Multicollinearity

Meaning:

A crucial condition for the application of least square is that the explanatory variables are not perfectly linearly correlated ie $r_{xixj} \neq 1$. The term multicollinearity is used to denote the presence of linear relationships among explanatory variables. If the explanatory variables are perfectly linearly correlated, ie if the correlation coefficient for these variables is equal to unity, the parameters become indeterminate. It is impossible to obtain numerical values for each parameter separately and the method of least-square breaks down.

In practice there is some degree of intercorrelation among the explanatory variables. In these case the simple correlation coeff. for each pair of explanatory variables will have a value between 0 and 1. So multicollinearity is a phenomenon which is there in most relationships due to the nature of economic magnitudes.

Example :

Suppose that consumption expenditure of an individual depends on his income and liquid assets. If over a period of time income and the liquid assets change by the same proportion the influence on consumption of one of these variables may be erroneously attributed to the other. The effects of these variables on consumption cannot be sensibly investigated due to their interconelation.

Question :

Define the problem of multicollineanity. Give an example. (2/5 marks)

More examples :

Multicollineanity may arise for various reasons :

There is a tendency of economic variables to move together over time. Economic magnitudes are influenced by the same factors and in consequence once these determining factors become operative the economic variables show the same broad pattern of behaviour over time.

For example in periods of booms or rapid economic growth the basic economic magnitudes grow, although some tend to lag behind others. Thus income, consumption, savings, investment, prices, employment tend to rise in periods of economic expansion and decrease in periods of recession.

Next the use of lagged values of some explanatory variables as separate independent factors in the relationship may result in multicollinearity.

for example:

$$C_t = \beta_0 Y_t + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + u_t$$

Here consumption is a fn of Income (Y) Now consumption in time period 't' depends on Income of time period 't' and ~~also~~ also on previous period Income ie for period t-1, t-2, etc. Now income of different periods can be corelated. So multicollinearity is almost certain in distributed lag models.

Consequences of Multicollinearity:

If the inter correlation between the explanatory variables is perfect ($r_{xixj} = 1$) then

i) the estimates of the coefficients are indeterminate.

ii) The standard error of these estimes become infinitely large.

Proof- The relationship to be estimates is $Y = b_0 + b_1 X_1 + b_2 X_2 + u$ here X_1 and X_2 are related with the exact relation

$X_2 = kX_1$, where k is any arbitrary const.

The formula for the estimation of the coefficient- $\hat{b}_1$ & $\hat{b}_2$ are

$$\hat{b}_1 = \frac{(\Sigma x_1 y)(\Sigma x_2^2) - (\Sigma x_2 y)(\Sigma x_1 x_2)}{(\Sigma x_1^2)(\Sigma x_2^2) - (\Sigma x_1 x_2)^2}$$

$$= \frac{k^2(\Sigma x_1 y)(\Sigma x_1^2) - k^2(\Sigma x_1 y)(\Sigma x_1^2)}{k^2(\Sigma x_1^2)^2 - k^2(\Sigma x_1^2)^2}$$

$$= \frac{0}{0}$$

* Model in reduced form:

$$y = b_1 x_1 + b_2 x_2 + u$$

$$\hat{b}_2 = \frac{(\Sigma x_2 y)(\Sigma x_1^2) - (\Sigma x_1 y)(\Sigma x_1 x_2)}{(\Sigma x_1^2)(\Sigma x_2^2) - (\Sigma x_1 x_2)^2}$$

$$= \frac{k^2(\Sigma x_1 y)(\Sigma x_1^2) - k^2(\Sigma x_1 y)(\Sigma x_1^2)}{k^2(\Sigma x_1^2)^2 - k^2(\Sigma x_1^2)^2}$$

$$= \frac{0}{0}.$$

So here the parameters are indeterminate.

Proof: If $r_{xixj} = 1$, the standard errors of the estimates become infinitely large. Let the model be given by

$$Y = b_0 + b_1 X_1 + b_2 X_2 + u$$

if X_1 and X_2 are perfectly correlated $x_2 = k x_1$ the var of $\hat{b}_1$ and $\hat{b}_2$ will be

$$\text{Var}(\hat{b}_1) = \sigma^2_u \frac{\sum x_2^2}{\sum x_1^2 \sum x_2^2 - (\sum x_1 x_2)^2}$$

and

$$\text{Var}(\hat{b}_2) = \sigma^2_u \frac{\sum x_1^2}{\sum x_1^2 \sum x_2^2 - (\sum x_1 x_2)^2}$$

Substituting kX_1 for x_2 we obtain

$$\text{Var}(\hat{b}_1) = \sigma^2_u \frac{\sum x_1^2}{k^2(\sum x_1^2)^2 - k^2(\sum x_1^2)^2}$$

$$= \sigma^2_u \frac{\sum x_1^2}{0} = \infty$$

Thus the variance of the estimates become infinite in presence of perfect multicollinearity unless $\sigma^2_u = 0$ However there is no prior reason why σ^2_u should tend to zero when intercorrelation of the explanatory variables increases.

