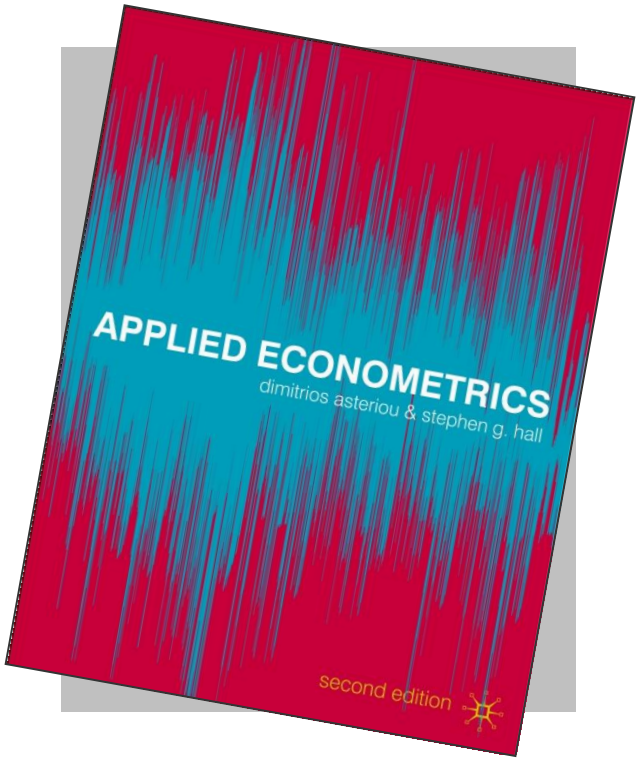
Applied Econometrics
Second edition

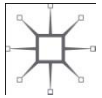
Dimitrios Asteriou and
Stephen G. Hall



palgrave
macmillan

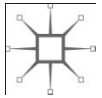
MULTICOLLINEARITY

1. Perfect Multicollinearity
2. Consequences of Perfect Multicollinearity
3. Imperfect Multicollinearity
4. Consequences of Imperfect Multicollinearity
5. Detecting Multicollinearity
6. Resolving Multicollinearity



Learning Objectives

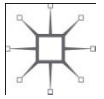
1. Recognize the problem of multicollinearity in the CLRM.
2. Distinguish between perfect and imperfect multicol-linearity.
3. Understand and appreciate the consequences of perfect and imperfect multicollinearity on OLS estimates.
4. Detect problematic multicollinearity using econometric software.
5. Find ways of resolving problematic multicollinearity.



Multicollinearity

Assumption number 8 of the CLRM requires that there are no exact linear relationships among the sample values of the explanatory variables (the Xs).

So, when the explanatory variables are very highly correlated with each other (correlation coefficients either very close to 1 or to -1) then the problem of multicollinearity occurs.



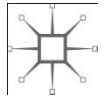
Perfect Multicollinearity

- When there is a perfect linear relationship.
- Assume we have the following model:

$$Y=\beta_1+\beta_2X_2+\beta_3X_3+e$$

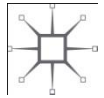
where the sample values for X_2 and X_3 are:

X_2	1	2	3	4	5	6
X_3	2	4	6	8	10	12



Perfect Multicollinearity

- We observe that $X_3=2X_2$
- Therefore, although it seems that there are two explanatory variables in fact it is only one.
- This is because X_2 is an exact linear function of X_3 or because X_2 and X_3 are perfectly collinear.



Perfect Multicollinearity

When this occurs then the equation:

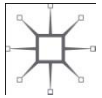
$$\delta_1 X_1 + \delta_2 X_2 = 0$$

can be satisfied for non-zero values of both δ_1 and δ_2.

In our case we have that

$$(-2)X_1 + (1)X_2 = 0$$

So $\delta_1 = -2$ and $\delta_2 = 1$.

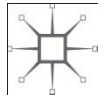


Perfect Multicollinearity

Obviously if the only solution is

$$\delta_1=\delta_2=0$$

(usually called as the trivial solution) then the two variables are linearly independent and there is no problematic multicollinearity.



