Multicollinearity

Seema Gupta
Department of Statistics
Ram lal Anand College

Remedies

1. **Obtain more data** The harmful multicollinearity arises essentially because 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 to various reasons as follows.

- The experiment and process have finished and no longer available.
- The economic constrains may also not allow to collect 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 same for such a long period.

2. Omitting the Variables

If possible, identify the variables which seems to causing 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 that how the model will exhibit less multicollinearity.

3. Model re-specification

- One approach to model respecification is to redefine the regressors. For example, if x1, x2 and x3 are nearly linearly dependent, it may be possible to find some function such as x = (x1 +x2)/ x3 or x = x1x2x3 that preserves the information content in the original regressors but reduces the ill conditionin

4. Use of prior information

- Suppose we consider the model
- $Yi = \beta 1 + \beta 2X2i + \beta 3X3i + ui$
- where *Y = consumption, X2 = income, and X3 = wealth. As noted before,* income and wealth variables tend to be highly collinear. But suppose a priori we believe that $\beta 3 = 0.10\beta 2$*; that is, the rate of change of consumption* with respect to wealth is one-tenth the corresponding rate with respect to income. We can then run the following regression:

$$Yi = \beta 1 + \beta 2X2i + 0.01\beta 2X3i + ui$$

$$Yi = \beta 1 + \beta 2Xi$$

where *Xi = X2i + 0.1X3i. Once we obtain estimate of* $\beta 2$ *we can estimate* $\beta 3$.

How does one obtain a priori information?

- It could come from previous empirical work in which the collinearity problem happens to be less serious
- or from the relevant theory underlying the field of study

5. Combining cross-sectional and time series data

A variant of the extraneous or a priori information technique is the combination of crosssectional and time-series data, known as *pooling the data*

- To study the demand for automobiles in the United States and assume we have time series data on the number of cars sold, average price of the car, and consumer income. Suppose also that.

- $lnYi = \beta 1 + \beta 2\ ln\ Ii + \beta 3\ ln\ Pi + ui$
- where *Y = number of cars sold, P = average price, I = income, and t = time.* Out objective is to estimate the price elasticity $\beta 2$ *and income elasticity* $\beta 3$*.*
- In time series data the price and income variables generally tend to be highly collinear. Therefore, if we run the preceding regression, we shall be faced with the usual multicollinearity problem.

- If we have cross-sectional data (for example, data generated by consumer panels, or budget studies conducted by various private and governmental agencies), we can obtain a fairly reliable estimate of the income elasticity $\beta 3$ *because in such data, which are at a* point in time, the prices do not vary much. Then using an estimate of $\beta 3$ *we can write the preceding time series regression as*
- $Yi^* = \beta 1 + \beta 3\ ln\ Pi + ui$ *where* $Y^* = ln\ Y - \beta \wedge 2\ ln\ I$

- That is, *Y* represents that value of Y after removing* from it the effect of income. We can now obtain an estimate of the price elasticity $\beta 2$ *from the preceding regression.*
- Although it is an appealing technique, pooling the time series and crosssectional data in the manner just suggested may create problems of interpretation, because we are assuming implicitly that the cross-sectionally estimated income elasticity is the same thing as that which would be obtained from a pure time series analysis

6. Additional or new data

- Since multicollinearity is a sample feature, it is possible that in another sample involving the same variables collinearity may not be so serious as in the first sample. Sometimes simply increasing the size of the sample (if possible) may attenuate the collinearity problem.

7. Transforming the model

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

holds at time *t, it must also hold at time t − 1 because the origin of time is*

arbitrary anyway. Therefore, we have

$Y_{t-1} = \beta_1 + \beta_2 X_{2,t-1} + \beta_3 X_{3,t-1} + u_{t-1} \quad (2)$

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

$Y_t - Y_{t-1} = \beta_2(X_{2t} - X_{2,t-1}) + \beta_3(X_{3t} - X_{3,t-1}) + v_t \quad (4)$

where $v_t = u_t - u_{t-1}$. *Equation (4) is known as the* ***first difference***

form because we run the regression, not on the original variables, but on the differences of successive values of the variables.

