MORE ON MULTICOLLINEARITY (MC)

Variance Inflation Factor (VIF) and **Tolerance** are two measures that can guide a researcher in identifying MC. Before developing the concepts, it should be noted that the variance of the OLS estimator for a typical regression coefficient (say b_i) can be shown to be the following [see Wooldridge (2000), Chapter 3 appendix for proof].

$$\text{Var}(\hat{b}_i) = \frac{s^2}{S_{ii}(1-R_i^2)}$$

where $S_{ii} = \sum_{j=1}^{n}(X_{ij} - \overline{X}_i)^2$ and R_i^2 is the unadjusted R^2 when you regress X_i against all the other explanatory variables in the model, that is, against a constant, X_2, X_3, …., X_{i-1}, X_{i+1}, …., X_k. Suppose there is *no* linear relation between X_i and the other explanatory variables in the model. Then, R_i^2 will be zero and the variance of $\hat{b}_i$ will be s^2 / S_{ii}. Dividing this into the above expression for Var($\hat{b}_i$), we obtain the variance inflation factor and tolerance as

$$\text{VIF}(\hat{b}_i) = \frac{1}{1-R_i^2} \qquad \text{Tolerance}(\hat{b}_i) = 1/\text{VIF} = 1-R_i^2$$

It is readily seen that the higher VIF or the lower the tolerance index, the higher the variance of $\hat{b}_i$ and the greater the chance of finding b_i insignificant, which means that severe MC effects are present. Thus, these measures can be useful in identifying MC. The procedure is to choose each right hand side variable (that is, explanatory variable) as the dependent variable and regress it against a constant and the remaining explanatory variables. We would thus get k–1 values for VIF. If any of them is high, then MC is indicated. Unfortunately, however, there is no theoretical way to say what the threshold value should be to judge that VIF is “high.” Also, there is no theory that tells you what to do if MC is found.

Example

This example revisits the application in Section 5.4 using DATA4-6 (see Table 5.3) to illustrate how the above methodology can be applied. The original model is

$$\text{povrate} = b_1 + b_2\text{urb} + b_3\text{famsize} + b_4\text{unemp} + b_5\text{highschl} + b_6\text{college} + b_7\text{medinc} + \text{u}$$

The estimates for this model are in Table 5.3 as Model 1 and are reproduced below. As can be seen, several coefficients are insignificant suggesting the possibility of MC.

MODEL 1: Dependent variable: povrate

VARIABLE	COEFFICIENT	STDERROR	T STAT	2Prob(t > |T|)	
const	16.8176	8.5026	1.978	0.053350	*
urb	-0.0187	0.0148	-1.270	0.210010	
famsize	6.0918	1.8811	3.238	0.002116	***
unemp	-0.0118	0.1195	-0.099	0.921724	
highschl	-0.1186	0.0681	-1.741	0.087742	*
college	0.1711	0.0982	1.743	0.087355	*
medinc	-0.5360	0.0704	-7.619	0.000000	***

Mean of dep. var.	9.903	S.D. of dep. variable	3.955
Error Sum of Sq (ESS)	146.0911	Std Err of Resid. (sgmahat)	1.6925
Unadjusted R-squared	0.836	Adjusted R-squared	0.817
F-statistic (6, 51)	43.3875	p-value for F()	0.000000

MODEL SELECTION STATISTICS

SGMASQ	2.86453	AIC	3.20646	FPE	3.21025
HQ	3.53259	SCHWARZ	4.11172	SHIBATA	3.1268
GCV	3.2577	RICE	3.32025		

It was noted in Table 5.3 that perhaps medium income (medinc), though significant, does not belong in the model because it is determined by famsize, unemp, highschl, and college. It therefore makes sense to omit this variable from the model specification. The revised model estimates are given below.

