Final Exam
Applied Econometrics
Prof. Leo Feler

Fall 2011

This exam is worth 100 points. It is worth 50% of your total grade in the class. You have 3 hours to complete this exam.

This exam is closed books and closed notes. You may use calculators. You must sign and adhere to the honor code below. Please write directly on this exam.

Good luck!

Honor Code

I, ________________________________, certify that all work on this exam is my own work. I have not consulted with others nor referenced any notes or books, nor have I engaged in any activities that could be construed as cheating. I have not received from anyone nor will I share with anyone information about the contents of this exam. I understand that some students may be taking this same exam at a later time, and by disseminating any information about this exam, I may be biasing their outcome. I will therefore not discuss or distribute the contents of this exam with or to anyone. I also understand that this exam is not officially proctored. This is because I and my classmates are trustworthy people: I will not cheat on this exam, and if I do observe or have knowledge of anyone cheating, I will report it to the professor, who will take appropriate action. I understand the maximum penalty for being found guilty of honor code violations by the Honor Code Board is expulsion from SAIS.

____
Signature

1. OLS and Standard Errors [20 points]. The estimating equation is $y_i = \beta S_i + \varepsilon_i$.

   a. Show that $\hat{\beta}_{OLS}$ minimizes the sum of squared residuals.

   b. What is the intuition for an estimate that minimizes the sum of squared residuals?

c. What are the assumptions for $\hat{\beta}_{OLS}$ to be unbiased? Why do we care about bias?

d. What are the assumptions for $\hat{\beta}_{OLS}$ to be efficient? Why do we care about efficiency?

e. Given our estimating equation, if S_i is years of schooling and y_i is $\ln(wage_i)$, why might $\hat{\beta}_{OLS}$ be biased? Give an example (and show the calculation) for how $\hat{\beta}_{OLS}$ might *overestimate* the true β. Give an example (and show the calculation) for how $\hat{\beta}_{OLS}$ might *underestimate* the true β.

f. What are two reasons why we might incorrectly estimate the variance of $\hat{\beta}_{OLS}$? How do we correct for these in Stata (what are the commands), and what is Stata doing when you insert these commands (i.e., how is Stata estimating the variance of $\hat{\beta}_{OLS}$)? Why do we care about the possibility of underestimating the true variance, and so how do we choose which standard errors to report?

2. Omitted Variables, Measurement Error, and Panel Data [25 points]. Suppose you have panel data on working-age individuals' schooling and wages.
   a. What is panel data? How does it differ from a cross section?

   b. Given that you have a panel, what can you do to minimize the possibility that an individual's characteristics, both observed and unobserved, jointly determine both wages and schooling? What assumption do you have to make about these characteristics? What is the drawback of your solution to controlling for both observed and unobserved individual characteristics?

c. When you're estimating the returns to schooling controlling for these observed and unobserved individual characteristics using your solution in part (b), what are you estimating β off of? Let me help you in answering this question: when you estimate from only a cross section of individuals, how do you obtain your estimate of returns to schooling, β [i.e., off of what kind of variation is Stata estimating β]? Now, with panel data and given your solution in part (b), how do you obtain your estimate of returns to schooling, β [off of what kind of variation]?

d. For your panel of working-age individuals and with your solution from (b), do you expect much variation in schooling? Do you expect this variation to be random? How might this bias your results?

e. If schooling is measured with error, and you apply your solution from (b), what might happen to your estimate of returns to schooling? Why? Relate this to your answer from parts (c) and (d).

f. We have discussed two instruments that try to address omitted variable bias in measuring the returns to schooling: quarter of birth and distance to a college immediately prior to being of college age (in this case, before working age). Can you use these instruments with your panel and your solution from (b)? Why or why not?

g. For any estimation you do with this panel, what should you do to your standard errors? Why?

3. Instrumental Variables [25 points]. Let's go back to a cross section. The estimating equation is $y_i = \beta S_i + \gamma X_i + \varepsilon_i$. You have two instruments for schooling S_i, the quarter of birth (call this Z_1) and the distance to a college immediately prior to being of college age (call this Z_2).
   a. What conditions must your instruments satisfy in order to be "good"? What do these conditions mean? How do you determine that these conditions are satisfied (if it's even possible to do)?

b. You instrument for S_i using both Z_1 and Z_2. What statistics do you look at to see if your conditions from part (a) are satisfied or at least not violated? How does Stata calculate these statistics?

