LINEAR REGRESSION ANALYSIS

MODULE – IX

Lecture - 29

Multicollinearity

Dr. Shalabh

Department of Mathematics and Statistics

Indian Institute of Technology Kanpur

Multicollinearity diagnostics

An important question that arises is how to diagnose the presence of multicollinearity in the data on the basis of given sample information. Several diagnostic measures are available and each of them is based on a particular approach. It is difficult to say that which of the diagnostic is the best or ultimate. Some of the popular and important diagnostics are described further.

The detection of multicollinearity involves 3 aspects:

(i) Determining its presence.

(ii) Determining its severity.

(iii) Determining its form or location.

1. Determinant of $X'X$ $(|X'X|)$

This measure is based on the fact that the matrix $X'X$ becomes ill conditioned in the presence of multicollinearity. The value of determinant of $X'X$, i.e., $|X'X|$ declines as degree of multicollinearity increases.

If rank $(X'X) < k$ then $|X'X|$ will be singular and so $|X'X| = 0$ So as $|X'X| \to 0$, the degree of multicollinearity increases and it becomes exact or perfect at $|X'X| = 0$. Thus $|X'X|$ serves as a measure of multicollinearity and $|X'X| = 0$ indicates that perfect multicollinearity exists.

Limitations:

This measure has following limitations

i. It is not bounded as $0 < |X'X| < \infty$.

ii. It is affected by dispersion of explanatory variables. For example, if $k = 2$, then

$$|X'X| = \begin{vmatrix} \sum_{i=1}^{n} x_{1i}^2 & \sum_{i=1}^{n} x_{1i}x_{2i} \\ \sum_{i=1}^{n} x_{2i}x_{1i} & \sum_{i=1}^{n} x_{2i}^2 \end{vmatrix}$$

$$= \left(\sum_{i=1}^{n} x_{1i}^2\right)\left(\sum_{i=1}^{n} x_{2i}^2\right)\left(1 - r_{12}^2\right)$$

where r_{12} is the correlation coefficient between X_1 and X_2. So $|X'X|$ depends on correlation coefficient and variability of explanatory variable. If explanatory variables have very low variability, then $|X'X|$ may tend to zero which will indicate the presence of multicollinearity and which is not the case so.

iii. It gives no idea about the relative effects on individual coefficients. If multicollinearity is present, then it will not indicate that which variable in $|X'X|$ is causing multicollinearity and is hard to determine.

2. Inspection of correlation matrix

The inspection of off-diagonal elements r_{ij} in $X'X$ gives an idea about the presence of multicollinearity. If X_i and X_j are nearly linearly dependent then $|r_{ij}|$ will be close to 1. Note that the observations in X are standardized in the sense that each observation is subtracted from mean of that variable and divided by the square root of corrected sum of squares of that variable.

When more than two explanatory variables are considered and if they are involved in near-linear dependency, then it is not necessary that any of the r_{ij} will be large. Generally, pairwise inspection of correlation coefficients is not sufficient for detecting multicollinearity in the data.

3. Determinant of correlation matrix

Let D be the determinant of correlation matrix then $0 \leq D \leq 1$.

If $D = 0$ then it indicates the existence of exact linear dependence among explanatory variables.

If $D = 1$ then the columns of X matrix are orthonormal.

Thus a value close to 0 is an indication of high degree of multicollinearity. Any value of D between 0 and 1 gives an idea of the degree of multicollinearity.

Limitation:

It gives no information about the number of linear dependencies among explanatory variables.

Advantages over $|X'X|$

(i) It is a bounded measure $0 \le D \le 1.$

(ii) It is not affected by the dispersion of explanatory variables. For example, when *k* = 2,

$$|X'X| = \begin{vmatrix} \sum_{i=1}^{n} x_{1i}^2 & \sum_{i=1}^{n} x_{1i}x_{2i} \\ \sum_{i=1}^{n} x_{1i}x_{2i} & \sum_{i=1}^{n} x_{2i}^2 \end{vmatrix} = (1 - r_{12}^2).$$

4. Measure based on partial regression

A measure of multicollinearity can be obtained on the basis of coefficients of determination based on partial regression. Let R^2 be the coefficient of determination in the full model, i.e., based on all explanatory variables and R_j^2 be the coefficient of determination in the model when i^{th} explanatory variable is dropped, *i*=1,2,…,*k,* and $R_L^2 = Max(R_1^2, R_2^2, ..., R_k^2)$.

Procedure:

i. Drop one of the explanatory variable among k variables, say X_1.

ii. Run regression of y over rest of the (k - 1) variables $X_2, X_3, \ldots, X_k$.

iii. Calculate R_1^2.

iv. Similarly calculate $R_2^2, R_3^2, ..., R_k^2$

v. Find $R_L^2 = Max(R_1^2, R_2^2, ..., R_k^2)$.

vi. Determine $R^2 - R_L^2$.

The quantity $(R^2 - R_L^2)$ provides a measure of multicollinearity. If multicollinearity is present, R_L^2 will be high. Higher the degree of multicollinearity, higher the value of R_L^2. So in the presence of multicollinearity, $(R^2 - R_L^2)$ be low.

