LINEAR REGRESSION ANALYSIS

MODULE – IX

Lecture - 28

Multicollinearity

Dr. Shalabh

Department of Mathematics and Statistics

Indian Institute of Technology Kanpur

A basic assumption in multiple linear regression model is that the rank of the matrix of observations on explanatory variables is same as the number of explanatory variables. In other words, such 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 not remain independent due to various reasons. The situation where the explanatory variables are highly intercorrelated is referred to as **multicollinearity.**

Consider the multiple regression model

$$y = \underset{n\times k}{X}\ \underset{k\times 1}{\beta} + \underset{n\times 1}{\varepsilon},\ \varepsilon \sim N(0, \sigma^2 I)$$

with k explanatory variables $X_1, X_2, \ldots, X_k$ with usual assumptions including 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 coefficients 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 $\ell_1, \ell_2, \ldots, \ell_k$, not all zero, such that

$$\sum_{j=1}^{k} \ell_j X_j = 0.$$

If this holds exactly for a subset of the $X_1, X_2, \ldots, X_k$, then rank ($X'X$) < k. Consequently ($X'X$)$^{-1}$ does not exist. If the condition $\sum_{j=1}^{k} \ell_j X_j = 0$ is approximately true for some subset of $X_1, X_2, \ldots, X_k$, 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.**

Source of multicollinearity

1. Method of data collection

It is expected that the data is collected over the whole cross-section of variables. It may happen that the data is collected over a subspace of the explanatory variables where the variables are linearly dependent. For example, sampling is done only over a limited range of explanatory variables in the population.

2. Model and population constraints

There may exists some constraints on the model or on the population from where the sample is drawn. The sample may be generated from that part of population having linear combinations.

3. Existence of identities or definitional relationships

There may exist some relationships among the variables which may be due to the definition of variables or any identity relation among them. For example, if data is collected on the variables like income, saving and expenditure, then

income = saving + expenditure.

Such relationship will not change even when the sample size increases.

4. Imprecise formulation of model

The formulation of the model may unnecessarily be complicated. For example, the quadratic (or polynomial) terms or cross product terms may appear as explanatory variables. For example, let there be 3 variables X_1, X_2 and X_3, so $k = 3$. Suppose their cross-product terms $X_1 X_2$, $X_2 X_3$ and $X_1 X_3$ are also added. Then k rises to 6.

5. An over-determined model

Sometimes, due to over enthusiasm, large number of variables are included in the model to make it more realistic and consequently the number of observations (n) becomes smaller than the number of explanatory variables (k). Such situation can arise in medical research where the number of patients may be small but information is collected on a large number of variables. In another example, if there is time series data for 50 years on consumption pattern, then it is expected that the consumption pattern does not remain same for 50 years. So better option is to choose smaller number of variables and hence it results into $n < k$. But this is not always advisable. For example, in microarray experiments, it is not advisable to choose smaller number of variables.

Consequences of multicollinearity

To illustrate the consequences of presence of multicollinearity, consider a model

$$y = \beta_1 x_1 + \beta_2 x_2 + \varepsilon,\ E(\varepsilon) = 0,\ V(\varepsilon) = \sigma^2 I$$

where x_1, x_2 and y are scaled to length unity.

The normal equation $(X'X)b = X'y$ in this model becomes

$$\begin{pmatrix} 1 & r \\ r & 1 \end{pmatrix}\begin{pmatrix} b_1 \\ b_2 \end{pmatrix} = \begin{pmatrix} r_{1y} \\ r_{2y} \end{pmatrix}$$

where r is the correlation coefficient between x_1 and x_2; r_{jy} is the correlation coefficient between x_j and y (j = 1, 2) and $b = (b_1 \quad b_2)'$ is the OLSE of β .

$$(X'X)^{-1} = \left(\frac{1}{1-r^2}\right)\begin{bmatrix} 1 & -r \\ -r & 1 \end{bmatrix}$$

$$\Rightarrow b_1 = \frac{r_{1y} - r\, r_{2y}}{1-r^2}$$

$$b_2 = \frac{r_{2y} - r\, r_{1y}}{1-r^2}.$$

So the covariance matrix is $V(b) = \sigma^2 (X'X)^{-1}$

$$\Rightarrow Var(b_1) = Var(b_2) = \frac{\sigma^2}{1-r^2}$$

$$Cov(b_1, b_2) = -\frac{r\sigma^2}{1-r^2}.$$

If x_1 and x_2 are uncorrelated, then $r = 0$ and

$$X'X = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

rank($X'X$) = 2.

If x_1 and x_2 are perfectly correlated, then $r = \pm 1$ and rank($X'X$) = 1.

