

EE325 HW 6 (Multicollinearity, Heteroscedasticity, and Autocorrelation)

Answers

Multicollinearity

1. State with reason whether the following statements are true, false, or uncertain:
 - a. Despite perfect multicollinearity, OLS estimators are BLUE.
False. If exact linear relationship(s) exist among variables, we cannot even estimate the coefficients or their standard errors.
 - b. In cases of high multicollinearity, it is not possible to assess the individual significance of one or more partial regression coefficients.
False. One may be able to obtain one or more significant t values.
 - c. High pair-wise correlations do not suggest that there is high multicollinearity.
Uncertain. If a model has only two regressors, high pairwise correlation coefficients may suggest multicollinearity. If one or more regressors enter non-linearly, the pairwise correlations may give misleading answers.
 - d. You will not obtain a high R-squared value in a multiple regression if all the partial slope coefficients are individually statistically insignificant on the basis of the usual t-test.
False. One usually obtains high R²'s in models with highly correlated regressors.
 - e. Ceteris Paribus, the higher the VIF is, the larger the variances of OLS estimators.
False. The variance of an OLS estimator is given by the following formula:

$$\text{var}(\hat{\beta}_j) = \frac{\sigma^2}{\sum x_j^2} \left(\frac{1}{1 - R_j^2} \right)$$

A high R_j² can be counterbalanced by a low σ² or high ∑ x_j²

Empirical Exercises (STATA exercise) ☺☺

2. Table 10.3 gives data on import, GDP, and the Consumer Price Index (CPI) for the United States over the period 1975-2005. You are asked to consider the following model:

$$\ln \text{imports}_t = \beta_1 + \beta_2 \ln \text{GDP}_t + \beta_3 \ln \text{CPI}_t + u_t$$

a. Estimate the parameters of this model using the data

Source	SS	df	MS			
Model	17.2842182	2	8.64210911	Number of obs =	31	
Residual	.139293473	28	.004974767	F(2, 28) =	1737.19	
Total	17.4235117	30	.580783723	Prob > F =	0.0000	
				R-squared =	0.9920	
				Adj R-squared =	0.9914	
				Root MSE =	.07053	

ln imports	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
ln gdp	1.850098	.1829119	10.11	0.000	1.47542	2.224777
ln cpi	-.873369	.2848058	-3.07	0.005	-1.456767	-.2899708
_cons	1.409415	.2700745	5.22	0.000	.8561926	1.962638

b. Do you suspect that there is multicollinearity in the data?

Judged by the high R^2 value and the negative coefficient on the log CPI variable, there *might* some multicollinearity in the data.

c. Regress:

$$(1) \ln imports_t = A_1 + A_2 \ln GDP_t$$

$$(2) \ln imports_t = B_1 + B_2 \ln CPI_t$$

$$(3) \ln GDP_t = C_1 + C_2 \ln CPI_t$$

On the basis of these regressions, what can you say about the nature of multicollinearity in the data?

(1) $\ln imports_t = A_1 + A_2 \ln GDP_t$

Source	SS	df	MS			
Model	17.2374371	1	17.2374371	Number of obs =	31	
Residual	.1860746	29	.006416366	F(1, 29) =	2686.48	
Total	17.4235117	30	.580783723	Prob > F =	0.0000	
				R-squared =	0.9893	
				Adj R-squared =	0.9890	
				Root MSE =	.0801	

ln imports	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
ln gdp	1.293252	.0249512	51.83	0.000	1.242221	1.344283
_cons	2.00191	.2143083	9.34	0.000	1.5636	2.440219

$$(2) \ln imports_t = B_1 + B_2 \ln CPI_t$$

Source	SS	df	MS			
Model	16.7752643	1	16.7752643	Number of obs =	31	
Residual	.648247373	29	.022353358	F(1, 29) =	750.46	
Total	17.4235117	30	.580783723	Prob > F =	0.0000	
				R-squared =	0.9628	
				Adj R-squared =	0.9615	
				Root MSE =	.14951	

ln imports	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
ln cpi	1.986499	.0725145	27.39	0.000	1.83819	2.134808
_cons	3.578294	.3480571	10.28	0.000	2.866437	4.290151

$$(3) \ln GDP_t = C_1 + C_2 \ln CPI_t$$

Source	SS	df	MS			
Model	10.1576897	1	10.1576897	Number of obs =	31	
Residual	.148692396	29	.005127324	F(1, 29) =	1981.09	
Total	10.3063821	30	.34354607	Prob > F =	0.0000	
				R-squared =	0.9856	
				Adj R-squared =	0.9851	
				Root MSE =	.07161	

ln gdp	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
ln cpi	1.545792	.0347295	44.51	0.000	1.474763	1.616822
_cons	1.172304	.1666957	7.03	0.000	.8313734	1.513236

The auxiliary regression of LN_GDP on LN_CPI shows that the two variables are highly correlated, perhaps suggesting that the data suffer from the collinearity problem.

- d. Suppose there is multicollinearity in the data but $\hat{\beta}_2$ and $\hat{\beta}_3$ are individually significant at the 5 percent level and the overall F-test is also significant. In this case should we worry about the collinearity problem?

