



Chapter Four

Violations Of The Assumptions of The Classical Model



4.1 MULTICOLLINEARITY

WHAT HAPPENS IF THE REGRESSORS ARE CORRELATED?

- £ One of the assumption of the classical linear regression model (CLRM) is that there is no multicollinearity among the regressors included in the regression model.
- £ Here we take a critical look at this assumption by seeking answers to the following questions:
 - 1.What is the nature of multicollinearity?
 - 2.Is multicollinearity really a problem?
 - 3.What are its practical consequences?
 - 4.How does one detect it?
 - 5.What remedial measures can be taken to alleviate the problem of multicollinearity?



Why does the classical linear regression model assume that there is no multicollinearity among the

- £ The reasoning is this: **If multicollinearity is perfect, the regression coefficients of the X variables are indeterminate and their standard errors are infinite.**
- £ **If multicollinearity is less than perfect, the regression coefficients, although determinate, possess large standard errors** (in relation to the coefficients themselves), **which means the coefficients cannot be estimated with great precision or accuracy**



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) and house size (X_3) there is a physical constraint in the population in that families with higher incomes generally have larger homes than



Continued.....

- 3. Model specification**, 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.
- 5. An additional reason for multicollinearity, especially in time series data**, may be that the regressors included in the model share a common trend, that is, they all increase or

Continued.....

£ Thus, in the regression of consumption expenditure on income, wealth, and population, the regressors income, wealth, and population may all be growing over time at more or less the same rate, leading to collinearity among these variables.



PRACTICAL CONSEQUENCES OF MULTICOLLINEARITY

In cases of near or high multicollinearity, one is likely to encounter the following consequences:

1. Although BLUE, the OLS estimators have large variances and covariances, 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” (i.e., the true population coefficient is zero) more readily.



Continued.....

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 is statistically insignificant, R^2 , the overall measure of goodness of fit, can be very high.
5. The OLS estimators and their standard errors can be sensitive to small changes in the data.



Detection of Multicollinearity

- £ The presence of multicollinearity makes it difficult to separate the individual effects of the collinear variables on the dependent variable.
- £ Explanatory variables are rarely uncorrelated with each other and multicollinearity is a matter of degree not of presence!!!!!!



Detection Of Multicollinearity

1. High R^2 but few significant t ratios. As noted, this is the “classic” symptom of multicollinearity. If R^2 is high, say, in excess of 0.8, the F test in most cases will reject the hypothesis that the partial slope coefficients are simultaneously equal to zero, but the individual t tests will show that none or very few of the partial slope coefficients are statistically different from zero. This fact was clearly demonstrated by our consumption–income–wealth example.

Although this diagnostic is sensible, its disadvantage is that “it is too strong in the sense that multicollinearity is considered as harmful only when all of the influences of the explanatory variables on Y cannot be disentangled.”¹⁸



2. High pair-wise correlations among regressors.

- £ Another suggested rule of thumb is that if the pair-wise or zero-order correlation coefficient between two regressors is high, say, in excess of 0.8, then multicollinearity is a serious problem. The problem with this criterion is that, although high zero-order correlations may suggest collinearity, it is not necessary that they be high to have collinearity in any specific case.

3. Wrong signs of the regression coefficients

4. Examination of Partial correlation coefficients among independent variables.



5. Using variance inflation factor (VIF)

- VIF>10 is used to indicate the presence of multicollinearity between continuous

The speed with which variances and covariances increase can be seen with the **variance-inflating factor (VIF)**, which is defined as

$$\text{VIF} = \frac{1}{(1 - r_{23}^2)} \quad (10.5.1)$$

VIF shows how the variance of an estimator is *inflated* by the presence of multicollinearity. As r_{23}^2 approaches 1, the VIF approaches infinity. That is, as the extent of collinearity increases, the variance of an estimator increases, and in the limit it can become infinite. As can be readily seen, if there is no collinearity between X_2 and X_3 , VIF will be 1.



