**Multicollinearity**

**Nature of Multicollinearity:** The term multicollinearity is introduced in Economic analysis by Economist “Ranger Frisch”. Multicollinearity refers to the existence of a perfect or exact linear relationship among some or all explanatory variables of a regression model.

For the k-variables regression involving explanatory variable $X_1, X_2, X_3, \ldots, X_k$ having the following linear relationship

$$\alpha_1 X_1 + \alpha_2 X_2 + \ldots + \alpha_k X_k = 0$$

**Example:** In demand function of a commodity suppose the quantity demanded of commodity ‘A’ depends upon the price of commodity ‘B’. If there are two prices are correlated to each other it will be difficult to separate the influence of two prices on the demand of the commodity. Such a problem is known as multicollinearity problem.

**Types of Multicollinearity:**

There are two types of multicollinearity. They are-

1. Exact/Perfect Multicollinearity.
2. Near or less than perfect Multicollinearity.

**Exact Multicollinearity:** If exist perfect linear relationship among the explanatory variables then it is treated as exact multicollinearity. In case of exact multicollinearity the design matrix as data matrix ‘X’ is not of full rank & consequently $(X'X)^{-1}$ does not exist. In this case $|X'X|=0$.

**Example:** For the K-variables regression model involving explanatory variable $X_1, X_2, \ldots, X_k$ (where $X_1=1$ for all observations to allow for the intercept term) an exact linear relationship is said to exist if the following condition is satisfied.

$$\lambda_1 X_1 + \lambda_2 X_2 + \lambda_k X_k = 0$$

Where $\lambda_1, \lambda_2, \ldots, \lambda_k$ are constants such that not all of them are zero simultaneously.

Assume that $\lambda_1 \neq 0$ then the equation can be written as

$$X_1 = -(\lambda_2/\lambda_1)X_2 - (\lambda_3/\lambda_1)X_3 - \ldots - (\lambda_k/\lambda_1)X_k$$

Which show that $X_1$ is linearly related with other explanatory variables(X’s).
**Near multicollinearity**: If the explanatory variables (x’s) are strongly as highly correlated but not perfectly then it’s called near multicollinearity. In this case $(X'X)^{-1}$ is exist but with related large diagonal elements i.e. $|X'X| \neq 0$.

**Example**: When the explanatory variables (X’s) are inter correlated but not perfectly then a linear relationship is said to be exist if

$$\lambda_1X_1 + \lambda_2X_2 + \ldots + \lambda_kX_k + v_i = 0$$

where $V_i$ is a stochastic error term & $\lambda$’s are constant such that not all of them are zero simultaneously.

Assume that $\lambda_1 \neq 0$, then the equation can be written as

$$X_1 = -(\lambda_2/\lambda_1)X_2-(\lambda_3/\lambda_1)X_3-\ldots-(\lambda_k/\lambda_1)X_k - v_i/\lambda_1$$

Which shows that $X_1$ is not exactly linearly related with other explanatory variables.
Sources of multicollinearity: There are several sources of multicollinearity.

1. The data collection method employed, for example, sampling over a limited range of the values taken by the regressors in the population.
2. Constraints on the model or in the population being sampled. For example, in the regression of electricity consumption on income ($X_2$) & house size ($X_3$) there is a physical constraint in the population in the families with higher income generally have larger homes than families with lower incomes.
3. Model specifications, for example, adding polynomial terms to a regression model, especially when the range of the $X$ variable is small.
4. An over determined model. This happens when the model has more explanatory variables than the number of observations. This could happen in medical research where there may be a small number of patients about whom information is collected on a large number of variables.

Consequences of multicollinearity: In case of near or high multicollinearity one is likely to encounter the following consequences.

1. Although BLUE, the OLS estimators have the large variance & covariance’s making precise estimation difficult.
2. Because of consequence 1, the confidence intervals tend to be much wider, leading to the acceptance of the “zero null hypothesis” more readily.
3. Also because of consequence 1, the t ratio of one or more coefficients tends to be statistically insignificant.
4. Although the t ratio of one or more coefficients statistically insignificant, $R^2$, the overall measure of goodness of fit, can be very high.
5. The OLS estimators & their standard error can be sensitive to small changes in the data.