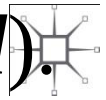
Perfect Multicollinearity

In case of more than two explanatory variables the case is that one variable can be expressed as an exact linear function of one or more or even all of the other variables.

So, if we have 5 explanatory variables we have:

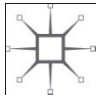
$$\delta_1 X_1 + \delta_2 X_2 + \delta_3 X_3 + \delta_4 X_4 + \delta_5 X_5 = 0$$

An application to better understand this situation is the Dummy variables trap (*explain on board*).



Consequences of Perfect Multicollinearity

- Under Perfect Multicollinearity, the OLS estimators simply **do not exist**. (*prove on board*)
- If you try to estimate an equation in Eviews and your equation specifications suffers from perfect multicollinearity Eviews will not give you results but will give you an error message mentioning multicollinearity in it.

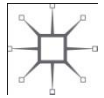


Imperfect Multicollinearity

- Imperfect multicollinearity (or near multicollinearity) exists when the explanatory variables in an equation are correlated, but this correlation is **less than** perfect.
- This can be expressed as:

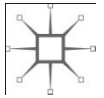
$$X_3 = X_2 + v$$

where v is a random variable that can be viewed as the 'error' in the exact linear releationship.



Consequences of Imperfect Multicollinearity

- In cases of imperfect multicollinearity the OLS estimators can be obtained and they are also BLUE.
- However, although linear unbiassed estimators with the minimum variance property to hold, the OLS variances are often larger than those obtained in the absence of multicollinearity.



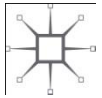
Consequences of Imperfect Multicollinearity

To explain this consider the expression that gives the variance of the partial slope of variable X_j:

$$\text{var}(\hat{\beta}_2) = \frac{\sigma^2}{\sum (X_2 - \bar{X}_2)^2 (1 - r^2)}$$

$$\text{var}(\hat{\beta}_3) = \frac{\sigma^2}{\sum (X_3 - \bar{X}_3)^2 (1 - r^2)}$$

where r^2 is the square of the sample correlation coefficient between X_2 and X_3.



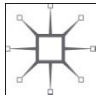
Consequences of Imperfect Multicollinearity

Extending this to more than two explanatory variables, we have:

$$\text{var}(\hat{\beta}_j) = \frac{\sigma^2}{\sum (X_2 - \bar{X}_2)^2 (1 - R_j^2)}$$

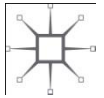
$$\text{var}(\hat{\beta}_3) = \frac{\sigma^2}{\sum (X_3 - \bar{X}_3)^2} \frac{1}{(1 - R_j^2)}$$

and therefore, what we call the Variance Inflation Factor (VIF)



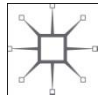
Variance Inflation

R^2_j	VIF_j
0	1
0.5	2
0.8	5
0.9	10
0.95	20
0.075	40
0.99	100
0.995	200
0.999	1000



The Variance Inflation Factor

- VIF values that exceed 10 are generally viewed as evidence of the existence of problematic multicollinearity.
- This happens for $R^2_j > 0.9$ (explain auxiliary reg)
- So large standard errors will lead to large confidence intervals.
- Also, we might have t-stats that are totally wrong.



Consequences of Imperfect Multicollinearity (Again)

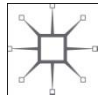
Concluding when imperfect multicollinearity is present we have:

(a) Estimates of the OLS may be imprecise because of large standard errors.

(b) Affected coefficients may fail to attain statistical significance due to low t-stats.

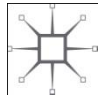
(c) Sing reversal might exist.

(d) Addition or deletion of few observations may result in substantial changes in the estimated coefficients.



Detecting Multicollinearity

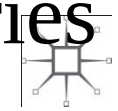
- The easiest way to measure the extent of multicollinearity is simply to look at the matrix of correlations between the individual variables.
- In cases of more than two explanatory variables we run the auxiliary regressions. If near linear dependency exists, the auxiliary regression will display a small equation standard error, a large R^2 and statistically significant F-value.



