{bmatrix} \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n} \hat{y}_i x_{1i} \\ \sum_{i=1}^{n} \hat{y}_i x_{2i} \end{bmatrix}$$

By using Cramer's Rule;

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}_1 - \hat{\beta}_2 \bar{X}_2 \text{-------------------------}(10)$$

$$\hat{\beta}_1 = \begin{vmatrix} \sum_{i=1}^{n} \hat{y}_i x_{1i} & \sum_{i=1}^{n} x_{1i} x_{2i} \\ \sum_{i=1}^{n} \hat{y}_i x_{2i} & \sum_{i=1}^{n} x_{2i}^2 \end{vmatrix} \div \begin{vmatrix} \sum_{i=1}^{n} x_{1i}^2 & \sum_{i=1}^{n} x_{1i} x_{2i} \\ \sum_{i=1}^{n} x_{1i} x_{2i} & \sum_{i=1}^{n} x_{2i}^2 \end{vmatrix}$$

$$\hat{\beta}_1 = \frac{(\sum_{i=1}^{n} \hat{y}_i x_{1i})(\sum_{i=1}^{n} x_{2i}^2) - (\sum_{i=1}^{n} \hat{y}_i x_{2i})(\sum_{i=1}^{n} x_{1i} x_{2i})}{(\sum_{i=1}^{n} x_{1i}^2)(\sum_{i=1}^{n} x_{2i}^2) - (\sum_{i=1}^{n} x_{1i} x_{2i})^2} \text{-----------}(11)$$

$$\hat{\beta}_2 = \begin{vmatrix} \sum_{i=1}^{n} x_{1i}^2 & \sum_{i=1}^{n} \hat{y}_i x_{1i} \\ \sum_{i=1}^{n} x_{1i}x_{2i} & \sum_{i=1}^{n} \hat{y}_i x_{2i} \end{vmatrix} \div \begin{vmatrix} \sum_{i=1}^{n} x_{1i}^2 & \sum_{i=1}^{n} x_{1i}x_{2i} \\ \sum_{i=1}^{n} x_{1i}x_{2i} & \sum_{i=1}^{n} x_{2i}^2 \end{vmatrix}$$

$$\hat{\beta}_2 = \frac{(\sum_{i=1}^{n} \hat{y}_i x_{2i})(\sum_{i=1}^{n} x_{1i}^2) - (\sum_{i=1}^{n} \hat{y}_i x_{1i})(\sum_{i=1}^{n} x_{1i}x_{2i})}{(\sum_{i=1}^{n} x_{1i}^2)(\sum_{i=1}^{n} x_{2i}^2) - (\sum_{i=1}^{n} x_{1i}x_{2i})^2} ----------- (12)$$

The variances of the parameter estimates:

$$var(\hat{\beta}_0) = \hat{\sigma}_e^2 \left[\frac{1}{n} + \frac{\bar{X}_1^2 \sum_{i=1}^{n} x_{2i}^2 + \bar{X}_2^2 \sum_{i=1}^{n} x_{1i}^2 - 2\bar{X}_1\bar{X}_2 \sum_{i=1}^{n} x_{1i}x_{2i}}{\sum_{i=1}^{n} x_{1i}^2 \sum_{i=1}^{n} x_{2i}^2 - (\sum_{i=1}^{n} x_{1i}x_{2i})^2}\right]$$

$$var(\hat{\beta}_1) = \hat{\sigma}_e^2 \left[\frac{\sum_{i=1}^{n} x_{2i}^2}{\sum_{i=1}^{n} x_{1i}^2 \sum_{i=1}^{n} x_{2i}^2 - (\sum_{i=1}^{n} x_{1i}x_{2i})^2}\right]$$

$$var(\hat{\beta}_2) = \hat{\sigma}_e^2 \left[\frac{\sum_{i=1}^{n} x_{1i}^2}{\sum_{i=1}^{n} x_{1i}^2 \sum_{i=1}^{n} x_{2i}^2 - (\sum_{i=1}^{n} x_{1i}x_{2i})^2}\right]$$

$where;\ \hat{\sigma}_e^2 = {\sum_{i=1}^{n} e_i^2}/{n-k}\,;\ k = no{:}\,of\ parameters$

2. RESULTS AND FINDINGS

When more than one explanatory variable are used to estimate the dependent variable then this process is known as Multiple Linear Regression and Correlation analysis. In here this study takes statistical model for Multiple Linear Regression with two explanatory variables.