Thus if $(R^2 - R_L^2)$ is close to 0, it indicates the high degree of multicollinearity.

Limitations:

i. It gives no information about the underlying relations about explanatory variables, i.e., how many relationships are present or how many explanatory variables are responsible for the multicollinearity.

ii. Small value of $(R^2 - R_L^2)$ may occur because of poor specification of the model also and it may be inferred in such situation that multicollinearity is present.

5. Variance inflation factors (*VIF*)

The matrix $X'X$ becomes ill-conditioned in the presence of multicollinearity in the data. So the diagonal elements of $C = (X'X)^{-1}$ helps in the detection of multicollinearity. If R_j^2 denotes the coefficient of determination obtained when X_j is regressed on the remaining (k - 1) variables excluding X_j, then the j^{th} diagonal element of C is

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

If X_j is nearly orthogonal to remaining explanatory variables, then R_j^2 is small and consequently C_{jj} is close to 1.

If X_j is nearly linearly dependent on a subset of remaining explanatory variables, then R_j^2 is close to 1 and consequently C_{jj} is large.

Since the variance of j^{th} OLSE of β_j is $Var(b_j) = \sigma^2 C_{jj}$.

So C_{jj} is the factor by which the variance of b_j increases when the explanatory variables are near linear dependent. Based on this concept, the variance inflation factor for the j^{th} explanatory variable is defined as

$$VIF_j = \frac{1}{1 - R_j^2}.$$

This is the factor which is responsible for inflating the sampling variance. The combined effect of dependencies among the explanatory variables on the variance of a term is measured by the *VIF* of that term in the model.

One or more large *VIFs* indicate the presence of multicollinearity in the data.

In practice, usually a $VIF > 5$ or 10 indicates that the associated regression coefficients are poorly estimated because of multicollinearity. If regression coefficients are estimated by OLSE and its variance is $\sigma^2 (X'X)^{-1}$. So *VIF* indicates that a part of this variance is given by VIF_j.

Limitations:

(i) It sheds no light on the number of dependencies among the explanatory variables.

(ii) The rule of *VIF* > 5 or 10 is a rule of thumb which may differ from one situation to another situation.

Another interpretation of VIF_j

The *VIF*s can also be viewed as follows.

The confidence interval of j^{th} OLSE of β_j is given by

$$\left(b \pm \sqrt{\hat{\sigma}^2 C_{jj}}\, t_{\frac{\alpha}{2}, n-k-1} \right).$$

The length of the confidence interval is

$$L_j = 2\sqrt{\hat{\sigma}^2 C_{jj}}\, t_{\frac{\alpha}{2}, n-k-1} \quad .$$

Now consider a situation where *X* is an orthogonal matrix, i.e., $X'X = I$ so that $C_{jj} = 1$, sample size is same as earlier and same root mean squares $\left(\frac{1}{n} \sum_{i=1}^{n} (x_{ij} - \bar{x}_j)^2 \right)$, then the length of confidence interval becomes $L^* = 2\hat{\sigma} t_{\frac{\alpha}{2}, n-k-1}$.

Consider the ratio $\frac{L_j}{L^*} = \sqrt{C_{jj}}$.

Thus $\sqrt{VIF_j}$ indicates the increase in the length of confidence interval of j^{th} regression coefficient due to the presence of multicollinearity.

6. Condition number and condition index

Let $\lambda_1, \lambda_2, ..., \lambda_k$ be the eigenvalues (or characteristic roots) of $X'X$. Let

$$\lambda_{\max} = Max(\lambda_1, \lambda_2, ..., \lambda_k)$$
$$\lambda_{\min} = Min(\lambda_1, \lambda_2, ..., \lambda_k).$$

The condition number *(CN)* is defined as

$$CN = \frac{\lambda_{\max}}{\lambda_{\min}},\ 0 < CN < \infty.$$

The small values of characteristic roots indicates the presence of near-linear dependencies in the data. The *CN* provides a measure of spread in the spectrum of characteristic roots of *X'X*.

The condition number provides a measure of multicollinearity.

- If *CN* < 100, then it is considered as **non-harmful multicollinearity**.
- If 100 < *CN* < 1000, then it indicates that the multicollinearity is moderate to severe (or strong). This range is referred to as **danger level.**
- If *CN* > 1000, then it indicates a **severe** (or **strong**) **multicollinearity**.

The condition number is based only or two eigenvalues: $\lambda_{\min}$ and $\lambda_{\max}$. Another measures are condition indices which use information on other eigenvalues as well.

The **condition indices** of *X'X* are defined as $C_j = \frac{\lambda_{\max}}{\lambda_j},\ j = 1, 2, ..., k.$

In fact, largest $C_j = CN$.

The number of condition indices that are large, say more than 1000, indicate the number of near-linear dependencies in *X'X*. A limitation of *CN* and C_j is that they are unbounded measures as $0 < CN < \infty$, $0 < C_j < \infty$.

