Chapter 9

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 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 $X_1, X_2, ..., X_k$ 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 coefficients between explanatory and study variables.

Let $X = [X_1, X_2, ..., X_k]$ where X_j is the j^{th} column of X denoting the *n* observations on X_j. The column vectors $X_1, X_2, ..., X_k$ are linearly dependent if there exists a set of constants $\ell_1, \ell_2, ..., \ell_k$, not all zero, such that

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

If this holds exactly for a subset of the $X_1, X_2, ..., X_k$, then $\text{rank}(X'X) < k$. 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 $X_1, X_2, ..., X_k$, 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 exist some constraints on the model or on the population from where the sample is drawn. The sample may be generated from that part of the 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 a 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_1X_2, X_2X_3 and X_1X_3 are also added. Then k rises to 6.

5. An over-determined model

Sometimes, due to over-enthusiasm, a large number of variables are included in the model to make it more realistic. Consequently, the number of observations (n) becomes smaller than the number of explanatory variables (k). Such a situation can arise in medical research where the number of patients may be small, but the 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 the same for 50 years. So better option is to choose a smaller number of variables, and hence it results in $n < k$.

Consequences of multicollinearity

To illustrate the consequences of the 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 $b = (b_1, b_2)'$ 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

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

$$\text{rank}(X'X) = 2.$$

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

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 regarded as harmful.

In case of near or high multicollinearity, the 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, essential 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 $\widehat{Var}(b_1)$ it becomes large, then the 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 the inclusion or deletion of variables. Thus OLSE loses stability and robustness.

When the number of explanatory variables is more than two, say k as $X_1, X_2, ..., X_k$ 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 the 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, the 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

$$\begin{aligned} L^2 &= (b-\beta)'(b-\beta) \\ E(L^2) &= \sum_{j=1}^{k} E(b_j - \beta_j)^2 \\ &= \sum_{j=1}^{k} Var(b_j) \\ &= \sigma^2 tr(X'X)^{-1}. \end{aligned}$$

The trace of a matrix is the 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 substantial. Thus

$$E(L^2) = E(b-\beta)'(b-\beta)$$
$$\sigma^2 tr(X'X)^{-1} = E(b'b - 2b'\beta + \beta'\beta)$$
$$\Rightarrow E(b'b) = \sigma^2 tr(X'X)^{-1} + \beta'\beta$$
$$\Rightarrow b \text{ is generally longer than } \beta$$
$$\Rightarrow \text{OLSE is too large in absolute value.}$$

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

Multicollinearity diagnostics

An important question arises about 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 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 $X'X$ $\left(|X'X|\right)$

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

If $\text{Rank}(X'X) < k$ 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 the following limitations

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

(ii) It is affected by the 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 the correlation coefficient and variability of the 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 the mean of that variable and divided by the square root of the 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, a 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 the correlation matrix then $0 \le D \le 1$.

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 a 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_i^2 be the coefficient of determination in the model when the 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 variables among k variables, say X_1.

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

(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) A 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 *VIF*s 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

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

Now consider a situation where X is an orthogonal matrix, i.e., $X'X = I$ so that $C_{jj} = 1,$ sample size is the same as earlier and the 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 the 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 indicate the presence of near-linear dependency 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 the information on other eigenvalues.

The **condition indices** of $X'X$ are defined as

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

In fact, the 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 a $k \times k$ 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 a 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 the associated eigenvector V_j.

The covariance matrix of OLSE is

$$\begin{aligned} V(b) &= \sigma^2 (X'X)^{-1} \\ &= \sigma^2 (V\Lambda V')^{-1} \\ &= \sigma^2 V \Lambda^{-1} V' \\ \Rightarrow Var(b_i) &= \sigma^2 \left(\frac{v_{i1}^2}{\lambda_1} + \frac{v_{i2}^2}{\lambda_2} + ... + \frac{v_{ik}^2}{\lambda_k} \right) \end{aligned}$$

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 the 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., an 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 the matrix as follows:

$$X = UDV'$$

where U is $n \times k$ matrix, V is $k \times k$ matrix, $U'U = I$, $V'V = I$, D is $k \times k$ 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 $\mu_j^2 = \lambda_j$,

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

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

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

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 helps in diagnosing the role of intercept term in near-linear dependence. No unique guidance is available in the 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.

Remedies for multicollinearity:

Various techniques have been proposed to deal with the problems resulting from the presence of multicollinearity in the data.

