Introductory Econometrics
Tutorial 10 Solutions

PART A: To be done before you attend the tutorial. The solutions will be made available at the end of the week.

1. a) Which Classical Linear Regression Model assumption is violated under the presence of heteroskedasticity? Briefly explain.

i. No perfect multicollinearity

ii. The error term has equal variance

iii. The error term has zero conditional mean

iv. The OLS estimators have equal variance

b) What is the effect of the presence of heteroskedasticity on the properties of OLS estimators? Briefly explain.

i. The OLS estimators are inconsistent.

ii. The usual F statistic no longer has an F distribution.

iii. The OLS estimators are no longer efficient.

Answer

1. a) ii. The CLRM assumptions in i and iii are not affected by the presence of heteroskedasticity. It is the assumption of constant variance in the error term that is violated. iv is not an CLRM assumption and is irrelevant.

b) iii. The usual standard errors produced via OLS are no longer valid. This implies that the OLS estimates no longer have minimum variance or are no longer efficient. That is, they are not BLUE.

2. We wish to study the hypothesis that the effect on trade tax revenue from higher trade volume is positive, while an increase in income would result in the government collecting more direct taxes (e.g. income tax) than rely on trade taxes. For this purpose, we run the following regression of the ratio of trade (import and export) taxes to total government revenue, on the ratio of the sum of exports and imports to GNP (Gross National Product) and GNP per capita, based on cross-sectional data on 41 countries (all in log form):

$$\ln Taxes_i = \beta_0 + \beta_1 \ln Trade_i + \beta_2 \ln GNP_i + u_i. \tag{1}$$

a) Intuitively, given the information provided would you expect heteroskedasticity in the error term?

b) You are asked to apply White's heteroskedasticity test in order to formally verify the existence of heteroskedasticity in this setting. What regression would you run?

c) Assuming that the R^2 of the regression in b) is equal to 0.11148, compute White's heteroskedasticity test statistic. What conclusion do you draw?

d) Given the number of restrictions imposed, what issue might you detect when applying White's heteroskedasticity test?

Answer

2. a) One would expect to find heteroskedasticity in the error variance since the data are cross-sectional involving a heterogeneity of countries.

b) We first run (1) and obtain residuals denoted by $\hat{u}_i$. In order to apply White's heteroskedasticity test we run the following regression:

$$\begin{aligned}\hat{u}_i^2 &= a_0 + a_1 \ln Trade_i + a_2 \ln GNP_i \\ &\quad + a_3 (\ln Trade_i)^2 + a_4 (\ln GNP_i)^2 \\ &\quad + a_5 (\ln Trade_i)(\ln GNP_i) + e_i.\end{aligned}$$

in other words, it is a regresion of the squared residuals from (1) on the explanatory variables, their squares and pairwise cross-products.

c) White's heteroskedasticity test statistic is given by $n \times R^2_{\hat{u}_i^2}$, where n is the number of observations in the regression. Thus, we have $n \times R^2_{\hat{u}_i^2} = 41 \times 0.11148 = 4.7068$, which asymptotically has a chi-square distribution with 5df. The corresponding critical value at 5% level is 11.0705 and at 10% level is 9.2363. One concludes that on the basis of the White test, there is no heteroskedasticity.

d) In this case we have 5 regressors in the auxiliary regression, thus consuming degrees of freedom. Instead, one can run the auxiliary regression using fitted values of $\ln Taxes_i$ from (1) and its squares as proxies of the original regressors, i.e. $\ln \widehat{Taxes}_i$ and $(\ln \widehat{Taxes}_i)^2$.

3. We wish to study the relationship between compensation and employment size using the data in wages.wf1. Compensation is measured as average compensation per employee in \$. Employment size comprises of 9 categories such that 1 (1-4 employees), 2 (5-9 employees), ..., 9 (1000-2499 employees). We are also given, σ_i, the standard deviations of wages.

a) Run the OLS regression of compensation on employment size. Comment on your results.

b) Rerun the regression in a) using weighted least squares. How are your variables of a) transformed? Compare your regression output with that in a).

