 Assignment 4: GR Model, Panel Data, IV Estimation

**Professor William Greene**
Phone: 212.998.0876
Office: KMC 7-90
Home page: people.stern.nyu.edu/wgreene
Email: wgreene@stern.nyu.edu
Course web page: people.stern.nyu.edu/wgreene/Econometrics/Econometrics.htm

**Part I. Econometric Theory**

1. A researcher reports a set of ordinary least squares to a colleague. The colleague remarks that the data are obviously drawn from a generalized regression model, and that they should be using generalized least squares, not ordinary least squares. The researcher returns to the lab and computes a two step generalized least squares estimator using one of the methods we have discussed in class. They are horrified to find that the GLS estimated standard errors are larger, not smaller than the OLS estimated standard errors. They conclude that something must be wrong with the computer program used to do the computation. Do you agree? Explain.

2. Suppose that a heteroscedastic regression model applies;

\[ y_i = x' \beta + \epsilon_i, \ E[\epsilon_i] = 0, \ \text{Var}[\epsilon_i] = \sigma^2 w_i \]

where \( w_i \) is an observed variable - that is, except for the scale factor, the disturbance variances are known. Suppose, however, that instead of using \( w_i \), you compute weighted least squares using another variable, \( z_i \) to form the weights. What is the true covariance matrix of this estimator? (Hint: See page 279 of your text – Equation (9-32).) Is the estimator computed using the “wrong” weights consistent? Is it unbiased? Explain your answers

3. Show analytically that 2SLS (two stage least squares) is an instrumental variable estimator by the definition we used in class.

4. Show analytically why the 2SLS estimator cannot be computed if there are fewer instrumental variables than there are endogenous variables in the model. Interpret this condition.
Part II. Applications: Generalized Regression

1. Using the 14 following observations on y and x,

\[
\begin{array}{cccccccccccccc}
  y & 0 & 0 & 0 & 1 & 2 & 0 & 0 & 4 & 2 & 0 & 1 & 1 & 0 & 0 \\
  x & 41 & 35 & 24 & 54 & 34 & 45 & 43 & 35 & 36 & 22 & 33 & 25 & 26 & 46 \\
\end{array}
\]

compute the ordinary least squares regression of y on x (without a constant) for the regression model

\[ y = \beta x + \epsilon \]

and report your results. Now, compute the generalized least squares estimator of \( \beta \) assuming that \( \text{Var}[\epsilon|x] = \theta_x \), and report your results. Which estimator appears to be more efficient? Does the finding actually give evidence about efficiency? Explain.

2. This and the next exercise are based on the GSOEP health care data, healthcare.csv. We will base the regressions on the model

\[ \text{hhinc} = \beta_1 + \beta_2 \text{age} + \beta_3 \text{age}^2 + \beta_4 \text{educ} + \beta_5 \text{married} + \beta_6 \text{female} + \beta_7 \text{hhkids} + \epsilon \]

We are interested in a (possibly) groupwise heteroscedastic model in which the variance of \( \epsilon \) depends on whether or not one has public health insurance. (Public is a dummy variable that equals 1 for those who have the insurance and 0 for those who do not.) One way to model this simple specification would be

\[ \text{Var}[\epsilon|\text{public}] = \sigma^2 \exp(\gamma \text{public}) \]

where married is the dummy variable for marital status. Thus, the two variances are \( \text{var}[\text{not public}] = \sigma^2 \) and \( \text{var}[\text{public}] = \sigma^2 \tau \) where \( \tau = \exp(\gamma) \). Thus, the disturbance is heteroscedastic if \( \gamma \) is not equal to zero.

a. Compute the least squares regression based on the full sample and report your results.

b. Compute separate regressions for married=0 and married=1. Report the values of s, the standard deviations of the residuals, from the two regressions. What do you conclude?

c. The experiment in b. is not quite consistent with the model, since it uses different regression coefficients for the two standard deviations. A more appropriate calculation would be to compute the residuals from the original pooled regression, then examine the two sets of residuals separately to see if the variances appear to be the same. What do you find?

d. Now, let’s approach this statistically. Consistent with the model for the variance of \( \epsilon \), compute the squares of the residuals from the pooled regression, then compute the logs of the squares. (Note, this is equivalent to twice the log of the absolute value, but not twice the log of the residuals, some of which are negative.) Finally, linearly regress the logs of the squared residuals on a constant and the Public dummy variable. What do you find? What is the conclusion. Are the disturbances heteroscedastic or not?

e. Weighted least squares: Your regression in part d. produces an estimate of \( \gamma \). So, you can now compute an estimate of the variance of \( \epsilon \) as \( \exp(c^*\text{public}) \) where \( c \) is the estimate. With this variance, you can compute a weighted least squares estimate of the coefficient vector. Use this two step approach to compute the weighted least squares estimates. Does the weighting change things much? Compare the weighted to the ordinary least squares estimates.