1. Obtain more data

The harmful multicollinearity arises essentially because the rank of $X'X$ falls below k and $|X'X|$ is close to zero. Additional data may help in reducing the sampling variance of the estimates. The data need to be collected such that it helps in breaking up the multicollinearity in the data.

It is always not possible to collect additional data for various reasons as follows.

- The experiment and process have finished and no longer available.
- The economic constraints may also not allow collecting the additional data.
- The additional data may not match with the earlier collected data and may be unusual.
- If the data is in time series, then longer time series may force to take ignore data that is too far in the past.
- If multicollinearity is due to any identity or exact relationship, then increasing the sample size will not help.
- Sometimes, it is not advisable to use the data even if it is available. For example, if the data on consumption pattern is available for the years 1950-2010, then one may not like to use it as the consumption pattern usually does not remains the same for such a long period.

2. Drop some variables that are collinear:

If possible, identify the variables which seem to cause multicollinearity. These collinear variables can be dropped so as to match the condition of fall rank of X – matrix. The process of omitting the variables way be carried out on the basis of some kind of ordering of explanatory variables, e.g., those variables can be deleted first which have smaller value of t-ratio. In another example, suppose the experimenter is not interested in all the parameters. In such cases, one can get the estimators of the parameters of interest which have smaller mean squared errors them the variance of OLSE by dropping some variables.

If some variables are eliminated, then this may reduce the predictive power of the model. Sometimes there is no assurance of how the model will exhibit less multicollinearity.

3. Use some relevant prior information:

One may search for some relevant prior information about the regression coefficients. This may lead to the specification of estimates of some coefficients. The more general situation includes the specification of some exact linear restrictions and stochastic linear restrictions. The procedures like restricted regression and mixed regression can be used for this purpose. The relevance and correctness of information play an important role in such analysis, but it is challenging to ensure it in practice. For example, the estimates derived in the U.K. may not be valid in India.

4. Employ generalized inverse

If $\text{rank}(X'X) < k$, then the generalized inverse can be used to find the inverse of $X'X$. Then β can be estimated by $\hat{\beta} = (X'X)^{-}X'y$.

In such a case, the estimates will not be unique except in the case of use of Moore-Penrose inverse of $(X'X)$. Different methods of finding generalized inverse may give different results. So applied workers will get different results. Moreover, it is also not known that which method of finding generalized inverse is optimum.

5. Use of principal component regression

The principal component regression is based on the technique of principal component analysis. The k explanatory variables are transformed into a new set of orthogonal variables called principal components. Usually, this technique is used for reducing the dimensionality of data by retaining some levels of variability of explanatory variables which is expressed by the variability in the study variable. The principal components involve the determination of a set of linear combinations of explanatory variables such that they retain the total variability of the system, and these linear combinations are mutually independent of each other. Such obtained principal components are ranked in the order of their importance. The importance being judged in terms of variability explained by a principal component relative to the total variability in the system. The procedure then involves eliminating some of the principal components which contribute to explaining relatively less variation. After elimination of the least important principal components, the set up of multiple regression is used by replacing the explanatory variables with principal components. Then study variable is regressed against the set of selected principal components using the ordinary least squares method. Since all the principal components are orthogonal, they are mutually independent, and so OLS is used without any problem. Once the estimates of regression coefficients for the reduced set of orthogonal variables (principal

components) have been obtained, they are mathematically transformed into a new set of estimated regression coefficients that correspond to the original correlated set of variables. These new estimated coefficients are the principal components estimators of regression coefficients.

Suppose there are k explanatory variables $X_1, X_2, ..., X_k$. Consider the linear function of $X_1, X_2, .., X_k$ like

$$Z_1 = \sum_{i=1}^{k} a_i X_i$$

$$Z_2 = \sum_{i=1}^{k} b_i X_i \text{ etc.}$$

The constants $a_1, a_2, ..., a_k$ are determined such that the variance of Z_1 is maximized subject to the normalizing condition that $\sum_{i=1} a_i^2 = 1$. The constant $b_1, b_2, ..., b_k$ are determined such that the variance of Z_2 is maximized subject to the normality condition that $\sum_{i=1} b_i^2 = 1$ and is independent of the first principal component.