Answer

3. a) We run the following unweighted OLS regression:

$$Y_i = a_0 + a_1 X_i + u_i,$$

where Y_i represents average compensation per employee and X_i is employment size. The Eviews output is given below:

Dependent Variable: COMP
Method: Least Squares
Date: 08/22/17 Time: 11:41
Sample: 1 9
Included observations: 9

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	3417.833	81.13632	42.12458	0.0000
SIZE	148.7667	14.41630	10.31790	0.0000

R-squared	0.938304	Mean dependent var	4161.667
Adjusted R-squared	0.929490	S.D. dependent var	420.5954
S.E. of regression	111.6837	Akaike info criterion	12.46235
Sum squared resid	87312.73	Schwarz criterion	12.50618
Log likelihood	-54.08057	Hannan-Quinn criter.	12.36777
F-statistic	106.4591	Durbin-Watson stat	1.221934
Prob(F-statistic)	0.000017		

From these results, it appears that average compensation and employment size have a positive association. However, since we have been given the variances of wages we can see that these are not constant.

b) Given the presence of heteroskedasticity and since we know σ_i we run weighted least squares on the following transformed regression:

$$\frac{Y_i}{\sigma_i} = a_0^* \left(\frac{1}{\sigma_i}\right) + a_1^* \left(\frac{X_i}{\sigma_i}\right) + \frac{u_i}{\sigma_i},$$

or

$$Y_i^* = a_0^* C_{0i}^* + a_1^* X_i^* + u_i^*.$$

The Eviews output of the WLS regression is given below:

Dependent Variable: COMP_S
Method: Least Squares
Date: 08/22/17 Time: 11:42
Sample: 1 9
Included observations: 9

Variable	Coefficient	Std. Error	t-Statistic	Prob.
W	3406.640	80.98322	42.06600	0.0000
SIZE_S	154.1526	16.95929	9.089565	0.0000

R-squared	0.964537	Mean dependent var	4.373494
Adjusted R-squared	0.959471	S.D. dependent var	0.671406
S.E. of regression	0.135167	Akaike info criterion	-0.971485
Sum squared resid	0.127890	Schwarz criterion	-0.927658
Log likelihood	6.371684	Hannan-Quinn criter.	-1.066065
Durbin-Watson stat	1.183941		

Compared with the OLS regression output, the standard error of the transformed employment size variable is larger than that of the original OLS regression. Consequently, the estimated t values obtained by WLS are smaller than those obtained by OLS. Despite these differences, the employment size regressor is statistically significant at 1%, 5% and 10% confidence level under both approaches.

Do not forget to bring your answers to PART A and a copy of the tutorial questions to your tutorial.

<u>Part B:</u> This part will be covered in the tutorial. It is still a good idea to attempt these questions before the tutorial.

This question is based on the work of James Tobin (winner of the Nobel prize in economics in 1981) on family food consumption, which was published in the Journal of the Royal Statistical Society in 1950. It is based on the US 1941 Family Budget Survey data set, which contains data on food consumption expenditure, income and number of people in the family, for some randomly selected families. However, data for each individual family was not reported. Instead, families were grouped according to the number of people in the family (1,2,3,4,5+) and their income range (0-500, 500-1000, 1000-1500, 1500-2000, 2000-2500, 2500-3000, 3000-4000, 4000+), and only average number of people in the family (nf), average income (inc) and average expenditure on food ($food$) for families in each group were reported (a total of 37 observations). The number of families in each group (ng) was also known. In a conference in his honour in 1997, James Tobin said that it took him two to three days in 1949 to run a regression with 3 explanatory variables! The data is in Tobin.wf1.

a. Suppose that food consumption is related to family income and number of people in each family according to the following model.

$$food_i = \beta_0 + \beta_1\, inc_i + \beta_2\, nf_i + u_i \tag{2}$$

Argue that even if the above model is a fair description of each family's demand for food and all Gauss-Markov assumptions are satisfied for this model, we may expect heteroskedastic errors when we estimate the model using group averaged data. Show that it is logical to expect that variance of errors are inversely proportional to the number of families in each group, i.e., $\text{Var}(u_i) = \frac{\sigma^2}{ng_i}$.