MODEL 2: Dependent variable: povrate

VARIABLE	COEFFICIENT	STDERROR	T STAT	2Prob(t > |T|)	
const	39.0423	11.5651	3.376	0.001399	***
urb	-0.0340	0.0212	-1.607	0.114191	
famsize	-2.1526	2.2281	-0.966	0.338450	
unemp	0.2044	0.1680	1.217	0.229239	
highschl	-0.2980	0.0925	-3.221	0.002204	***
college	-0.3759	0.0969	-3.878	0.000297	***

Error Sum of Sq (ESS)	312.3529	Std Err of Resid. (sgmahat)	2.4509
Unadjusted R-squared	0.650	Adjusted R-squared	0.616
F-statistic (5, 52)	19.2931	p-value for F()	0.000000
Durbin-Watson stat.	2.070	First-order autocorr. coeff	-0.044

MODEL SELECTION STATISTICS

SGMASQ	6.00679	AIC	6.62326	FPE	6.62818
HQ	7.19663	SCHWARZ	8.19674	SHIBATA	6.49961
GCV	6.69988	RICE	6.79028		

MC might still be present and hence the next step is to regress each explanatory variable against all the other right hand side variables and compute the tolerance $(1-R^2)$ and VIF. The following table has these values.

Dependent Variable	Independent Variables	Tolereance	VIF
urb	constant, famsize, unemp, highschl, college, medinc	0.608	1.645
famsize	constant, urb, unemp, highschl, college, medinc	0.245	4.082
unemp	constant, urb, famsize, highschl, college, medinc	0.228	4.386
highschl	constant, urb, famsize, unemp, college, medinc	0.280	3.571
college	constant, urb, famsize, unemp, highschl, medinc	0.088	11.364
medinc	constant, urb, famsize, unemp, highschl, college	0.164	6.098

All the regressions except the first one have low tolerance and high values for VIF indicating a high degree of MC. It therefore makes sense to omit variables with insignificant coefficients, but one at a time. In Model 2, the coefficient for famsize is the least significant and hence it is omitted first (in the belief that the coefficient is closest to zero and that the "omitted variable bias" will be minimal). yielding the following results.

MODEL 3: Dependent variable: povrate

VARIABLE	COEFFICIENT	STDERROR	T STAT	2Prob(t > |T|)	
const	30.1267	6.9665	4.324	0.000068	***
urb	-0.0438	0.0186	-2.360	0.022014	**
unemp	0.1860	0.1668	1.115	0.269949	
highschl	-0.2415	0.0716	-3.372	0.001399	***
college	-0.3554	0.0945	-3.760	0.000425	***

Error Sum of Sq (ESS)	317.9597	Std Err of Resid. (sgmahat)	2.4493
Unadjusted R-squared	0.643	Adjusted R-squared	0.617

Excluding the constant, p-value was highest for variable 4 (unemp).

Next omit unemp from the model.

MODEL 4: Dependent variable: povrate

VARIABLE	COEFFICIENT	STDERROR	T STAT	2Prob(t > |T|)	
const	36.7290	3.6771	9.989	0.000000	***
urb	-0.0493	0.0180	-2.744	0.008227	***
highschl	-0.2910	0.0563	-5.173	0.000003	***
college	-0.4466	0.0476	-9.390	0.000000	***

Error Sum of Sq (ESS)	325.4159	Std Err of Resid. (sgmahat)	2.4548
Unadjusted R-squared	0.635	Adjusted R-squared	0.615

MODEL SELECTION STATISTICS

SGMASQ	6.02622	AIC	6.44041	FPE	6.44182
HQ	6.80694	SCHWARZ	7.42381	SHIBATA	6.3845
GCV	6.47261	RICE	6.50832		

Comparing this Model 4 with Model 4 in Table 5.3, we note two things. First all model selection statistics are better here than in Table 5.3. Second, the above model has a very strongly significant coefficient for urb whereas in Table 5.3 urb was replaced by famsize with a relatively week significance. Therefore, overall this Model 4 is superior.

References

Greene, W.H., *Econometric Analysis*, Fourth Edition, Prentic-Hall, Upper Saddle River, New Jersey, 2000.

Wooldridge, J. M., *Introductory Econometrics: A Modern Approach*, South Western, 2000.