Problem set 4: Weighted Least Squares and Instrumental Variables Estimators

September 18, 2013

1 Introduction

This problem set accompanies the Youtube lecture series, and roughly corresponds to videos 125-155, covering Weighted Least Squares and Instrumental Variables analysis.

2 Returns to education - practical

2.1 Weighted Least Squares

1. In this section we are going to investigate how education, (and various other factors), affects future wages. The dataset we are going to use is prebuilt into Gretl, and comes from Jeffrey Wooldridge's great textbook[1]. To get hold of the Wooldrige, 'Introductory Econometrics'. In order to load the dataset, navigate to File → Open data → Sample file... Then select the Wooldridge tab (if you can see it - if you cannot I will explain how to access this in a minute) and open the 'card' dataset. If you cannot see the Wooldridge tab you may need to go to:

   `http://gretl.sourceforge.net/gretl_data.html#jw`

   and download the wooldridge_data.exe or wooldridge.tar.gz file (I recommend the former method because it handles the installation for you). Alternatively, you can load in the sample dataset from clicking on the 'Look on server' button at the top of the sample datasets box. Alternatively, I have provided a link to download a .xls file containing the dataset on: (however note that this data is missing the descriptive labels of the native Wooldridge Gretl file)

[1] The actual data source is D.Card (1995), "Using Geographic Variation in College Proximity to Estimate the Return to Schooling," in Aspects of Labour Market Behavior: Essays in Honour of John Vanderkamp. Ed. L.N. Christophides, E.K. Grant, and R. Swidinsky, 201-222. Toronto: University of Toronto Press. You can find the paper here `ttp://davidcard.berkeley.edu/papers/geo_var_schooling.pdf - I recommend it; it makes for a good read.`

`http://www.oxbridge-tutor.co.uk/#!datasets/culy`

This data contains hourly wages, years of schooling, IQ test scores, age, father and mother education level, as well as host of other potential explanatory factors for a sample of 3010 individuals in 1976 (although some factors are from 1966).

(a) Let's start with the most important part of any econometric analysis: graphing the data. Which have the highest correlation with *wage*?

(b) Build an ordinary least squares model of the form:

$$wage_i = \alpha + \beta_1 educ_i + \varepsilon_i$$

What is the interpretation of the constant in this model?

(c) What does the R-squared imply about the explanatory power of education in this model?

(d) Test for the presence of heteroscedasticity using an appropriate test. What does this show?

(e) Can you graphically display your results to the previous test? What are the problems posed by this?

(f) Take the log of *wage*, and rerun the regression with this as a dependent variable. What is the interpretation of the coefficient on *educ* now?

(g) Compare the fit of the non-logged and logged models. Which specification do you prefer?

(h) Test for the presence of heteroscedasticity in the logged model. Has logging had the desired effect? Why/why not?

(i) One solution to heteroscedasticity is to use robust standard errors which take it into account. If you re-estimate your model for non-logged *wage* using robust standard errors, what changes?

(j) Assuming we have a well-specified model (which likely isn't the case here), another way of dealing with heteroscedasticity is to estimate a weighted least squares model. The problem however is that we do not know the specific weights to adequately re-weight our model. All is not lost however, since we are able to estimate them from OLS residuals, and then use these to weight our variables in feasible Generalised Least Squares. Luckily, Gretl is set up to conduct the entirety of the procedure, so we don't have to worry about estimating the covariance matrix ourselves. We are going to estimate this type of model using (non-logged) *wage* as a dependent variable. In order to estimate an fGLS model, type 'hsk depvar const indepdentvar1 indepdentvar2....' into to console. Alternatively, if you are using the GUI mode go to 'Model' → 'Other linear models' → 'Heteroskedasticity corrected...' You can then select the dependent variables in the same way as you did for estimating OLS models. How do the coefficient estimates compare to those estimated for OLS?

(k) How do the standard errors compare? On this basis why should we prefer fGLS estimates over OLS?

2.2 Instrumental variables estimation

(l) Do you think that the coefficient on *educ* in the log *wage* reflects the return to one more year of schooling? Why/why not?