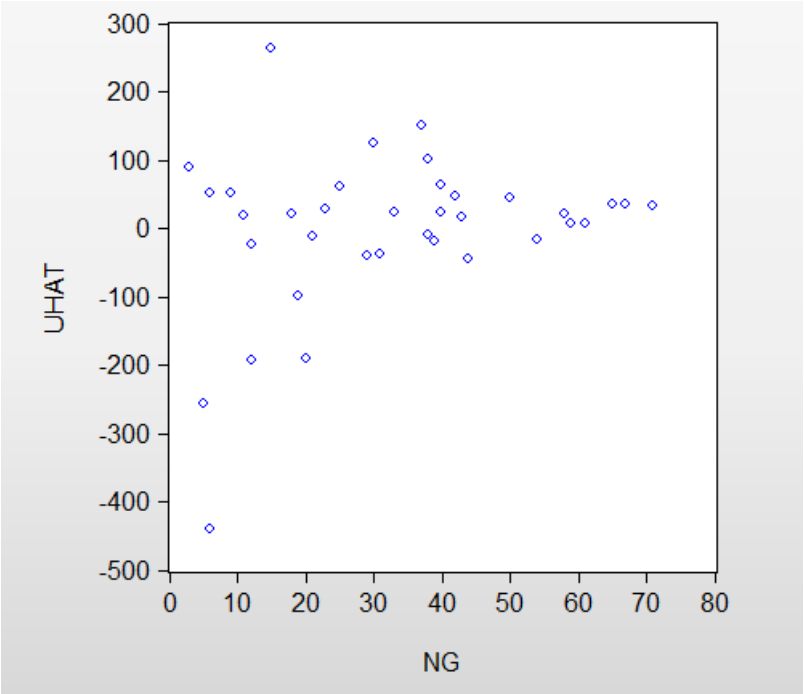
We know that the variance of average of n independent observations is $\frac{\sigma^2}{n}$. Since $food_i$ is the average food consumption for a group comprising ng_i households, the variance of $food_i$ will be $\frac{\sigma^2}{ng_i}$. And since different groups have different number of households (i.e. ng_i is not the same across i), then we are likely to have heteroskedasticity here caused by the way that data are reported.

b. If errors are heteroskedastic, would OLS be unbiased? Explain.
Yes. The assumption of homoskedasticity is not needed for unbiasedness, therefore OLS is unbiased even when errors are heteroskedastic.

c. Estimate equation (2) using OLS. Then explore visually if the error variance is likely to be inversely proportional to the number of families in each group (ng), and then formally test the hypothesis of no heteroskedasticity against the alternative that $\text{Var}(u_i) = \frac{\sigma^2}{ng_i}$. (Hint: Breusch-Pagan auxiliary regression would be a regression of $\hat{u}^2$ on a constant and $\frac{1}{ng}$)

Dependent Variable: FOOD
Method: Least Squares
Sample: 1 37
Included observations: 37

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	73.14384	48.74665	1.500490	0.1427
INC	0.164541	0.012276	13.40395	0.0000
NF	59.62959	11.42616	5.218691	0.0000
R-squared	0.872122	Mean dependent var		662.0270



Heteroskedasticity Test: Breusch-Pagan-Godfrey

F-statistic	6.577799	Prob. F(1,35)	0.0148
Obs*R-squared	5.853570	Prob. Chi-Square(1)	0.0155
Scaled explained SS	15.78144	Prob. Chi-Square(1)	0.0001