7. Measure based on characteristic roots and proportion of variances

Let $\lambda_1, \lambda_2, .., \lambda_k$ be the eigenvalues of $X'X$, $\Lambda = diag(\lambda_1, \lambda_2, ..., \lambda_k)$ is $k \times k$ matrix and *V* is a $k \times k$ matrix constructed by the eigenvectors of *X'X*. Obviously, *V* is an orthogonal matrix. Then *X'X* can be decomposed as $X'X = V\Lambda V'$. Let $V_1, V_2, ..., V_k$ be the column of *V*. If there is near-linear dependency in the data, then λ_j is close to zero and the nature of linear dependency is described by the elements of associated eigenvector V_j.

The covariance matrix of OLSE is

$$\begin{aligned} V(b) &= \sigma^2 (X'X)^{-1} \\ &= \sigma^2 (V\Lambda V')^{-1} \\ &= \sigma^2 V \Lambda^{-1} V' \end{aligned}$$

$$\Rightarrow Var(b_i) = \sigma^2 \left(\frac{v_{i1}^2}{\lambda_1} + \frac{v_{i2}^2}{\lambda_2} + ... + \frac{v_{ik}^2}{\lambda_k} \right)$$

where $v_{i1}, v_{i2}, ..., v_{ik}$ are the elements in *V*.

The condition indices are

$$C_j = \frac{\lambda_{\max}}{\lambda_j}, \; j = 1, 2, ..., k.$$

Procedure:

i. Find condition index $C_1, C_2, ..., C_k$.

ii. (a) Identify those λ_i's for which C_j is greater than the danger level 1000.

(b) This gives the number of linear dependencies.

(c) Don't consider those C_j's which are below the danger level.

iii. For such λ's with condition index above the danger level, choose one such eigenvalue, say λ_j.

iv. Find the value of proportion of variance corresponding to λ_j in $Var(b_1), Var(b_2), ..., Var(b_k)$ as

$$p_{ij} = \frac{(v_{ij}^2 / \lambda_j)}{VIF_j} = \frac{v_{ij}^2 / \lambda_j}{\sum_{j=1}^{k} (v_{ij}^2 / \lambda_j)}.$$

Note that $\left(\frac{v_{ij}^2}{\lambda_j} \right)$ can be found from the expression

$$Var(b_i) = \sigma^2 \left(\frac{v_{i1}^2}{\lambda_1} + \frac{v_{i2}^2}{\lambda_2} + ... + \frac{v_{ik}^2}{\lambda_k} \right)$$

i.e., corresponding to j^{th} factor.

The proportion of variance p_{ij} provides a measure of multicollinearity.

If $p_{ij} > 0.5$, it indicates that b_i is adversely affected by the multicollinearity, i.e., estimate of β_i is influenced by the presence of multicollinearity.

It is a good diagnostic tool in the sense that it tells about the presence of harmful multicollinearity as well as also indicates the number of linear dependencies responsible for multicollinearity. This diagnostic is better than other diagnostics.

The condition indices are also defined by the singular value decomposition of X matrix as follows:

$$X = UDV'$$

where U is $n \times k$ matrix, V is matrix, is $k \times k$ matrix $U'U = I$, $V'V = I$, D is $p \times p$ matrix, $D = diag(\mu_1, \mu_2, ..., \mu_k)$ and $\mu_1, \mu_2, ..., \mu_k$ are the singular values of X, V is a matrix whose columns are eigenvectors corresponding to eigenvalues of $X'X$ and U is a matrix whose columns are the eigenvectors associated with the k nonzero eigenvalues of $X'X$.

The condition indices of X matrix are defined as

$$\eta_j = \frac{\mu_{\max}}{\mu_j}, \; j = 1, 2, ..., k$$

where $\mu_{\max} = Max(\mu_1, \mu_2, ..., \mu_k)$.

If $\lambda_1, \lambda_2, ..., \lambda_k$ are the eigenvalues of $X'X$ then

$$X'X = (UDV')'UDV' = VD^2V' = V\Lambda V',$$

so $\mu_j^2 = \lambda_j, \; j = 1, 2, ..., k.$

Note that with $\mu_j^2 = \lambda_j$,

$$Var(b_j) = \sigma^2 \sum_{i=1}^{k} \frac{v_{ji}^2}{\mu_i^2}$$

$$VIF_j = \sum_{i=1}^{k} \frac{v_{ji}^2}{\mu_i^2}$$

$$p_{ij} = \frac{(v_{ji}^2 / \mu_i^2)}{VIF_j}.$$

The ill-conditioning in X is reflected in the size of singular values. There will be one small singular value for each non-linear dependency. The extent of ill conditioning is described by how small is μ_j relative to $\mu_{\max}$.

It is suggested that the explanatory variables should be scaled to unit length but should not be centered when computing p_{ij}. This will help in diagnosing the role of intercept term in near-linear dependence. No unique guidance is available in literature on the issue of centering the explanatory variables. The centering makes the intercept orthogonal to explanatory variables. So this may remove the ill conditioning due to intercept term in the model.