(m) One way to get round the issue of endogenous regressors is to use an appropriate instrument. Would *enroll* (whether an individual was enrolled in school in 1976) be an appropriate instrument? Why/why not?

(n) Evaluate whether the distance an individual lives from college is likely a reasonable instrument to use for education.

(o) There are two measures of whether an individual lived close to a college in 1966: *nearc2* and *nearc4*, for whether an individual lived near to a two year and/or four year college. If we are going to choose a single instrument for *educ*, which of these variables should we use and why?

(p) We are going to re-estimate our model with the log of wage as a dependent variable, with *educ* as the independent variable, using a single appropriate IV. First of all we are going to do this manually, then using the built in Gretl option. The first step in estimating a model manually via instrumental variables is a first step regression of the endogenous variable on the relevant instrument. Carry out an appropriate first stage regression, and present its results. Is the instrument sufficiently strong?

(q) Now use your first stage regression to get predicted values of the endogenous variable by selecting 'Save' → 'Fitted values' from your model window. Now use these fitted values in place of the endogenous variable in the regression. How do your estimates of the return to education compare with those obtained via OLS/fGLS?

(r) Now replicate your results using the 'Model' → 'Instrumental variables' → 'Two Stage Least Squares...' option in Gretl. Check that your results are the same.

(s) What does the Hausman test statistic suggest about OLS estimates?

(t) How do the standard errors compare with those of the equivalent OLS model?

(u) Since both indicator variables are correlated individually with education, we could include them as instruments for *educ*. What happens if you re-estimate the model using both of these?

(v) Do you prefer your answer obtained using *nearc2* and *nearc4* as instruments, or just using a single instrument?

2.3 Freestyle - estimate the return to education!

(w) I thought it would be fun to see what different approaches to this data yield in terms of the return to education. See if you can produce a model which you are happy with, and compare it with my efforts in the *answers* PDF.

3 Theory

2. A researcher is interested in estimating the effect of civil war on economic growth. For a sample of 100 countries he estimates the following regression.

$$\widehat{\text{growth}} = \underset{(0.23)}{2.12} - \underset{(0.21)}{1.35}\,\text{civil} - \underset{(0.45)}{3.01}\,\text{landlocked}$$

$$N = 100 \quad R^2 = 0.09 \quad F = 4.84$$

(standard errors in parentheses)

Where *growth* is the country's 2012 real gdp growth, *civil* is a dummy variable taking on a value 1 if the country has experienced a civil war in the past ten years, and *landlocked* is a dummy variable equal to 1 if the country is landlocked.

(a) What is wrong with the estimation of the effect of civil war by the above specification?

(b) A student suggests using *landlocked* as an instrument for *civil* since it is likely correlated with this variable. Evaluate this methodology.

(c) It is suggested that a dummy variable *mountains*, taking on the value '1' if a country has mountainous terrain over 5,000 feet, may be a good instrument for *civil*. Evaluate its attractiveness in this role.

(d) Assuming that the researcher opts to use *mountains* as an instrument for *civil*, outline the first stage regression.

(e) 'Since civil is a dummy variable, the researcher should really use a logit or probit model to estimate it.' Evaluate this statement.

(f) Another variable *ethnic*, representing the ethnic fractionalisation of the country is found to be strongly correlated with the residual from the first stage regression above. What does this suggest for an improved estimation technique?

(g) The Hausman chi-statistic from the 2SLS regression of the form indicated in the last part is found to be 56.74. What does this suggest about the results from OLS regression as to the effect of civil war on growth?

3. A model for the relationship between y and x is found to be the following:

$$y = \alpha + \beta x + \varepsilon$$

Where $\mathbb{E}[\varepsilon|x] = 0$ and $Var(\varepsilon|x) = log(x)\sigma^2$.

(a) What properties does the ordinary least squares estimator for β have?

(b) Suggest a better estimator. (*Hint*: what weights would it apply to the original regression equation?)

(c) Prove that this estimator is BLUE.