Multicolinearity

A crucial condition for the application of least squares is that the explanatory variable are not perfectly linearly co-related

i,e $\Sigma x_i x_j \neq 1$

If there is no linear relationship between the regression of Ev than they are said to be orthogonal. When the regressions are orthogonal inferences like —

A) Identifying relative effects of the regresso variables.

B) Prediction and/or estimation, and

C) Selection of an appropiate set of variables for the model.

Can be made relatively very easily. Unfortunately in most applications of regression regressors are not orthogonal. However in some situation the regressors are nearly perfectly linearly correlated and in such cases, the inferences based on the regression model can be misleading or erroneous. When there are near interrelationship or dependencie

among the EV, then there exist the problem of multicolenearity. Hence, the term multicolinearity is ~~issued~~ used to denote the presence of linear relationship or linear relationship among the EV.

If the EV are perfectly linearly co related i,e if co-relation co-efficient of the variables is equal to unity then the parameters became ~~interdet~~ ~~de~~ indeterminent and it becames impossible to obtain numerical values of each parameter separately and the method of least square breaks down.

When $Uxixj = 1$ then there is complete vial-ation of the assumption, while $Uxixj$ lies between 0 and 1, then there is near correlation.

* $0 < rxixj < 1$ —> near correlation
  $rxixj = 1$ —> Perfectly linearly correlat-ed
  $rxixj = 0$ —> no correlation.
  $rxixj \neq 1$ —> (assumption) not linearly perfectly correlated.

In practice neither of the extreme cases, there is some degree of i.e of orthogonal x's is often met. In most cases, there is some degree of inter correlation among the explanatory variables due to the interdependence of many economic magnitudes overtime. In this cases the simple correlation co-efficient for each pair of explanatory variable will have a value between 0 and 1.

Multi co-linearity is not a condition that either exist or does not exist in econom-ic functions but it is rather a phenomena inherence in most relationships due to the nature of economic magnitudes. There is no conclusive evidence concerting the degree of co-linearity, which it present will effect seriously the parameter estimates. Intuitively, when two explanatory variables are changing in nearly the same way it becames extremely difficult to establish the influence ~~the~~ of each one of the regressers on Y separately eg, assume that consumption expenditure of an individual depends on his income and liquide asset change by the same proportion then, the influence of one of these variables on consump-tion may be erroniously attributed to the other and the effects of these variables one

consumption can not be sensively investig-ated due to high. Inter correlated multi-colinearity is thus a situation where because of strong interrelationship among the explanatory variables, it becames very difficult to thus untangle their separate effects on the dependent variable i,e it becomes difficult to ~~obtain~~ change precise estimates of the regression coefficients i,e,

$$\hat{\beta} = (x'x)^{-1} x'y$$

does not exist or we do not have an exact solution. Thus the term multicolinearity which was due to Ragnar Frisch oreginally meant "existence of a perfect or exact or linear relationship among some or all the explanatory variables of a regression model". Today however multicolinearity is used in a broader sense to include the case of perfect multicolinearity as well as the case where the explanatory variables are inter-corelated but not perfectly inter correlated.

causes of multicolinearity / sources :

① the data collection method can lead to multicolinearity problems when an analyst samples only a sub space of the region of the regressions defined.

② Constraints on the model or in the population being samples can cause multicolinearity.

eg - Suppose that an electric utility is investigating the effect of family income (x_1) and house size (x_2) on residential electricity consumption, the levels of the two regressor obtain in the sample data share that if the data lie approximately along the straight line thus indicating potential multicolinearity problem.

In this example of physical constraint in the population has caused this phenomena i,e families with higher income generally have larger homes than families with lower. when physical constraint such regardless of the sampling method employed, constraints after occur in problems involving production or chemical process whether regressor are the component of a product and these components add to a constraint.

③ Multicolinearity may be induced by the choice of model. Moreover, adding polynomial terms to the regression model can result in significant multicolinearity.

④ An over (adding unnecessary explanatory variables) defined model with more regressor variables than observations may be a source of multicolinearity. These models are sometimes encountered in medical and behavioural research where there may be small number of sample units, eg- the small number of patients, and information is collected for a large number of regressors on each subject. The usual approach is dealing with multicolinearity in this context is to eliminate some of the regressor variables from consideration.

⑤ There is a tendency of economic variables to move together overtime which is a cause of multicolinearity. Economic magnitudes are influenced by the same factors and in consequences once these determining factors become operative, the economic variables show the same broad. Pattern of

behaviour over time eg - In periods of boom or rapid economic growth, the basic economic magnitudes grow although some tend to lag behind others. Thus income, consumption, saving investment, prices, employment tend to rise in the periods of economic expansion and decrease in periods of recession. Growth and trend factors in time series are the most serious cause of multicolinearity.