Test Equation:
Dependent Variable: RESID^2
Method: Least Squares
Included observations: 37

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	1372.979	7123.362	0.192743	0.8483
1/NG	212500.3	82855.09	2.564722	0.0148

R-squared	0.158205	Mean dependent var	13550.23
F-statistic	6.577799		
Prob(F-statistic)	0.014773		

$$
\begin{aligned}
H_0 &: E(u_i^2) = \sigma^2 \\
H_1 &: E(u_i^2) = \frac{\sigma^2}{ng_i} \\
\textit{Aux. regress} &: \hat{u}_i^2 = \hat{\delta}_0 + \hat{\delta}_1 \times \frac{1}{ng_i} + \hat{v}_i,\ R^2_{\hat{u}} = 0.1582 \\
&\quad n \times R^2_{\hat{u}} \overset{a}{\sim} \chi^2_1 \text{ under } H_0
\end{aligned}
$$

We reject the null because the value of the $n \times R^2_{\hat{u}}$ *here is* $37 \times 1582 = 5.85$*, which is larger than* 3.84*, the 5% critical value of the* χ^2_1 *distribution. Please caution the students that this is a large sample test, and a sample of 37 observations may not be sufficient, and we only take that as corroborating evidence for our theoretical reasoning in part (a).*

d. Suppose we reject homoskedasticity in favor of Var $(u_i) = \frac{\sigma^2}{ng_i}$. Describe how we should transform the variables and which regression we should run to obtain best linear unbiased estimator for

β_0, β_1 and β_2.

$$\sqrt{ng_i} \times food_i = \beta_0\sqrt{ng_i} + \beta_1 \sqrt{ng_i} \times inc_i + \beta_2 \sqrt{ng_i} \times nf_i + \sqrt{ng_i} \times u_i \Longrightarrow Var(u_i) = \sigma^2$$

regress $\sqrt{ng_i} \times food_i$ on $\sqrt{ng_i}, \sqrt{ng_i} \times inc_i, \sqrt{ng_i} \times nf_i$ with no constant

e. Based on the weighted least squares estimates test the hypothesis that, other things staying the same, a 1 dollar increase in family income will result in 20 cents increase in family food consumption, against the alternative that it will result in a less than 20 cents increase in food consumption. Perform this test at the 5% level of significance.

Dependent Variable: @SQRT(NG)*FOOD
Method: Least Squares
Included observations: 37

Variable	Coefficient	Std. Error	t-Statistic	Prob.
@SQRT(NG)	70.31890	28.51754	2.465813	0.0189
@SQRT(NG)*INC	0.177184	0.010268	17.25529	0.0000
@SQRT(NG)*NF	57.70058	8.056305	7.162164	0.0000

$$\begin{aligned} H_0 &: \beta_1 = 0.20 \\ H_1 &: \beta_1 < 0.20 \\ t &= \frac{\hat{\beta}_1 - \beta_1}{se\left(\hat{\beta}_1\right)} \sim t_{34} \text{ under } H_0 \\ \text{reject if } t_{calc} &< -1.6909 \\ t_{calc} &= \frac{0.1772 - 0.20}{0.0103} = -2.2136 \end{aligned}$$

We reject the null hypothesis and conclude that data suggests that keeping the number of family members constant, households spend less than 20 cents per dollar on food. You can explain that in this context, it would make sense to consider also a log-log form (log(food) on log(inc) and nf), however, there may not be time to explore it in the tutorial.

Note that you can perform the Breusch-Pagan test by either running the auxiliary regression yourself and calculating the test statistic using the output of the auxiliary regression, or you can use the eviews built in commands. If you need help with this, please refer to the eviews screen shots in the lecture slides. Similarly, you can run the weighted least squares by multiplying each variable by the appropriate weight and running an OLS regression with these weighted variables, or you can use eviews' weighted least squares option as shown in the lecture.