c. Here's some output from an IV procedure. You don't know what these variables are, and it doesn't matter. Is the IV procedure legit? Can you determine if it is or not? Why or why not?

```
Summary results for first-stage regressions
-------------------------------------------

                                           (Underid)            (Weak id)
Variable     | F(  1,   812)  P-val | AP Chi-sq(  1) P-val | AP F(  1,   812)
ShareTransfe |     179.57    0.0000 |       180.67   0.0000 |      179.57

NB: first-stage test statistics heteroskedasticity-robust

Stock-Yogo weak ID test critical values for single endogenous regressor:
                                   10% maximal IV size             16.38
                                   15% maximal IV size              8.96
                                   20% maximal IV size              6.66
                                   25% maximal IV size              5.53
Source: Stock-Yogo (2005).  Reproduced by permission.
NB: Critical values are for Cragg-Donald F statistic and i.i.d. errors.

Underidentification test
Ho: matrix of reduced form coefficients has rank=K1-1 (underidentified)
Ha: matrix has rank=K1 (identified)
Kleibergen-Paap rk LM statistic          Chi-sq(1)=74.88    P-val=0.0000

Weak identification test
Ho: equation is weakly identified
Cragg-Donald Wald F statistic                                     414.58
Kleibergen-Paap Wald rk F statistic                               179.57

Stock-Yogo weak ID test critical values for K1=1 and L1=1:
                                   10% maximal IV size             16.38
                                   15% maximal IV size              8.96
                                   20% maximal IV size              6.66
                                   25% maximal IV size              5.53
Source: Stock-Yogo (2005).  Reproduced by permission.
NB: Critical values are for Cragg-Donald F statistic and i.i.d. errors.
```

```
Number of observations                  N  =        817
Number of regressors                    K  =          5
Number of endogenous regressors         K1 =          1
Number of instruments                   L  =          5
Number of excluded instruments          L1 =          1

IV (2SLS) estimation
--------------------

Estimates efficient for homoskedasticity only
Statistics robust to heteroskedasticity

                                                      Number of obs =       817
                                                      F(  4,    812) =     40.41
                                                      Prob > F      =    0.0000
Total (centered) SS     =   44.0061878                Centered R2   =    0.1454
Total (uncentered) SS   =  100.7510703                Uncentered R2 =    0.6267
Residual SS             =  37.60558497                Root MSE      =     .2145

------------------------------------------------------------------------------
             |               Robust
dltotinc0604 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
Sha~iDir2006 |   1.107051   .6540936     1.69   0.091    -.1749493     2.38905
dmedyrs~0604 |   .0430033   .0086306     4.98   0.000     .0260875     .059919
  dlnpop0604 |   .6918022   .0738875     9.36   0.000     .5469853    .8366191
 dshurban0604 |  -.4185882   .2854034    -1.47   0.142    -.9779685    .1407921
       _cons |   .2050826   .0140577    14.59   0.000     .1775301    .2326352
------------------------------------------------------------------------------
Underidentification test (Kleibergen-Paap rk LM statistic):             74.885
                                                   Chi-sq(1) P-val =    0.0000
------------------------------------------------------------------------------
Weak identification test (Cragg-Donald Wald F statistic):              414.579
                         (Kleibergen-Paap rk Wald F statistic):        179.568
Stock-Yogo weak ID test critical values: 10% maximal IV size             16.38
                                         15% maximal IV size              8.96
                                         20% maximal IV size              6.66
                                         25% maximal IV size              5.53
Source: Stock-Yogo (2005).  Reproduced by permission.
NB: Critical values are for Cragg-Donald F statistic and i.i.d. errors.
------------------------------------------------------------------------------
Hansen J statistic (overidentification test of all instruments):         0.000
                                                 (equation exactly identified)
------------------------------------------------------------------------------
Instrumented:         ShareTransferMuniDir2006
Included instruments: dmedyrsofschooling0604 dlnpop0604 dshurban0604
Excluded instruments: GenAgShockInst0103
------------------------------------------------------------------------------

.
```