⑥ The use of lagged values of some explanatory variables as seperate independent factors in the model is another cause of multicolinearity. eg - In consumption function, it is customary to include, among the explanatory variables, past as well as present values of income. Similarly, in investment functions distributed lags concerning past levels of economic activities are introduced as separate explanatory variables. Naturally, the successive values of these variables are inter-correlated and hence multicolinearity is almost certain to exist in distribution lag models.

Consequences of multicolinearity

(A) Perfect multicolinearity : If the inter-co-relation between the explanatory variables is perfect. i.e $\sum x_i x_j = 1$, then consequences are as follows –

(a) the estimates of the coefficients are inde-terminant.

(b) the standard errors of these estimates becames infinitely large.

Proof of (a)

Suppose that the relation be estimated is $y = b_0 + b_1 x_1 + b_2 x_2 + u$

and that x_1 and x_2 are related with the exact relation, $x_2 = k x_1$, where k is any arbitrary constant.

The formula for the estimation of the coefficients $\hat{b}_1$ and $\hat{b}_2$ are —

$$\hat{b}_1 = \frac{(\Sigma x_1 y)(\Sigma x_2^2) - (\Sigma x_2 y)(\Sigma x_1 x_2)}{(\Sigma x_1^2)(\Sigma x_2^2) - (\Sigma x_1 x_2)^2} \quad \text{——— (1)}$$

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

Now putting, $x_2 = kx_1$ in eq^n (1)

$$\hat{b}_1 = \frac{(\Sigma x_1 y)\,\Sigma (kx_1)^2 - (\Sigma kx_1 y)(\Sigma x_1 kx_1)}{(\Sigma x_1^2)\,\Sigma (kx_1)^2 - (\Sigma x_1 kx_1)^2}$$

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

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

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

Now putting $x_2 = kx_1$ in eq^n (2)

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

$$= \frac{(\Sigma k x_1 y)(\Sigma x_1^2) - (\Sigma x_1 y)(\Sigma x_1 kx_1)}{(\Sigma x_1^2)\,\Sigma (kx_1)^2 - (\Sigma k x_1 x_1)^2}$$

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

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

Therefore the parameters are indeterminant and there is no way of finding separate values of these coefficients.

Proof of (b)

If $r_{x_i x_j} = 1$, then the standard errors of the estimates became infinitely large.

In a two variable model, $Y = b_0 + b_1 x_1 + b_2 x_2 + u$

$$\text{var}(\hat{b_1}) = \sigma_u^2 \frac{\sum x_2^2}{\sum x_1^2 \sum x_2^2 - (\sum x_1 x_2)^2} \quad \text{---(1)}$$

$$\text{var}(\hat{b_2}) = \sigma_u^2 \frac{\sum x_1^2}{\sum x_1^2 \sum x_2^2 - (\sum x_1 x_2)^2} \quad \text{---(2)}$$

Now substituting, $x_2 = kx_1$ in eq^n (1) & (2)

$$\text{var}(\hat{b_1}) = \sigma_u^2 \frac{\sum (kx_1)^2}{\sum x_1^2 \sum (kx_1)^2 - (\sum x_1 kx_1)^2}$$

$$= \sigma_u^2 \frac{k^2 (\sum x_1)^2}{\sum x_1^2 k^2 x_1^2 - k^2 (\sum x_1^2)^2}$$

$$= \sigma_u^2 \frac{k^2 \sum x_1^2}{k^2 (\sum x_1^2)^2 - k^2 (\sum x_1^2)^2}$$

$$= \sigma_u^2 \frac{k^2 \sum x_1^2}{0} = \infty$$

$$\text{Var}(\hat{b}_2) = \sigma_u^2 \frac{\sum x_1^2}{\sum (kx_1)^2 \cdot \sum x_1^2 - (\sum x_1 kx_1)^2}$$

$$= \sigma_u^2 \frac{\sum x_1^2}{k^2 (\sum x_1^2)^2 - k^2 (\sum x_1^2)^2}$$

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

Thus the variances of the estimates became infinite unless $\sigma_u^2 = 0$.