Resolving Multicollinearity

- Approaches, such as the ridge regression or the method of principal components. But these usually bring more problems than they solve.

- Some econometricians argue that if the model is otherwise OK, just ignore it. Note that you will always have some degree of multicollinearity, especially in time series data.



Resolving Multicollinearity

- The easiest ways to "cure" the problems are

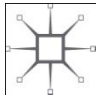
(a) drop one of the collinear variables

(b) transform the highly correlated variables into a ratio

(c) go out and collect more data e.g.

(d) a longer run of data

(e) switch to a higher frequency



Examples

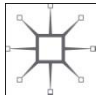
We have quarterly data for

Imports (IMP)

Gross Domestic Product (GDP)

Consumer Price Index (CPI) and

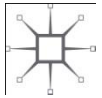
Producer Price Index (PPI)



Examples

Correlation Matrix

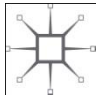
	IMP	GDP	CPI	PPI
IMP	1	0.979	0.916	0.883
GDP	0.979	1	0.910	0.899
CPI	0.916	0.910	1	0.981
PPI	0.883	0.8998	0.981	1



Examples – only CPI

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.631870	0.344368	1.834867	0.0761
LOG(GDP)	1.926936	0.168856	11.41172	0.0000
LOG(CPI)	0.274276	0.137400	1.996179	0.0548

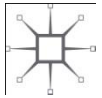
R-squared	0.966057	Mean dependent var	10.81363
Adjusted R-squared	0.963867	S.D. dependent var	0.138427
S.E. of regression	0.026313	Akaike info criterion	-4.353390
Sum squared resid	0.021464	Schwarz criterion	-4.218711
Log likelihood	77.00763	F-statistic	441.1430
Durbin-Watson stat	0.475694	Prob(F-statistic)	0.000000



Examples –CPI with PPI

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.213906	0.358425	0.596795	0.5551
LOG(GDP)	1.969713	0.156800	12.56198	0.0000
LOG(CPI)	1.025473	0.323427	3.170645	0.0035
LOG(PPI)	-0.770644	0.305218	-2.524894	0.0171

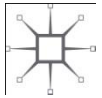
R-squared	0.972006	Mean dependent var	10.81363
Adjusted R-squared	0.969206	S.D. dependent var	0.138427
S.E. of regression	0.024291	Akaike info criterion	-4.487253
Sum squared resid	0.017702	Schwarz criterion	-4.307682
Log likelihood	80.28331	F-statistic	347.2135
Durbin-Watson stat	0.608648	Prob(F-statistic)	0.000000



Examples – only PPI

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.685704	0.370644	1.850031	0.0739
LOG(GDP)	2.093849	0.172585	12.13228	0.0000
LOG(PPI)	0.119566	0.136062	0.878764	0.3863

R-squared	0.962625	Mean dependent var	10.81363
Adjusted R-squared	0.960213	S.D. dependent var	0.138427
S.E. of regression	0.027612	Akaike info criterion	-4.257071
Sum squared resid	0.023634	Schwarz criterion	-4.122392
Log likelihood	75.37021	F-statistic	399.2113
Durbin-Watson stat	0.448237	Prob(F-statistic)	0.000000



Examples – the auxiliary regression

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-0.542357	0.187073	-2.899177	0.0068
LOG(CPI)	0.974766	0.074641	13.05946	0.0000
LOG(GDP)	0.055509	0.091728	0.605140	0.5495

R-squared	0.967843	Mean dependent var	4.552744
Adjusted R-squared	0.965768	S.D. dependent var	0.077259
S.E. of regression	0.014294	Akaike info criterion	-5.573818
Sum squared resid	0.006334	Schwarz criterion	-5.439139
Log likelihood	97.75490	F-statistic	466.5105
Durbin-Watson stat	0.332711	Prob(F-statistic)	0.000000