Introductory Econometrics: Computer Problem Sheet 2

****I will not provide detailed Excel spreadsheet answers to this question, but rather just a few comments (in italics) after each question. Note that parts of this question are very similar to the example given in the textbook in Chapter5 in the Heteroskedasticity section and, hence, precise numerical answers for some parts are given there******

You can download the data from course website.

Question 1

This data set has been collected from the UNESCO yearbook and it contains data from 1997 for 38 countries on their educational spending, GDP and population. In particular, it contains three variables:

EDUC = public recurrent expenditure on education (millions of US dollars)
GDP = gross domestic product (millions of US dollars)
POP = population (millions)

1. *Testing for heteroskedasticity*. Sort the data by GDP and perform a Goldfeld-Quandt test, running regressions using the subsamples of fourteen countries with the smallest and largest values for GDP. Repeat this exercise sorting the data by POP.

Note that I am asking you to do two heteroskedasticity tests. One is to see whether there is heteroskedasticity associated with GDP, the other with POP. For each, you have to sort the data appropriately and then run 2 regressions (one with the first 14 countries, one with the last 14) and calculate the GQ stat. It turns out that there is heteroskedasticity associated with GDP (but not POP).

2. *The White test*. Perform White's test for heteroskedasticity on a regression of EDUC on GDP and POP. Do the results of this test agree with the results of the Goldfeld-Quandt test?

   Interestingly, the White test indicates heteroskedasticity is not present. This is what you can sometimes get in econometrics: one test indicates one thing, but another indicates the opposite. There is nothing you can do in such cases other than to note that they do happen and to go with one test or the other. The conservative strategy here is to assume heteroskedasticity does exist so I will assume it does for the rest of this exercise sheet.

   It is also worth noting that you only need to do the White test once (in part 1, you had to do 2 GQ tests). However, the White test offers no guidance as to the form of the heteroskedasticity.

3. *Correcting for heteroskedasticity: weighted least squares*. Based on your results from question 1, investigate whether scaling by POP or GDP corrects the heteroskedasticity problem. That is, run a regression of EDUC/POP on

GDP/POP and test for heteroskedasticity using the White test. Run a regression of EDUC/GDP on 1/GDP and test for heteroskedasticity using the White test. Does either of these transformations solve the heteroskedasticity problem. How do you interpret the coefficients (including the intercept) in these transformed regressions?

Remember that, if heteroskedasticity is present, you should try and use the GLS estimator if possible. But the GLS estimator is just an OLS regression using an appropriately transformed model (by "appropriately transformed" I mean a regression which satisfies the classical assumptions – something you must verify after doing the transformation). The textbook example using this data has details about this. As discussed there, it does seem that the transformation "divide all variables by GDP" fixes the heteroskedasticity problem. The interpretation of coefficients is described near the top of page 137 of the textbook.

4. *Correcting for heteroskedasticity by logging.* Investigate whether the heteroskedasticity problem can be corrected by logging the variables. That is, use a White test investigate whether the a heteroskedasticity problem exists in a regression of ln(EDUC) on ln(GDP) and ln(POP). How do you interpret the coefficients in this regression?

Given that the transformation in question 3 solves the problem, in a real empirical exercise, there would be no need to answer this question. That is, once you have found a transformation that solves the heteroskedasticity problem there is no need to seek for another one. But I put this question on just to emphasize that there are many popular transformations that can be experimented with to solve a heteroskedasticity problem. It is often the case that logging some or all of your variables will solve a heteroskedasticity problem (although note that, when you are working with logged variables, the interpretation of coefficients is different)

Question 2

In Computer Problem Sheet 1 OLS methods were used in a regression analysis involving the house price data set. Was the use of OLS appropriate? Hint: Use a White test to check if heteroskedasticity is present in the regression you ran in Computer Problem Sheet 1.

If you do a White test involving the regression of part i) of Computer Problem Sheet 1, a White stat of 61.95 is obtained. The critical value associated with the 5% level of significance is taken from the Chi-squared(11) distribution (since there are 11 explanatory variables in this regression) and is 19.675. Therefore, the null hypothesis of homoskedasticity is rejected and heteroskedasticity does seem to be present. Hence, the use of OLS in Computer Problem Sheet 1 was not appropriate (the theory tells us that OLS estimates are unbiased but confidence intervals and hypothesis tests will be incorrect).

If you are feeling ambitious you might want to see if you can fix up this heteroskedasticity problem. If I log the dependent variable (but do not log the *explanatory variables) and then do a White test on the resulting regression, I find a White stat of 20.36. This (barely) is greater than the 19.675 critical value so this does not solve the problem. Perhaps logging some of the explanatory variables as well will solve the problem? I leave you to experiment (but remember you cannot log dummy variables since the log of zero is not defined, so you can only log explanatory variables like lot size, number of bedrooms, etc. that only take on positive values).*