We continue with such process and obtain k such linear combinations such that they are orthogonal to their preceding linear combinations and satisfy the normality condition. Then we obtain their variances. Suppose such linear combinations are $Z_1, Z_2, .., Z_k$ and for them, $Var(Z_1) > Var(Z_2) > ... > Var(Z_k)$. The linear combination having the largest variance is the first principal component. The linear combination having the second largest variance is the second-largest principal component and so on. These principal components have the property that $\sum_{i=1}^{k} Var(Z_i) = \sum_{i=1}^{k} Var(X_i)$. Also, the $X_1, X_2, ..., X_k$ are correlated but $Z_1, Z_2, .., Z_k$ are orthogonal or uncorrelated. So there will be zero multicollinearity among $Z_1, Z_2, .., Z_k$.

The problem of multicollinearity arises because $X_1, X_2, ..., X_k$ are not independent. Since the principal components based on $X_1, X_2, ..., X_k$ are mutually independent, so they can be used as explanatory variables, and such regression will combat the multicollinearity.

Let $\lambda_1, \lambda_2, ..., \lambda_k$ be the eigenvalues of $X'X$, $\Lambda = diag(\lambda_1, \lambda_2, ..., \lambda_k)$ is $k \times k$ diagonal matrix, V is a $k \times k$ orthogonal matrix whose columns are the eigenvectors associated with $\lambda_1, \lambda_2, ..., \lambda_k$. Consider the canonical form of the linear model

$$\begin{aligned} y &= X\beta + \varepsilon \\ &= XVV'\beta + \varepsilon \\ &= Z\alpha + \varepsilon \end{aligned}$$

where $Z = XV, \alpha = V'\beta, V'X'XV = Z'Z = \Lambda$.

Columns of $Z = (Z_1, Z_2, ..., Z_k)$ define a new set of explanatory variables which are called as **principal components.**

The OLSE of α is

$$\begin{aligned} \hat{\alpha} &= (Z'Z)^{-1}Z'y \\ &= \Lambda^{-1}Z'y \end{aligned}$$

and its covariance matrix is

$$\begin{aligned} V(\hat{\alpha}) &= \sigma^2 (Z'Z)^{-1} \\ &= \sigma^2 \Lambda^{-1} \\ &= \sigma^2 diag\left(\frac{1}{\lambda_1}, \frac{1}{\lambda_2}, ..., \frac{1}{\lambda_k}\right) \end{aligned}$$

Note that λ_j is the variance of j^{th} principal component and $Z'Z = \sum_{i=1}^{k}\sum_{j=1}^{k} Z_i Z_j = \Lambda$. A small eigenvalue of $X'X$ means that the linear relationship between the original explanatory variable exists and the variance of the corresponding orthogonal regression coefficient is large, which indicates that the multicollinearity exists. If one or more λ_j is small, then it indicates that multicollinearity is present.

Retainment of principal components:

The new set of variables, i.e., principal components are orthogonal, and they retain the same magnitude of variance as of the original set. If multicollinearity is severe, then there will be at least one small value of eigenvalue. The elimination of one or more principal components associated with the smallest eigenvalues will reduce the total variance in the model. Moreover, the principal components responsible for creating multicollinearity will be removed, and the resulting model will be appreciably improved.

The principal component matrix $Z = [Z_1, Z_2, ..., Z_k]$ with $Z_1, Z_2, ..., Z_k$ contains precisely the same information as the original data in X in the sense that the total variability in X and Z is the same. The difference between them is that the original data are arranged into a set of new variables which are uncorrelated with each other and can be ranked with respect to the magnitude of their eigenvalues. The j^{th} column vector Z_j corresponding to the largest λ_j accounts for the largest proportion of the variation in the original data. Thus the Z_j's are indexed so that $\lambda_1 > \lambda_2 > ... > \lambda_k > 0$ and λ_j is the variance of Z_j.

A strategy of elimination of principal components is to begin by discarding the component associated with the smallest eigenvalue. The idea behind to do so is that the principal component with the smallest eigenvalue is contributing least variance and so is least informative.

Using this procedure, principal components are eliminated until the remaining components explain some preselected variance is terms of percentage of the total variance. For example, if 90% of the total variance is needed, and suppose r principal components are eliminated which means that $(k-r)$ principal components contribute 90% variation, then r is selected to satisfy

$$\frac{\sum_{i=1}^{k-r} \lambda_i}{\sum_{i=1}^{k} \lambda_i} > 0.90.$$

Various strategies to choose the required number of principal components are also available in the literature.