(B) Near multicollinearity or (High multicollinearity)

(1) Large variance and co-variance of OLS estimators

Although BLUE, the OLS estimators have large variances and making estimation difficult. Say, in three variable linear regression model —

$$Y_i = \hat{\beta}_2 x_{2i} + \hat{\beta}_3 x_{3i} + \hat{u}_i \quad \text{——— (1)}$$

$$\text{var}(\hat{\beta}_2) = \frac{\sigma^2}{\sum x_{2i}^2 (1 - r_{23}^2)} \quad \text{——— (2)}$$

$$\& \quad \text{var}(\hat{\beta}_3) = \frac{\sigma^2}{\sum x_{3i}^2 (1 - r_{23}^2)} \quad \text{——— (3)}$$

$$\text{and. } \text{cov}(\hat{\beta}_2 \hat{\beta}_3) = \frac{-r_{23} \sigma^2}{(1 - r_{23}^2) \sqrt{\sum x_{2i}^2 \sum x_{3i}^2}} \quad \text{——— (4)}$$

Where, r_{23} is correlation coefficient between x_2 and x_3.

It is now apparent that from ② and ③ as $r_{23} \xrightarrow{\text{tends to}} 1$. i,e If collinearity increases, the variance of the two estimators increases and ~~Cov~~ the co-variance of the two estimators also increases in absolute value.

② Wider confidence interval : Because of large standard errors the confidence interval for the relavente population parameter also tends to be very large. Therefore in cases of high multicollinearity the sample data may be compatible with ~~diffrt~~ diverse set of hypothesis and hence the problem of accepting of all hypothesis increases.

③ Insignificant t-ratios : To test a null hypothesis we use the t-ratio, i,e $\frac{\hat{\beta}}{SE(\hat{\beta})}$ and compare the estimated t value with the critical t value from t table. But in case of high multicollinearity the estima- ~~test~~ standard error become dramatically large, thereby making the t-value smaller.

therefore in such cases there is the probability of accepting of false hypotheses.

④ A high R^2 but a few significant t-ratio :

If we consider a K variable linear regression model — $Y_i = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots\ldots + \beta_K x_K + u$ — (1) and if there is high and near multicollinearity defined as one and more partial slope co-efficient individually statistically insignificant on the bases of 't' test. However R^2 means such situations may be so high say in excess of 0.9 that on the bases of F-test we may reject the hypo-theses to zero ; i.e $\beta_2 = \beta_3 = \ldots\ldots = \beta_K = 0$. Infact insignificant t-ratios and a very high R^2 is one of the signals of multicollinearity.

⑤ As long as multicollinearity is not perfect, estimation of the regression coefficient is possible but their estimates and their standard errors become very sensitive to even a small change in data.

Test / methods of multicollinearity

(1) Test based on Frisch's confluence analysis:

The seriousness of the effects of multicollinea since the seems to depend on the effects of degree of inter-correlation i.e. r_{xixj} as well the overall co-relation co-efficient R^2. The standard error, the partial correlation co-effic and total correlation coefficient may be used for testing multi-collinearity. However, none of these, by itself is a satisfactory indicator of multicollinearity, because —

(i) Large standard error do not always appear with multicollinearity. Large standard error may arise for various other reasons.

(ii) A high r_{xixj} is sufficient condition but not a necessary condition of multicollinearity.

(iii) R^2 may be high and yet the estimates may not be significant and imprecise.

However the all these criterion should help the detection of multicollinearity and has been suggested by frisch's test.

Frisch's test

In this test it is required to estimate all the possible regressions between the variables which are present in relationship taking each variable successively as the dependent variable and considering all possible regression of each variables and all others which are introduce into the analysis.

(2) The Farrar - Glauber test of multicollinearity:

A statistical test of multicollinearity has been developed by Farrar - Glauber and it is a set of 3 tests.

(i) Firstly, chi-square test is used to detect the presence and severity of multicollinearity in a function of several explanatory variables. The basic hypotheses here are -

H_0 : The X's are orthogonal

H_a : The X's are not orthogonal.

It is observed that if the chi-square is greater than the theoritical chi-square value, with $\frac{1}{2}k(k-1)$ degrees of freedom then we reject the assumption of orthogonality and we say that there is multicollinearity in the

function. If the observed chi-square is less than theoritical chi-square ~~is less than theoritical chi-square~~ we accept the assumption of orthogonality and say there is no multicollinearity.