If multicollinearity is perfect among the explanatory variables then the regression coefficient of the $X$ variables are indeterminate & their standard errors are infinite. If multicollinearity is less than perfect, then the regression coefficients although determinate, possesses large standard errors (in relation to the coefficients) which mean
that the coefficient cannot be estimated with great precision or accuracy. But when there is no multicollinearity among the X's variables then we can easily estimate the regression coefficient.

For this reason CLRM assume that there is no multicollinearity among the X's.

**Detection of multicollinearity:**

The indicators for detecting multicollinearity are as follows:

1. Eigen Values & Conditional Index: Here we discuss the method of Eigen value & conditional index to detect the multicollinearity. At first we have to calculate the data matrix. Then using |X'X-λI| = 0 we get the values of λ which is eigen value. Now we have

   Conditional Index (CI)= \( \sqrt{\frac{\text{maximum eigen value}}{\text{minimum eigen value}}} \)

   After calculating CI, if CI lies between 10 to 30 then there is moderate multicollinearity. And if CI exceeds 30 then there have severe multicollinearity.

2. High R\(^2\) but few significant t-ratios: This is classic symptom of multicollinearity. If R\(^2\) is high, say excess of 0.8, the F-test in most cases reject the H\(_0\) that the partial slope coefficients are simultaneously equal to zero, but the individual t-test will show that none or very few of the partial coefficients are statistically different from zero. In short we can write when R\(^2\) is very high but none of the regression coefficients is statistically significant.

   To illustrate the various points made thus far, let us reconsider the consumption-income example of Chapter 3. In Table 10.5 we reproduce the data of Table 3.2 and add to it data on wealth of the consumer. If we assume that consumption expenditure is linearly related to income and wealth, then, from Table 10.5 we obtain the following regression:

   \[
   \hat{Y}_i = 24.7747 + 0.9415X_{3i} - 0.0424X_{3i}
   \]

   \[
   (6.7525) \quad (0.8229) \quad (0.0807)
   \]

   \[
   t = (3.6690) \quad (1.1442) \quad (-0.5261)
   \]

   \[
   R^2 = 0.9635 \quad \hat{R}^2 = 0.9531 \quad df = 7
   \]

   Regression (10.6.1) shows that income and wealth together explain about 96 percent of the variation in consumption expenditure, and yet neither of the slope coefficients is individually statistically significant. Moreover, not only is the wealth variable statistically insignificant but also it has the wrong
3. High pairwise correlation among Regression: Another suggested rule of thumb is that if the pairwise as zero order correlation coefficients between two regressors is high, say excess of 0.1 then multicollinearity is a serious problem.

4. Examination of partial correlation:
   If $R^2$ is high but the partial correlation are comparatively low may suggest that the explanatory variables are highly correlated.

6. Tolerance and Variance Inflation Factor:
The speed with which variance and covariance increase can be seen with the VIF, which is defined as

$$\text{VIF} = \frac{1}{1 - r_{23}^2}$$

VIF shows how the variance of an estimator is inflated by the presence of multicollinearity.