Table 1: Calculations of Multiple Linear Regression Analysis

Y	X_1	X_2	y_i	x_{1i}	x_{2i}	y_i^2	x_{1i}^2	x_{2i}^2	y_ix_{1i}	y_ix_{2i}	$x_{1i}x_{2i}$
1000	5	50	200	-1	-10	40000	1	100	-200	-2000	10
750	7	70	-50	1	10	2500	1	100	-50	-500	10
800	6	60	0	0	0	0	0	0	0	0	0
700	6	60	-100	0	0	10000	0	0	0	0	0
500	8	80	-300	2	20	90000	4	400	-600	-6000	40
650	7	70	-150	1	10	22500	1	100	-150	-1500	10
900	5	50	100	-1	-10	10000	1	100	-100	-1000	10
1000	4	40	200	-2	-20	40000	4	400	-400	-4000	40
1100	3	30	300	-3	-30	90000	9	900	-900	-9000	90
600	9	90	-200	3	30	40000	9	900	-600	-6000	90
8000	**60**	**600**	**0**	**0**	**0**	**345000**	**30**	**3000**	**-3000**	**-30000**	**300**

```
> cor(data)
             Y           X1          X2
Y     1.0000000  -0.9325048  -0.9325048
X1   -0.9325048   1.0000000   1.0000000
X2   -0.9325048   1.0000000   1.0000000
```

X_1 and X_2 are related with the exact relation $X_2 = 10.X_1$. $((\sum_{i=1}^{n} x_{2i}^2) = 10^2(\sum_{i=1}^{n} x_{1i}^2))$

Estimation of the coefficients of $\hat{\beta}_1$ and $\hat{\beta}_2$;

$$\hat{\beta}_1 = \frac{(\sum_{i=1}^{n} \hat{y}_i x_{1i})(\sum_{i=1}^{n} x_{2i}^2) - (\sum_{i=1}^{n} \hat{y}_i x_{2i})(\sum_{i=1}^{n} x_{1i}x_{2i})}{(\sum_{i=1}^{n} x_{1i}^2)(\sum_{i=1}^{n} x_{2i}^2) - (\sum_{i=1}^{n} x_{1i}x_{2i})^2}$$

$$\hat{\beta}_1 = \frac{(-3000)(3000) - (-30000)(300)}{(30)(3000) - (300)^2}$$

$$\boldsymbol{\hat{\beta}_1} = \frac{\mathbf{0}}{\mathbf{0}}$$

$$\hat{\beta}_2 = \frac{(\sum_{i=1}^{n} \hat{y}_i x_{2i})(\sum_{i=1}^{n} x_{1i}^2) - (\sum_{i=1}^{n} \hat{y}_i x_{1i})(\sum_{i=1}^{n} x_{1i}x_{2i})}{(\sum_{i=1}^{n} x_{1i}^2)(\sum_{i=1}^{n} x_{2i}^2) - (\sum_{i=1}^{n} x_{1i}x_{2i})^2}$$

$$\hat{\beta}_2 = \frac{(-30000)(30) - (-3000)(300)}{(30)(3000) - (300)^2}$$

$$\boldsymbol{\hat{\beta}_2} = \frac{\mathbf{0}}{\mathbf{0}}$$

The variances of the parameter estimates $\hat{\beta}_1$ and $\hat{\beta}_2$:

$$var(\hat{\beta}_1) = \hat{\sigma}_e^2\left[\frac{\sum_{i=1}^{n} x_{2i}^2}{\sum_{i=1}^{n} x_{1i}^2 \sum_{i=1}^{n} x_{2i}^2 - (\sum_{i=1}^{n} x_{1i}x_{2i})^2}\right]$$

$$var(\hat{\beta}_1) = \hat{\sigma}_e^2\left[\frac{3000}{(30)(3000) - (300)^2}\right]$$

$$\boldsymbol{var(\hat{\beta}_1) = \hat{\sigma}_e^2\left[\frac{3000}{0}\right]} = \infty$$

$$var(\hat{\beta}_2) = \hat{\sigma}_e^2\left[\frac{\sum_{i=1}^{n} x_{1i}^2}{\sum_{i=1}^{n} x_{1i}^2 \sum_{i=1}^{n} x_{2i}^2 - (\sum_{i=1}^{n} x_{1i}x_{2i})^2}\right]$$

$$var(\hat{\beta}_2) = \hat{\sigma}_e^2\left[\frac{30}{(30)(3000) - (300)^2}\right]$$

$$\boldsymbol{var(\hat{\beta}_2) = \hat{\sigma}_e^2\left[\frac{30}{0}\right]} = \infty$$

3. CONCLUSION

If the correlation between the explanatory variables is perfectly linearly correlated$(r_{x_ix_j} = 1\ or\ X_i = kX_j; i \neq j\ , k\ is\ constant)$, there is no proper way to find separate estimates of each regression coefficients. Due to this situation, the standard errors of these estimates also become infinitely large. That means, if the correlation between the explanatory variables is perfectly linearly correlated, the Ordinary Least Squares (OLS) Method also breaks down.

REFERENCES

[1]. FISHER, R. A. (1925) 1950 Theory of Statistical Estimation. Pages 11.699a-ll.725 in R. A. Fisher,Contributions to Mathematical Statistics. New York: Wiley.

[2]. FISHER, R. A. (1922) 1950 On the Mathematical Foundations of Theoretical Statistics. Pages 10.308a-10.368 in R. A. Fisher, Contributions to Mathematical Statistics. New York: Wiley.

[3]. Tanner, M. A. (1991). Tools for statistical inference (Vol. 3). New York: Springer.

[4]. Dismuke, C., & Lindrooth, R. (2006). Ordinary least squares. Methods and Designs for Outcomes Research, 93, 93-104.

[5]. Neter, J., Kutner, M. H., Nachtsheim, C. J., & Wasserman, W. (1996).Applied linear statistical models (Vol. 4, p. 318). Chicago: Irwin.