Detection of Multicollinearity

Multicollinearity is a question of degree and not of kind. The meaningful distinction is not between the presence and the absence of multicollinearity but between its various degrees.
Since multicollinearity refers to the condition of the explanatory variables it is a feature of the sample and not of the population.
Since multicollinearity is essentially a sample phenomenon, there is no unique method of detecting it or measuring its strength.

Farrer - Glauber test for Multicollinearity

A statistical test for multicollinearity has been recently developed by Farrer - Glauber. It is really a set of three tests.
Firstly: A chi-square (χ^2) test for the presence and severity of multicollinearity in a function with several explanatory variables.
Secondly: An F-test for the location of multicollinearity.
Thirdly: A 't' test for the pattern of multicollinearity.

Solutions for the incidence of multicollinearity.

The solution which may be adopted if multicollinearity exists in a fn. vary depending on the severity of multicollinearity, on available of other sources of data (large samples or cross-section samples etc) on the importance of factors which are multicollinear on the purpose for which the fn. is being estimated and other considerations.

1) Some writers have suggested that if multicollinearity affects some of the unimportant factors one may exclude these factors from the fn. Again specification error may well be expected to undermine the BLUE character of the OLS.

2) Some also have suggested that if multicollinearity does not seriously affect the estimates of the coefficients one may tolerate its presence in the system although the integrity of the least sq. estimates to a certain extent may be lost.

3) It has been suggested that multicollinearity may be avoided or

4)

reduced if we increase the size of the sample by gathering more observations.

4) Multicollineanity may be overcome if we introduce additional equations into our model to express meaningfully the relationships between the multicollinear X's. When looking at the set of explanatory variables one is able, in most cases, to find relationship between the X's (and other new variables) which make economic sense. By explicitly formulating these relationships one can form a simultaneous-equation technique. The reduced form method will by pass the problem of multicollineanity in the original equation provided the new model is exactly identified.

5) Different methods of estimation.
i) The method of restricted least square.
ii) The method of pooling cross-section and time series data (which is actually a special case of restricted least square)

Multicollinearity and Mis-specification in variables: specification bias.

Multicollinearity is often a serious source of error in the individual coefficients. Generally we tend to omit variables from various functions in order to avoid the consequences of multicollinearity. This procedure introduces a specification error in the model, since the omission of variables will affect the values of the parameters of the remaining variables. By omitting (or adding) variables one may avoid multicollinearity but one is bound to have a specification error in the parameters. Here we examine the consequences of a function from which a variable is wrongly excluded.

Assume that the true function explaining the variation in y is

$$y = b_1 x_1 + b_2 x_2 + u.$$

However, either due to ignorance of the true relation or because x_1 and x_2 are strongly multicollinear, we exclude x_2 from the f. We apply least-

Prime Merit
Page No. 11
Date

square to the equation:

$$y = b_1^* x_1 + u^*$$

clearly b_1^* will be different from b_1. The actual difference may be found as follows:

Applying OLS to the mis-specified f. we obtain $b_1^* = \Sigma y x_1 / \Sigma x_1^2$

Now the normal equations for the correctly specified model are

$$\Sigma y x_1 = b_1 \Sigma x_1^2 + b_2 \Sigma x_1 x_2$$
$$\Sigma y x_2 = b_1 \Sigma x_1 x_2 + b_2 \Sigma x_2^2$$

Dividing the first equation through by Σx_1^2 we obtain

$$\frac{\Sigma y x_1}{\Sigma x_1^2} = b_1 + b_2 \frac{\Sigma x_1 x_2}{\Sigma x_1^2}$$

We observe that $\Sigma y x_1 / \Sigma x_1^2 = b_1^*$. and $\Sigma x_1 x_2 / \Sigma x_1^2$ is the slope of the regression of x_2 on x_1, that is the the coefficient in fn.

$$x_2 = a x_1. \quad \& \quad a_1 = \frac{\Sigma x_1 x_2}{\Sigma x_1^2}.$$

Hence we may write

$$E(b_1^*) = b_1 + b_2 a_1$$

The specification error is

$$[\text{specification bias}] = [E(b_1^{*}) - b_1]$$

$$= b_2 \cdot a_1$$

By the definition of the correct model we know that $b_2 \neq 0$. Therefore b_1^{*} would be equal to b_1 only if $a_1 = 0$, that is if x_1 and x_2 are not corelated at all. In real life we cannot expect to have such variables in an econometric model since most of the economic variables are interdependent. Thus omission of variables from the fn. will yield biased estimates of the parameters of the included variables.