<table>
<thead>
<tr>
<th>Value of $r_{23}$</th>
<th>VIF</th>
<th>$\text{var}(\hat{\beta}_2)$</th>
<th>$\text{var}(\hat{\beta}_2)$</th>
<th>$\text{var}(\hat{\beta}_2)$</th>
<th>$\text{cov}(\hat{\beta}_2, \hat{\beta}_3)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.00</td>
<td>1.00</td>
<td>$\frac{\sigma^2}{\sum x_{2i}^2}$</td>
<td>$A$</td>
<td>$-$</td>
<td>0</td>
</tr>
<tr>
<td>0.50</td>
<td>1.33</td>
<td>$1.33 \times A$</td>
<td>$1.33$</td>
<td>$1.33$</td>
<td>$0.67 \times B$</td>
</tr>
<tr>
<td>0.70</td>
<td>1.98</td>
<td>$1.98 \times A$</td>
<td>$1.98$</td>
<td>$1.98$</td>
<td>$1.37 \times B$</td>
</tr>
<tr>
<td>0.80</td>
<td>2.78</td>
<td>$2.78 \times A$</td>
<td>$2.78$</td>
<td>$2.78$</td>
<td>$2.22 \times B$</td>
</tr>
<tr>
<td>0.90</td>
<td>5.76</td>
<td>$5.26 \times A$</td>
<td>$5.26$</td>
<td>$5.26$</td>
<td>$4.73 \times B$</td>
</tr>
<tr>
<td>0.95</td>
<td>10.26</td>
<td>$10.26 \times A$</td>
<td>$10.26$</td>
<td>$10.26$</td>
<td>$9.74 \times B$</td>
</tr>
<tr>
<td>0.97</td>
<td>16.92</td>
<td>$16.92 \times A$</td>
<td>$16.92$</td>
<td>$16.92$</td>
<td>$16.41 \times B$</td>
</tr>
<tr>
<td>0.99</td>
<td>50.25</td>
<td>$50.25 \times A$</td>
<td>$50.25$</td>
<td>$50.25$</td>
<td>$49.75 \times B$</td>
</tr>
<tr>
<td>0.995</td>
<td>100.00</td>
<td>$100.00 \times A$</td>
<td>$100.00$</td>
<td>$100.00$</td>
<td>$90.50 \times B$</td>
</tr>
<tr>
<td>0.999</td>
<td>500.00</td>
<td>$500.00 \times A$</td>
<td>$500.00$</td>
<td>$500.00$</td>
<td>$499.50 \times B$</td>
</tr>
</tbody>
</table>

Note: $A = \frac{\sigma^2}{\sum x_{2i}^2}$

$B = \sqrt{\sum x_{2i}^2 \sum x_{3i}^2}$
7. Low value of $|X'X|$ in case of exact multicollinearity i.e $|X'X|=0$.

**Remedial Measures:**

If multicollinearity has serious effects on coefficients estimates of important factor, one should adopt one of the following solutions –

1. **Increase The Sample Size:**
   The easiest way to overcome the problem of multicollinearity is to increase the sample size. Investigators are advised to collect more data to reduce the intensity of collinearity.

2. **Using Extraneous Estimate:**
   To eliminate the effects of multicollinearity the other commonly adopted method is the uses of extraneous information is estimating parameters. Suppose our model is
   \[ Y = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + u \]
   Where $X_1 \& X_2$ are correlated. If we know that $\alpha_2 = 0.5\alpha_1$, then the model will be
   \[ Y = \alpha_0 + \alpha_1 X_1 + 0.5\alpha_1 X_2 + u \]
   \[ = \alpha_0 + \alpha_1 (X_1 + 0.5X_2) + u \]
   \[ = \alpha_0 + \alpha_1 X' + u \]
   Now we can estimate $\alpha_1$ by OLS and hence $\alpha_2 = 0.5\alpha_1$. 

\[ A = \frac{\sigma^2}{\Sigma x^2} \]
3. Dropping Variables:
When we faced with severe multicollinearity one of the simplest things to do is to drop one of the collinear variables.

4. Combining Cross-Sectional And Time Series Data:
Generally time series data is affected by multicollinearity problem. So if we combine the cross-sectional data in time series data then the multicollinearity problem should be reduced.

5. Model Specification:
Multicollinearity may be overcome if we specify our model; this can be done in the following way.
   a) One approach is to redefine regressors.
   b) Re-specification of lagged variable or other explanatory variable in a distributed lagged variables.

Is multicollinearity necessarily bad?
It has been said that if the purpose of regression analysis is prediction or forecasting, then multicollinearity is not a serious problem because the higher the $R^2$, the better the prediction.
Moreover, if the objective of the analysis is not only prediction but also reliable estimation of the parameters, serious multicollinearity will be a problem because we have seen that it tends to large standard error of the estimators.
In one situation however, multicollinearity may not impose a serious problem. This is the case when $R^2$ is high and the regression coefficient are individually significant as revealed by the higher t-values.