(ii) Secondly, an F test is also used for the location of multicollinearity. To Locate the factors which are multicollinear, Farrar and Glauber computed the multiple correlation coefficients amongst the explanatory variables. ($R^2_{x_1 . x_2 x_3 \ldots x_k}$, $R^2_{x_2 . x_1 x_3 \ldots x_k}$) and in general $R^2_{x_i . x_1 x_2 \ldots x_k}$ and then test the statistical significance of these multiple correlation coefficient with an F test. For each multiple correlation coefficient, we compute the observed F^*.

i,e $$F^* = \frac{(R^2_{x_i . x_1 x_2 \ldots x_k}) / (K-1)}{(1 - R^2_{x_i . x_1 x_2 \ldots x_k}) / (n-K)}$$

where, n = size of the sample
K = Number of explanatory variables.

The hypothesis are :- $H_0 : R^2_{x_i . x_1 x_2 \ldots x_k} = 0$
$H_a : R^2_{x_i . x_1 x_2 \ldots x_k} \neq 0$

the observed F^* is compared with theoritical f value with degrees of freedom $u_1 = (k-1)$
$u_2 = (n-1)$

If $F^* > F$, we reject H_0 and say that there is multicollinearity.
And if $F^* < F$, we accept H_0 and say that there is no multicollinearity.

(iii) Thirdly, a t-test is used to detect the variables which cause multicollinearity. To find which variables are responsible for the multicollinearety, we compute the partial correlation coefficient among the explanatory variables and test their statistical significance with the t-statistics.

Remedial measures

1. A priori information:
Suppose we consider the model -

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i \quad \text{---------} \quad (1)$$

where, Y = Consumption
X_2 = Income
X_3 = Wealth

As noted income and wealth variables are highly collinear but suppose a priori we believe that $\beta_3 = \cdot 10\beta_2$. i,e the rate of change of consumption with respect to wealth is $\frac{1}{10}$ the corresponding rate with respect to income. We can then run the following regression; —

$$Y_i = \beta_1 + \beta_2 X_{2i} + \cdot 10\beta_2 X_{3i} + U_i$$

$$= \beta_1 + \beta_2 X_i + U_i$$

where, $X_i = X_{2i} + \cdot 10 X_{3i}$

Once we obtain $\hat{\beta}_2$ we can estimate $\hat{\beta}_3$ from the postulated relationship between β_2 and β_3. However a priori information comes from previous empirical work in which the collinearity problem happens to be less serious or from the relevant from underline the field of study.

2. Combining cross-section data and time-series data :

Pooling the data i,e combining the cross-section data and time series data is another remedial measure for the problem of multicollinearity. Suppose we want to study demand for the auto-mobiles in country

and assumed we have time series data on the numbers of cars sold, average prices of the cars and consumer income. Suppose we have here, —

$$\log Y_t = \beta_1 + \beta_2 \log P_t + \beta_3 \log I_t + u_t$$

where, Y = Number of cars sold
P = Average price
I = Income
T = Time

Now our aim is to estimate β_2 the price elasticity and β_3 the income elasticity. Now in time series data the price and the income variables tend to be collinear. Therefore if we run the preceeding regression we will be faced with the problem of multi-collinearity. Way out of this is that if we have a cross section data. we can obtain a fairly reliable estimate of the income elasticity of β_3 because in such data we are at on point of time and the prices do not change much.

3. Droping a variable :

When facial with severe multicollinearity one of the simplest way to drop one of the multi-collinear variables. But however in droping from

a model, we may be committing specification error which arises from the incorrect specification of the model used in the analysis.

4. Transformation of variables:

Suppose we have time-series data on consumption (Y), expenditure, income and wealth one reason for high multicollinearity between Y and w in such data is that over time both the variables tend to move in the same direction. If the relation,

$$Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + U_t \quad \text{——— (1)}$$

holds at time t, it must also hold at time, (t-1) because the origin of t is arbitrary any way. Therefore, we have

$$Y_{t-1} = \beta_1 + \beta_2 X_{2(t-1)} + \beta_3 X_{3(t-1)} + U_{(t-1)} \quad \text{——— (2)}$$

If we subtract (2) from (1), we have

$$Y_t - Y_{t-1} = \beta_2 (X_{2t} - X_{2,t-1}) + \beta_3 (X_{3t} - X_{3,t-1}) + V_t \quad \text{——— (3)}$$

where, $V_t = U_t - U_{t-1}$

Eq" ③ is known as the first differential from form because we run the regression not only original variables, but on the differential of successive values of the variables.

The 1st differential regression model reduces the severity of multicollinearity because although the levels of X_2 and X_3 are highly co-related, there is no a priori reason to believe their differential will also be highly collinear.

5. Additional / new data :

Since multicollinearity is a sample feature, it is possible that in another sample involving the same variables, collinearity may not be as serious as in the 1st sample. Sometimes simply increase size of the sample, we may get rid of collinearity problem.