Suppose after using such a rule, the r principal components are eliminated. Now only $(k-r)$ components will be used for regression. So Z matrix is partitioned as

$$Z = (Z_r \quad Z_{k-r}) = X(V_r \quad V_{k-r})$$

where Z_r submatrix is of order $n \times r$ and contains the principal components to be eliminated. The submatrix Z_{k-r} is of order $n \times (k-r)$ and contains the principal components to be retained.

The reduced model obtained after the elimination of r principal components can be expressed as

$$y = Z_{k-r}\alpha_{k-r} + \varepsilon^*.$$

The random error component is represented as ε^* just to distinguish with ε. The reduced coefficients contain the coefficients associated with retained $Z_j's$. So

$$Z_{k-r} = (Z_1, Z_2, ..., Z_{k-r})$$
$$\alpha_{k-r} = (\alpha_1, \alpha_2, ..., \alpha_{k-r})$$
$$V_{k-r} = (V_1, V_2, ..., V_{k-r}).$$

Using OLS on the model with retained principal components, the OLSE of α_{k-r} is

$$\hat{\alpha}_{k-r} = (Z_{k-r}' Z_{k-r})^{-1} Z_{k-r}' y .$$

Now it is transformed back to original explanatory variables as follows:

$$\alpha = V'\beta$$
$$\alpha_{k-r} = V_{k-r}'\beta$$
$$\Rightarrow \hat{\beta}_{pc} = V_{k-r}\hat{\alpha}_{k-r}$$

is the **principal component regression estimator** of β.

This method improves the efficiency as well as multicollinearity.

6. Ridge regression

The OLSE is the best linear unbiased estimator of regression coefficient in the sense that it has minimum variance in the class of linear and unbiased estimators. However, if the condition of unbiased can be relaxed, then it is possible to find a biased estimator of regression coefficient say $\hat{\beta}$ that has smaller variance them the unbiased OLSE b. The mean squared error (*MSE*) of $\hat{\beta}$ is

$$\begin{aligned} MSE(\hat{\beta}) &= E(\hat{\beta} - \beta)^2 \\ &= E\left[\left\{\hat{\beta} - E(\hat{\beta})\right\} + \left\{E(\hat{\beta}) - \beta\right\}\right]^2 \\ &= Var(\hat{\beta}) + \left[E(\hat{\beta}) - \beta\right]^2 \\ &= Var(\hat{\beta}) + \left[Bias(\hat{\beta})\right]^2 . \end{aligned}$$

Thus $MSE(\hat{\beta})$ can be made smaller than $Var(\hat{\beta})$ by introducing small bias is $\hat{\beta}$. One of the approach to do so is ridge regression. The ridge regression estimator is obtained by solving the normal equations of least squares estimation. The normal equations are modified as

$$(X'X + \delta I)\hat{\beta}_{ridge} = X'y$$
$$\Rightarrow \hat{\beta}_{ridge} = (X'X + \delta I)^{-1} X'y$$

is the **ridge regression estimator** of β and $\delta \geq 0$ is any characterizing scalar termed as **biasing parameter**.

As $\delta \to 0, \hat{\beta} \to b(OLSE)$ and as $\delta \to \infty$, $\hat{\beta} \to 0$.

So larger the value of δ, larger shrinkage towards zero. Note that the OLSE in inappropriate to use in the sense that it has very high variance when multicollinearity is present in the data. On the other hand, a very small value of $\hat{\beta}$ may tend to accept the null hypothesis $H_0 : \beta = 0$ indicating that the corresponding variables are not relevant. The value of the biasing parameter controls the amount of shrinkage in the estimates.

Bias of ridge regression estimator:

The bias of $\hat{\beta}_{ridge}$ is

$$\begin{aligned} Bias(\hat{\beta}_{ridge}) &= E(\hat{\beta}_{ridge}) - \beta \\ &= (X'X + \delta I)^{-1} X'E(y) - \beta \\ &= \left[(X'X + \delta I)^{-1} X'X - I \right] \beta \\ &= (X'X + \delta I)^{-1} \left[X'X - X'X - \delta I \right] \beta \\ &= -\delta (X'X + \delta I)^{-1} \beta. \end{aligned}$$

Thus the ridge regression estimator is a biased estimator of β.