Remedial Measures

£ Several methodologies have been proposed to overcome the problem of multicollinearity

1) Do Nothing: sometimes multicollinearity is not necessarily bad or unavoidable. If the R-squared of the regression exceeds the R-squared of any independent variable on other variables, there shouldn't be much worry. Also, if the t-statistics are greater than 2 there should not be much problem



Remedial Measures

2. **Drop a variable(s) from the model:** this however could lead to specification error
3. **Acquiring additional information: increasing the sample size**
4. **Rethinking of the model:** incorrect choice of functional form, specification errors, etc.
5. **Transformation of variables.** Eg in to logarithmic form, forming ratios, etc



IS MULTICOLLINEARITY NECESSARILY BAD? MAYBE NOT IF THE OBJECTIVE IS PREDICTION ONLY

It has been said that if the sole 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.

However, 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 leads to large standard errors of the estimators.



SUMMARY AND CONCLUSIONS

1. One of the assumptions of the classical linear regression model is that there is no multicollinearity among the explanatory variables, the X 's. Broadly interpreted, multicollinearity refers to the situation where there is either an exact or approximately exact linear relationship among the X variables.

2. The consequences of multicollinearity are as follows: If there is perfect collinearity among the X 's, their regression coefficients are indeterminate and their standard errors are not defined. If collinearity is high but not perfect, estimation of regression coefficients is possible but their standard errors tend to be large. As a result, the population values of the coefficients cannot be estimated precisely. However, if the objective is to estimate linear combinations of these coefficients, *the estimable functions*, this can be done even in the presence of perfect multicollinearity.

3. Although there are no sure methods of detecting collinearity, there are several indicators of it, which are as follows:

- (a) The clearest sign of multicollinearity is when R^2 is very high but none of the regression coefficients is statistically significant on the basis of the conventional t test. This case is, of course, extreme.
- (b) In models involving just two explanatory variables, a fairly good idea of collinearity can be obtained by examining the zero-order, or simple, correlation coefficient between the two variables. If this correlation is high, multicollinearity is generally the culprit.
- (c) However, the zero-order correlation coefficients can be misleading in models involving more than two X variables since it is possible to have low zero-order correlations and yet find high multicollinearity. In situations like these, one may need to examine the partial correlation coefficients.
- (d) If R^2 is high but the partial correlations are low, multicollinearity is a possibility. Here one or more variables may be superfluous. But if R^2 is high and the partial correlations are also high, multicollinearity may not be readily detectable. Also, as pointed out by C. Robert, Krishna Kumar, John O'Hagan, and Brendan McCabe, there are some statistical problems with the partial correlation test suggested by Farrar and Glauber.
- (e) Therefore, one may regress each of the X_i variables on the remaining X variables in the model and find out the corresponding coefficients of determination R_i^2 . A high R_i^2 would suggest that X_i

is highly correlated with the rest of the X 's. Thus, one may drop that X_i from the model, provided it does not lead to serious specification bias.

4. Detection of multicollinearity is half the battle. The other half is concerned with how to get rid of the problem. Again there are no sure methods, only a few rules of thumb. Some of these rules are as follows: (1) using extraneous or prior information, (2) combining cross-sectional and time series data, (3) omitting a highly collinear variable, (4) transforming data, and (5) obtaining additional or new data. Of course, which of these rules will work in practice will depend on the nature of the data and severity of the collinearity problem.

5. We noted the role of multicollinearity in prediction and pointed out that unless the collinearity structure continues in the future sample it is hazardous to use the estimated regression that has been plagued by multicollinearity for the purpose of forecasting.

6. Although multicollinearity has received extensive (some would say excessive) attention in the literature, an equally important problem encountered in empirical research is that of micronumerosity, smallness of sample size. According to Goldberger, "When a research article complains about multicollinearity, readers ought to see whether the complaints would be convincing if "micronumerosity" were substituted for "multicollinearity."⁴⁶ He suggests that the reader ought to decide how small n , the number of observations, is before deciding that one has a small-sample problem, just as one decides how high an R^2 value is in an auxiliary regression before declaring that the collinearity problem is very severe.