If $r \to \pm 1$, then

$$Var(b_1) = Var(b_2) \to \infty.$$

So if variables are perfectly collinear, the variance of OLSEs becomes large. This indicates highly unreliable estimates and this is an inadmissible situation.

Consider the following result

r	0.99	0.9	0.1	0
$Var(b_1) = Var(b_2)$	$50\sigma^2$	$5\sigma^2$	$1.01\sigma^2$	σ^2

The standard errors of b_1 and b_2 rise sharply as $r \to \pm 1$ and they break down at $r = \pm 1$ because $X'X$ becomes non-singular.

- If r is close to 0, then multicollinearity does not harm and it is termed as **non-harmful multicollinearity** .
- If r is close to +1 or -1 then multicollinearity inflates the variance and it rises terribly. This is termed as **harmful multicollinearity.**

There is no clear cut boundary to distinguish between the harmful and non-harmful multicollinearity. Generally, if *r* is low, the multicollinearity is considered as non-harmful and if *r* is high, the multicollinearity is considered as harmful.

In case of near or high multicollinearity, following possible consequences are encountered.

1. The OLSE remains an unbiased estimator of β but its sampling variance becomes very large. So OLSE becomes imprecise and property of BLUE does not hold anymore.
2. Due to large standard errors, the regression coefficients may not appear significant. Consequently, important variables may be dropped.

   For example, to test $H_0 : \beta_1 = 0,$ we use *t* - ratio as

   $$t_0 = \frac{b_1}{\sqrt{\widehat{Var}(b_1)}}.$$

   Since $\widehat{Var}(b_1)$ is large, so t_0 is small and consequently H_0 is more often accepted.

   Thus harmful multicollinearity intends to delete important variables.
3. Due to large standard errors, the large confidence region may arise. For example, the confidence interval is given by $\left(b_1 \pm t_{\frac{\alpha}{2}, n-1} \sqrt{\widehat{Var}(b_1)} \right)$. When $\widehat{Var}(b_1)$ becomes large, then confidence interval becomes wider.
4. The OLSE may be sensitive to small changes in the values of explanatory variables. If some observations are added or dropped, OLSE may change considerably in magnitude as well as in sign. Ideally, OLSE should not change with inclusion or deletion of few observations. Thus OLSE loses stability and robustness.

When the number of explanatory variables are more than two, say k as $X_1, X_2,\ldots,X_k$ then the j^{th} diagonal element of $C = (X'X)^{-1}$ is

$$C_{jj} = \frac{1}{1-R_j^2}$$

where R_j^2 is the multiple correlation coefficient or coefficient of determination from the regression of X_j on the remaining (k - 1) explanatory variables.

If X_j is highly correlated with any subset of other (k - 1) explanatory variables then R_j^2 is high and close to 1.

Consequently variance of j^{th} OLSE

$$Var(b_j) = C_{jj}\sigma^2 = \frac{\sigma^2}{1-R_j^2}$$

becomes very high. The covariance between b_i and b_j will also be large if X_i and X_j are involved in the linear relationship leading to multicollinearity.

The least squares estimates b_j become too large in absolute value in the presence of multicollinearity. For example, consider the squared distance between b and β as

$$L^2 = (b-\beta)'(b-\beta)$$

$$E(L^2) = \sum_{j=1}^{k} E(b_j - \beta_j)^2$$

$$= \sum_{j=1}^{k} Var(b_j)$$

$$= \sigma^2 tr(X'X)^{-1}.$$

The trace of a matrix is same as the sum of its eigenvalues. If $\lambda_1, \lambda_2, ..., \lambda_k$ are the eigenvalues of $(X'X)$, then $\frac{1}{\lambda_1}, \frac{1}{\lambda_2}, ..., \frac{1}{\lambda_k}$ are the eigenvalues of $(X'X)^{-1}$ and hence

$$E(L^2) = \sigma^2 \sum_{j=1}^{k} \frac{1}{\lambda_j}, \ \lambda_j > 0.$$

If $(X'X)$ is ill-conditioned due to the presence of multicollinearity then at least one of the eigenvalue will be small. So the distance between b and β may also be large. Thus

$$E(L^2) = E(b-\beta)'(b-\beta)$$

$$\sigma^2 tr(X'X)^{-1} = E(b'b - 2b'\beta + \beta'\beta)$$

$$\Rightarrow E(b'b) = \sigma^2 tr(X'X)^{-1} + \beta'\beta$$

$\Rightarrow b$ is generally larger in magnitude than β.

$\Rightarrow$ OLSE are too large in absolute value.

The least squares produces bad estimates of parameters in the presence of multicollinearity. This does not imply that the fitted model produces bad predictions also. If the predictions are confined to *X-space* with non-harmful multicollinearity, then predictions are satisfactory.