Covariance matrix:

The covariance matrix of $\hat{\beta}_{ridge}$ is defined as

$$V(\hat{\beta}_{ridge}) = E\left[\left\{ \hat{\beta}_{ridge} - E(\hat{\beta}_{ridge}) \right\} \left\{ \hat{\beta}_{ridge} - E(\hat{\beta}_{ridge}) \right\}' \right].$$

Since

$$\begin{aligned} \hat{\beta}_{ridge} - E(\hat{\beta}_{ridge}) &= (X'X + \delta I)^{-1} X'y - (X'X + \delta I)^{-1} X'X\beta \\ &= (X'X + \delta I)^{-1} X'(y - X\beta) \\ &= (X'X + \delta I)^{-1} X'\varepsilon, \end{aligned}$$

so

$$\begin{aligned} V(\hat{\beta}_{ridge}) &= (X'X + \delta I)^{-1} X'V(\varepsilon) X (X'X + \delta I)^{-1} \\ &= \sigma^2 (X'X + \delta I)^{-1} X'X (X'X + \delta I)^{-1}. \end{aligned}$$

Mean squared error:

The mean squared error of $\hat{\beta}_{ridge}$ is

$$\begin{aligned} MSE(\hat{\beta}_{ridge}) &= Var(\hat{\beta}_{ridge}) + \left[bias(\hat{\beta}_{ridge}) \right]^2 \\ &= tr\left[V(\hat{\beta}_{ridge}) \right] + \left[bias(\hat{\beta}_{ridge}) \right]^2 \\ &= \sigma^2 tr\left[(X'X + \delta I)^{-1} X'X (X'X + \delta I)^{-1} \right] + \delta^2 \beta' (X'X + \delta I)^{-2} \beta \\ &= \sigma^2 \sum_{j=1}^{k} \frac{\lambda_j}{(\lambda_j + \delta)^2} + \delta^2 \beta' (X'X + \delta I)^{-2} \beta \end{aligned}$$

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

Thus as δ increases, the bias in $\hat{\beta}_{ridge}$ increases but its variance decreases. Thus the trade-off between bias and variance hinges upon the value of δ. It can be shown that there exists a value of δ such that

$$MSE(\hat{\beta}_{ridge}) < Var(b)$$

provided $\beta'\beta$ is bounded.

Choice of δ:

The estimation of ridge regression estimator depends upon the value of δ. Various approaches have been suggested in the literature to determine the value of δ. The value of δ can be chosen on the bias of criteria like

- the stability of estimators with respect to δ.
- reasonable signs.
- the magnitude of residual sum of squares etc.

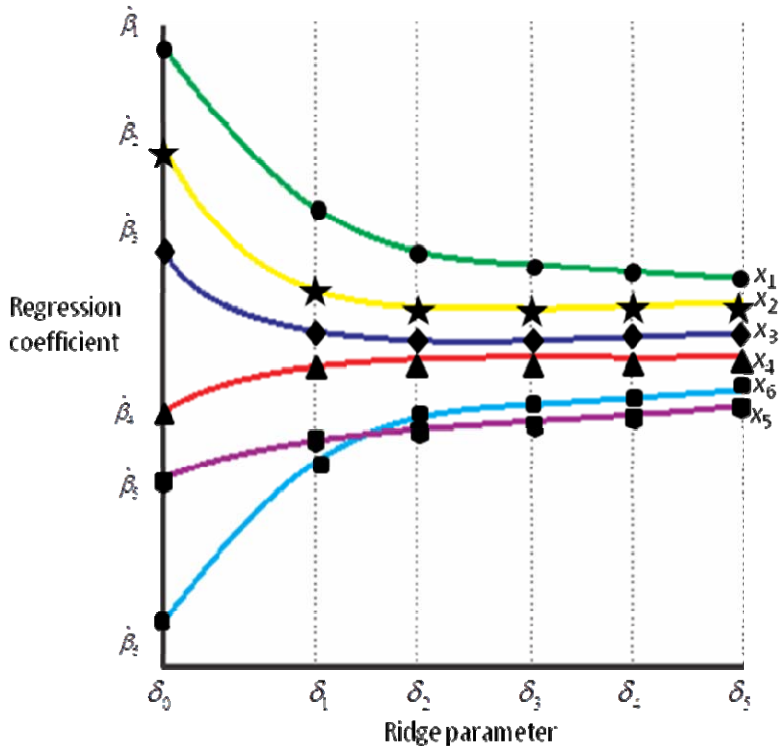
We consider here the determination of δ by the inspection of ridge trace.

Ridge trace:

Ridge trace is the graphical display of ridge regression estimator versus δ.

