

## Answer Keys To EC5040 Computer Exercise

1) Be careful in your summary of the data. The variable “edage” is not the years of education variable (this is “yearsed”)

There are some observations on years of education that are coded as -99. This is wrong. (Typically “minus” codes are used to indicate missing values). Should drop these.

N.B. There are some individuals in the data set coded with yearsed=0 as “never had an education” and coded as 97 in the variable edage. However the yearsed variable is correctly coded as zero, so there is no need to drop these observations.

Omitting those with zero years of education is a mistake. These are valid observations and will influence the estimated coefficient.

Also there are missing values for much of the sample in the quarter of birth variables. Again it is not necessary to drop them. We can assume the data is missing at random.

No marks were lost however if either of these two criteria were used to drop individuals

Do need to comment on the difference between the restricted sample and unrestricted sample estimates

2) In a semi-log (log-lin) model the coefficients on discrete variables are only approximations to the idea of a semi-elasticity that is the interpretation on a continuous right hand side variable. Hence the correction factor  $\exp(b)-1$ .

3) When testing a null hypothesis it is important to state what the null hypothesis being tested is.

4) Heteroskedasticity biases the OLS standard errors and hence linear hypothesis tests. To fix up the standard errors to make them consistent (if not efficient) to heteroskedasticity of unknown form use the “robust” correction

Should comment on the difference to the standard errors and t statistics use of the “robust” command makes.

5) The (Durbin-Wu) Hausman test for endogeneity is based on a comparison of the IV and OLS estimates. Under the null hypothesis of no endogeneity both OLS and IV will give consistent estimates of the true coefficient values, but OLS will be the most efficient. If endogeneity is present then only IV is consistent so would expect a big difference in the coefficient estimates

The asymptotic equivalent version of the test (Wu-Hausman) is to regress the endogenous variable on the **full set** of exogenous variables ie including the original exogenous right hand side variables, save the residuals and include these as an additional term in the original model. An insignificant t value suggests that the residuals are uncorrelated with the endogenous variables (see lecture notes)

6) The IV estimates using different instruments are very different

In particular the IV estimate using the school leaving age 15 dummy is estimated with a very wide confidence interval – even though it is significantly different from zero – and the point estimate 1.14 looks odd (we normally expect the returns to education to be in the range 0.05-0.15)

Even though the sla15 dummy passes the 1<sup>st</sup> stage F test with a value >10 (the “rule of thumb” threshold for a good instrument) it is still not enough to prevent a strange result occurring

The reason for the poor performance of this instrument lies in its high correlation with the variable age (over 0.8) The resulting multicollinearity makes it hard to detect an independent effect.

Moral: be careful in your choice of instruments. Sometimes the tests and thresholds used to help you are not good enough

7) Using both instruments together should improve the asymptotic efficiency of the estimate. In this case there is not much improvement. Make sure you know exactly what the null of the overidentifying test is and what the asymptotic equivalent form of the test is (see lecture notes and model answer). Since the reason for the poor performance of this instrument lies in its high correlation with the variable age (over 0.8) The resulting multicollinearity makes it hard to detect an independent effect. This also has implications for the IV estimate in the over-identified case which is in effect a weighted average of the effect of the 2 instruments.

Conclude should probably not use sla15 as an instrument

8) Make sure you can calculate the F test of linear restrictions correctly and calculate the statistic without relying on the “test” command in stata

9) Be careful with your presentation . Three decimal points when reporting or commenting on estimation results is more than enough. Spurious accuracy otherwise

10) Make sure you know the formula for the test of linear restrictions (see lecture notes and computer exercises)

11) make sure you know the difference between the regression based Wu-Hausman test and the Hausman test of endogeneity

12) The F value reported in the first stage of the IV regression using the command “ivreg” is **NOT** the correct F test since it is the test of goodness of fit of ALL the right hand side variables. The correct F test – of the predictive power of the instrument of other variables – is only given in the “ivreg2” command (as the question specified).

13) Make sure you do the rest of the analysis with the same sample having dropped the suspect observations

14) Make sure you can do the Breusch-Pagan test for heteroskedasticity manually without relying on the stata command

15) Make sure you can interpret a semi-elasticity estimate of a continuous variable correctly