3. Hypothesis test in the generalized regression. Since the disturbance in part 2 is heteroscedastic, the F statistic is no longer valid for testing linear restrictions. The general approach is to use a Wald statistic. I am interested in the joint hypothesis that \( \beta_5 = \beta_6 = \beta_7 = 0 \). Show how to carry out the Wald test. Using the weighted least squares estimation results in part e, carry out the test.
Part III. Applications: Instrumental Variables

This exercise is based on the Cornwell and Rupert Data. I am interested in the model

\[ \text{logwage} = \beta_1 + \beta_2 \text{fem} + \beta_3 \text{ed} + \beta_4 \text{occ} + \beta_5 \text{ind} + \beta_6 \text{exp} + \beta_7 \text{blk} + \beta_8 \text{union} + \text{eps} \]

I am concerned that the experience variable, \( \text{exp} \), might be endogenous, so as an alternative to OLS, I propose using 2SLS in which my instruments for \( \text{exp} \) are (wks, south, smsa). Compute the OLS and 2SLS estimators, and report the relevant results. I am also interested in testing for the endogeneity of \( \text{exp} \). I find the Hausman test to be complicated, so I propose to use Wu’s variable addition test, instead. Thus, I will regress \( \text{exp} \) on all of the exogenous variables in the model and compute the residuals. I will then recompute the OLS regression, but add these residuals to the model. A statistically significant coefficient on the residual suggests that \( \text{exp} \) is, indeed, endogenous. Carry out the test.

Part IV. Application: Panel Data

We continue the analysis of the previous model. Assume for now that \( \text{exp} \) is exogenous, so we are not concerned with simultaneous equations or two stage least squares

1. The C&R data are a balanced panel of 595 observations observed over 7 years. Thus, in estimation, I might be concerned about ignoring individual heterogeneity. We begin by attempting to compute an estimator for the covariance matrix of the least squares estimator that is ‘robust’ to the misspecification. What assumptions are needed to justify this computation? Specifically, does it work for a fixed effects model? Compute the estimator. Compare the results to the conventional that does not account for the ‘clustering.’

2. Show algebraically how the fixed effects estimator is computed. Note that the equation above contains three variables that do not have any time variation, FEM, ED, BLK. What is the implication of the presence of these variables in the equation for estimation of the fixed effects model. Estimate the logwage equation using a fixed effects estimator.

3. How do the random and fixed effects estimators differ in (1) their assumptions and (2) in the computations done to compute them. Compute a random effects estimator of the logwage equation and compare your results to the fixed effects estimator.

4. Is there a way to decide which is the preferred model, fixed or random effects? If your statistical results contain any evidence, use them to make the decision. Justify your choice.
import; file="H:\Courses-Stern\Econometrics\Problems\DataFiles\healthcare.csv"
sample; all
create; agesq = age*age
namelist; x = one, age, agesq, educ, married, female, hhkids $  
regress; lhs = hhninc; rhs = x; res = e $ Pooled regression and keep residuals
regress; for [public = 0, 1]; lhs = hhninc; rhs = x $ Separate regressions
create; esq = e*esq
regress; lhs = logesq; rhs = one, public $ Looking for groupwise heteroscedasticity
? A little experiment to show the heteroscedasticity effect.
draw; n=5000
plot; lhs = public; rhs = e; endpoints = -1, 2 $  
? Analysis of variance to see if the standard deviations are different.
sample; all
dstat; rhs = e; str = public $  
? Weighted least squares
regress; lhs = logesq; rhs = one, public
}
calc; gma = b(2) $
create; vare = exp(b(2)*public); wt = 1/vare $
regress; lhs = hhninc; rhs = x $
regress; lhs = hhninc; rhs = x; wts = wt $
regress; lhs = hhninc; rhs = x; hetero$
regress; lhs = hhninc; rhs = x; res = e $
create; esq = e*esq
matrix; white = <x'x> * x'
regress; lhs = hhninc; rhs = x $
regress; lhs = hhninc; rhs = x; wts = wt
regress; rhs = x; wts = wt $
\text{Using built in procedure}
\text{regress; lhs = hhninc; rhs = x; wts = wt} \text{; test: married = 0, female = 0, hhkids = 0}$
import; file="H:\Courses-Stern\Econometrics\Problems\DataFiles\cornwell-rupert.csv"

namelist; x = one, fem, ed, occ, ind, exp, blk, union $
namelist; z = one, fem, ed, occ, ind, blk, union, wks, south, smsa $
regress; lhs = lwage; rhs = x $
2sls; lhs = lwage; rhs = x; inst = z $
regress; lhs = lwage; rhs = x; res = ze $
regress; lhs = lwage; rhs = x; ze$
create; i = trn(7, 0)$
setpanel; group = i; pds = ni $
? Simple least squares
regress; lhs = lwage; rhs = x $
regress; lhs = lwage; rhs = x; cluster = 7 $
regress; lhs = lwage; rhs = x; panel; fixed $
regress; lhs = lwage; rhs = one, occ, ind, exp, union; panel; fixed $
regress; lhs = lwage; rhs = one, occ, ind, exp, union; panel; random $
regress; lhs = lwage; rhs = one, occ, ind, exp, union; panel $
regress; lhs = lwage; rhs = one, occ, ind, exp, union; panel