If multicollinearity is present and is severe, then the instability of regression coefficients is reflected in the ridge trace. As δ increases, some of the ridge estimates vary dramatically, and they stabilize at some value of δ. The objective in ridge trace is to inspect the trace (curve) and find the reasonable small value of δ at which the ridge regression estimators are stable. The ridge regression estimator with such a choice of δ will have smaller MSE than the variance of OLSE.

An example of ridge trace is as follows for a model with 6 parameters. In this ridge trace, the $\hat{\beta}_{ridge}$ is evaluated for various choices of δ and the corresponding values of all regression coefficients $\hat{\beta}_{j(ridge)}$'s, j=1,2,…,6 are plotted versus δ. These values are denoted by different symbols and are joined by a smooth curve. This produces a ridge trace for the respective parameter. Now choose the value of δ where all the curves stabilize and become nearly parallel. For example, the curves in the following figure become almost parallel, starting from $\delta = \delta_4$ or so. Thus one possible choice of δ is $\delta = \delta_4$ and parameters can be estimated as $\hat{\beta}_{ridge} = \left(X'X + \delta_4 I\right)^{-1} X'y$.



The figure drastically exposes the presence of multicollinearity in the data. The behaviour of $\hat{\beta}_{i(ridge)}$ at $\delta_0 \approx 0$ is very different than at other values of δ. For small values of δ, the estimates change rapidly. The estimates stabilize gradually as δ increases. The value of δ at which all the estimates stabilize gives the desired value of δ because moving away from such δ will not bring any appreciable reduction in the residual sum of squares. It multicollinearity is present, then the variation in ridge regression estimators is rapid around $\delta = 0$. The optimal δ is chosen such that after that value of δ, almost all traces stabilize.

Limitations:

1. The choice of δ is data-dependent and therefore is a random variable. Using it as a random variable violates the assumption that δ is constant. This will disturb the optimal properties derived under the assumption of constancy of δ.
2. The value of δ lies in the interval $(0,\infty)$. So a large number of values are required for exploration. This result is wasting of time. This is not a big issue when working with the software.
3. The choice of δ from graphical display may not be unique. Different people may choose different δ, and consequently, the values of ridge regression estimators will be changing. Another choice of δ is $\delta = \frac{k\hat{\sigma}^2}{b'b}$ where b and $\hat{\sigma}^2$ are obtained from the least-squares estimation.
4. The stability of numerical estimates of $\hat{\beta}_i's$ is a rough way to determine δ. Different estimates may exhibit stability for different δ, and it may often be hard to strike a compromise. In such a situation, generalized ridge regression estimators are used.
5. There is no guidance available regarding the testing of hypothesis and for confidence interval estimation.

Idea behind ridge regression estimator:

The problem of multicollinearity arises because some of the eigenvalues roots of $X'X$ are close to zero or are zero. So if $\lambda_1, \lambda_2, ..., \lambda_p$ are the characteristic roots, and if

$$X'X = \Lambda = diag(\lambda_1, \lambda_2, ..., \lambda_k)$$

then

$$\hat{\beta}_{ridge} = (I + \delta\Lambda^{-1})^{-1} b$$

where b is the OLSE of β given by

$$b = (X'X)^{-1}X'y = \Lambda^{-1}X'y.$$

Thus a particular element will be of the forms

$$\frac{1}{1+\frac{\delta}{\lambda_i}} b_i = \frac{\lambda_i}{\lambda_i + \delta} b_i.$$

So a small quantity δ is added to λ_i so that if $\lambda_i = 0,$ even then $\frac{\lambda_i}{\lambda_i + \delta}$ remains meaningful.

Another interpretation of ridge regression estimator:

In the model $y = X\beta + \varepsilon,$ obtain the least squares estimator of β when $\sum_{i=1}^{k} \beta_1^2 = C,$ where C is some constant. So minimize

$$\delta(\beta) = (y - X\beta)'(y - X\beta) + \delta(\beta'\beta - C)$$

where δ is the Lagrangian multiplier. Differentiating $S(\beta)$ with respect to β, the normal equations are obtained as

$$\frac{\partial S(\beta)}{\partial \beta} = 0 \Rightarrow -2X'y + 2X'X\beta + 2\delta\beta = 0$$

$$\Rightarrow \hat{\beta}_{ridge} = (X'X + \delta I)^{-1} X'y.$$

Note that if C is very small, it may indicate that most of the regression coefficients are close to zero and if C is large, then it may indicate that the regression coefficients are away from zero. So C puts a sort of penalty on the regression coefficients to enable its estimation.