The best solutions here would be to express imports and GDP in real terms by dividing each by CPI (the ratio method). The results are as follows:

Source	SS	df	MS			
Model	4.64571883	1	4.64571883	Number of obs =	31	
Residual	.139534936	29	.00481155	F(1, 29) =	965.53	
Total	4.78525377	30	.159508459	Prob > F =	0.0000	
				R-squared =	0.9708	
				Adj R-squared =	0.9698	
				Root MSE =	.06937	

ln mpcpi	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
ln gdp cpi	1.811942	.0583123	31.07	0.000	1.69268	1.931204
_cons	1.442444	.2210172	6.53	0.000	.9904132	1.894475

Heteroscedasticity

3. State with brief reason whether the following statements are true, false, or uncertain
- In the presence of heteroscedasticity OLS estimators are biased as well as inefficient.
False. The estimators are unbiased but are inefficient.
 - If heteroscedasticity is present, the conventional t and F tests are invalid
True.
 - In the presence of heteroscedasticity the usual OLS method always overestimates the standard errors of estimators.
False. Typically, but not always, will the variance be overestimated.
 - If residuals estimated from an OLS regression exhibit a systematic pattern, it means heteroscedasticity is present in the data
False. Besides heteroscedasticity, such a pattern may result from autocorrelation, model specification errors, etc.

Empirical Exercises (STATA exercise) ☺☺

4. Table 11.7 gives data on 81 cars about MPG (average miles per gallons), HP (engine horsepower), VOL (cubic feet of cab space), SP (top speed, miles per hour), and WT (vehicle weight in 100 lbs)
- Consider the following model:

$$MPG_i = \beta_1 + \beta_2 SP_i + \beta_3 HP_i + \beta_4 WT_i + u_i$$

Estimate the parameters of this model and interpret the results.

Source	SS	df	MS			
Model	7141.40496	3	2380.46832	Number of obs =	81	
Residual	947.498564	77	12.3051762	F(3, 77) =	193.45	
Total	8088.90352	80	101.111294	Prob > F =	0.0000	
				R-squared =	0.8829	
				Adj R-squared =	0.8783	
				Root MSE =	3.5079	

mpg	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
sp	-1.271697	.2331174	-5.46	0.000	-1.735894 - .8075013
hp	.3904334	.0762458	5.12	0.000	.2386086 .5422582
wt	-1.903273	.1855158	-10.26	0.000	-2.272682 -1.533864
_cons	189.9597	22.52879	8.43	0.000	145.0991 234.8202

As expected, MPG is positively related to HP and negatively related to speed and weight.

b. Use the white test to find out if the error variance is heteroscedastic.

H_0 : Homoscedastic

H_1 : Heteroscedastic

White's general test statistic: 37.65619

Compare with chi-square distribution table at degree of freedom 9, the critical value is 16.9190

Reject the null hypothesis

That is, there is heteroscedasticity.

Source	SS	df	MS			
Model	7141.40496	3	2380.46832	Number of obs =	81	
Residual	947.498564	77	12.3051762	F(3, 77) =	193.45	
Total	8088.90352	80	101.111294	Prob > F =	0.0000	
				R-squared =	0.8829	
				Adj R-squared =	0.8783	
				Root MSE =	3.5079	

mpg	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
sp	-1.271697	.2331174	-5.46	0.000	-1.735894	-.8075013
hp	.3904334	.0762458	5.12	0.000	.2386086	.5422582
wt	-1.903273	.1855158	-10.26	0.000	-2.272682	-1.533864
_cons	189.9597	22.52879	8.43	0.000	145.0991	234.8202

. whi tetst

White's general test statistic : 37.65619 Chi-sq(9) P-value = 2.0e-05

Autocorrelation

5. State whether the following statements are true or false. Briefly justify your answer
 - a. When autocorrelation is present, OLS estimators are biased as well as inefficient.
False. The estimators are unbiased but they are not efficient.
 - b. The Durbin-Watson d test assumes that the variance of the error term u_t is homoscedastic.
True. We are still retaining the other assumptions of CLRM.
 - c. The R-squared values of two models, one involving regression in the first-difference form and another in the level form, are not directly comparable.
True. To compare R^2 s, the regressand in the two models must be the same.
 - d. The exclusion of an important variable(s) from a regression model may give a significant d value.
True.

Empirical Exercises (STATA exercise) ☺☺

6. Refer to the data on the copper industry given in table 12.7
 a. From these data estimate the following regression model:

$$\ln C_t = \beta_1 + \beta_2 \ln I_t + \beta_3 \ln L_t + \beta_4 \ln H_t + \beta_5 \ln A_t + u_t$$

Interpret the results.

Source	SS	df	MS			
Model	5.42774163	4	1.35693541	Number of obs =	30	
Residual	.370572985	25	.014822919	F(4, 25) =	91.54	
Total	5.79831462	29	.199941883	Prob > F =	0.0000	
				R-squared =	0.9361	
				Adj R-squared =	0.9259	
				Root MSE =	.12175	

Inc	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
lni	.4675086	.1659868	2.82	0.009	.1256524 .8093647
lnl	.2794423	.1147258	2.44	0.022	.0431602 .5157244
lnh	-.0051515	.142947	-0.04	0.972	-.2995564 .2892534
lna	.4414491	.1065083	4.14	0.000	.222091 .6608071
_cons	-1.500441	1.00302	-1.50	0.147	-3.5662 .5653184

As you can see, the coefficients of I , L and A are individually statistically significant and have the economically meaningful impact on C .

- b. Estimate the Durbin-Watson d statistic and comment on the nature of autocorrelation present in the data

H_0 : No positive autocorrelation

H_0^* : No negative autocorrelation

H_1 : otherwise

H_1^* : otherwise

Durbin-Watson d -statistic(5, 30) = .9549404

The d statistic is 0.955. Now for $n = 30$, $k' = 4$ and $\alpha = 5\%$, the lower limit of d is 1.138. Since the computed d value is below this critical d value, there is evidence of positive first-order autocorrelation.