The first difference regression model often reduces the severity of multicollinearity because, although the levels of *X2 and X3 may be highly correlated,* there is no a priori reason to believe that their differences will also be highly correlated.

- Another commonly used transformation in practice is the **ratio transformation.**
- Consider : $Yi = \beta 1 + \beta 2X2i + \beta 3X3i + ui \quad (1)$
- where *Y is consumption expenditure in real dollars, X2 is GDP, and X3 is* total population. Since GDP and population grow over time, they are likely to be correlated. One "solution" to this problem is to express the model on a per capita basis, that is, by dividing (1) by *X3, to obtain*

- Such a transformation may reduce collinearity in the original variables. But the first-difference or ratio transformations are not without problems.
- For instance, the error term *vt in (2) may not satisfy one of the* assumptions of the classical linear regression model, namely, that the disturbances are serially uncorrelated.

- Therefore, the remedy may be worse than the disease. Moreover, there is a loss of one observation due to the differencing procedure, and therefore the degrees of freedom are reduced by one. In a small sample, this could be a factor one would wish at least to take into consideration. Furthermore, the first-differencing procedure may not be appropriate in cross-sectional data where there is no logical ordering of the observations.

- Similarly, in the ratio model, the error term(Ui/X3i) will be heteroscedastic, if the original error term *ut is homoscedastic*

8. Ridge regression

- Ridge Regression is a technique for analyzing multiple regression data that suffer from multicollinearity. In ridge regression, the first step is to standardize the variables (both dependent and independent) by subtracting their means and dividing by their standard deviations. Each diagonal element is then multiplied by (1+d) where d is small. We can start with very small value of d say .01 and keep on increasing it till the resulting estimates of the regression parameters are stable or do not vary much.

9. 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 as principal components. The principal components involves 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 in 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 ordinary least squares method. Since all the principal components are orthogonal, they are mutually independent and so OLS is used without any problem. Suppose there are k explanatory variables X1, X2, …, Xk. Consider the linear function li, l2, … lk say
- L1 = a1X1 +a2X2+… + akXk
- L2 = b1X1 + b2X2 +…. +bkXk etc.

We choose ai's such that variance of li is maximum subject to the condition that a1^2+a2^2+…+ak^2 = 1 (called normalization condition). Then l1 is said to be the first component.

- It is a linear function of x’s such that it has maximum variance. We consider l2 such that it is uncorrelated to l1 and it maximizes variance subject to condition b1^2+b2^2+…+bk^2 = 1 Then l2 is called second principal component. Following this procedure we find l1, l2, …, lk These principal components have the property that
- ∑ var (li) = ∑ var (Xi) Unlike Xi’s which are highly correlated li’s are mutually orthogonal.

- It is suggested that instead of regressing Y on Xi's regress Y on li's. But there are two problems:

(i) l1 though it picks up major portion of variance of X's need not necessarily be the one that is most correlated with Y. Infact there is no necessary relationship between order of principal components and the degree of their correlation with Y. Often l1, l2, …, lk have no meaningful economic interpretation. If we regress Y on li's and then substitute li's in

- Terms of Xi's we finally get the same answer as before. So there is a point in using principal component analysis only if we regress Y on a subset of li's.

10. Stepwise regression

11. **Reducing collinearity in polynomial regressions**

- polynomial regression models. A special feature of these models is that the explanatory variable(s) appear with various powers. Thus, in the total cubic cost function involving the regression of total cost on output, (output)2, and (output)3 the various output terms are going to be correlated, making it difficult to estimate the various slope coefficients precisely In practice though, it has been found that if the explanatory variable(s) are expressed in the deviation form (i.e., deviation from the mean value), multicollinearity is substantially reduced. But even then the problem may persist in which case one may want to consider techniques such as **orthogonal polynomials.**