d. You decide to generate a second instrument, called "randomcrap", which is just a random number distributed $N(0,1)$, and you include it in your IV procedure. You get the following output. Is the IV procedure legit? Why or why not? What can you deduce, from the Hansen J statistic, about your original instrument (i.e., not the randomcrap one)? Why?

```
Summary results for first-stage regressions
-------------------------------------------

                                           (Underid)            (Weak id)
Variable     | F(  2,   811)  P-val | AP Chi-sq(  2) P-val | AP F(  2,    811)
ShareTransfe |      92.05    0.0000 |      185.47   0.0000 |        92.05

NB: first-stage test statistics heteroskedasticity-robust

Stock-Yogo weak ID test critical values for single endogenous regressor:
                                   10% maximal IV size             19.93
                                   15% maximal IV size             11.59
                                   20% maximal IV size              8.75
                                   25% maximal IV size              7.25
Source: Stock-Yogo (2005).  Reproduced by permission.
NB: Critical values are for Cragg-Donald F statistic and i.i.d. errors.

Underidentification test
Ho: matrix of reduced form coefficients has rank=K1-1 (underidentified)
Ha: matrix has rank=K1 (identified)
Kleibergen-Paap rk LM statistic          Chi-sq(2)=75.80    P-val=0.0000

Weak identification test
Ho: equation is weakly identified
Cragg-Donald Wald F statistic                                     207.31
Kleibergen-Paap Wald rk F statistic                                92.05

Stock-Yogo weak ID test critical values for K1=1 and L1=2:
                                   10% maximal IV size             19.93
                                   15% maximal IV size             11.59
                                   20% maximal IV size              8.75
                                   25% maximal IV size              7.25
Source: Stock-Yogo (2005).  Reproduced by permission.
NB: Critical values are for Cragg-Donald F statistic and i.i.d. errors.

Weak-instrument-robust inference
Tests of joint significance of endogenous regressors B1 in main equation
Ho: B1=0 and orthogonality conditions are valid
Anderson-Rubin Wald test           F(2,811)=       1.55     P-val=0.2124
Anderson-Rubin Wald test           Chi-sq(2)=      3.13     P-val=0.2094
Stock-Wright LM S statistic        Chi-sq(2)=      3.04     P-val=0.2192

NB: Underidentification, weak identification and weak-identification-robust
    test statistics heteroskedasticity-robust
```

```
Number of observations               N  =        817
Number of regressors                 K  =          5
Number of endogenous regressors      K1 =          1
Number of instruments                L  =          6
Number of excluded instruments       L1 =          2

IV (2SLS) estimation
--------------------

Estimates efficient for homoskedasticity only
Statistics robust to heteroskedasticity

                                                      Number of obs =      817
                                                      F(  4,    812) =    40.43
                                                      Prob > F      =   0.0000
Total (centered) SS     =   44.0061878                Centered R2   =   0.1455
Total (uncentered) SS   =  100.7510703                Uncentered R2 =   0.6268
Residual SS             =  37.60388326                Root MSE      =    .2145

------------------------------------------------------------------------------
             |               Robust
dltotinc0604 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
Sha~iDir2006 |    1.10536   .6539053     1.69   0.091    -.1762713    2.386991
dmedyrs~0604 |   .0430017   .0086279     4.98   0.000     .0260914     .059912
  dlnpop0604 |   .6917985   .0738871     9.36   0.000     .5469824    .8366145
 dshurban0604 |  -.418558   .2854328    -1.47   0.143    -.9779959      .14088
       _cons |   .2051166   .0140292    14.62   0.000       .17762    .2326133
------------------------------------------------------------------------------
Underidentification test (Kleibergen-Paap rk LM statistic):             75.797
                                                   Chi-sq(2) P-val =    0.0000
------------------------------------------------------------------------------
Weak identification test (Cragg-Donald Wald F statistic):              207.312
                         (Kleibergen-Paap rk Wald F statistic):         92.055
Stock-Yogo weak ID test critical values: 10% maximal IV size             19.93
                                         15% maximal IV size             11.59
                                         20% maximal IV size              8.75
                                         25% maximal IV size              7.25
Source: Stock-Yogo (2005).  Reproduced by permission.
NB: Critical values are for Cragg-Donald F statistic and i.i.d. errors.
------------------------------------------------------------------------------
Hansen J statistic (overidentification test of all instruments):         0.011
                                                   Chi-sq(1) P-val =    0.9179
------------------------------------------------------------------------------
Instrumented:         ShareTransferMuniDir2006
Included instruments: dmedyrsofschooling0604 dlnpop0604 dshurban0604
Excluded instruments: GenAgShockInst0103 randomcrap
------------------------------------------------------------------------------
```

e. Suppose the IV procedure above, where the instruments are GenAgShockInst0103 and randomcrap, is perfectly legit, regardless of whether this is actually true. The dependent variable is the change in the natural log of total income in a municipality between 2004 and 2006. The independent variable of interest is the share of the municipality's income in 2006 that comes from federal government conditional cash-transfers. In 2004, this share was zero. How do you interpret the coefficient on the independent variable of interest [Sha~iDir2006]? If the share of a municipality's income in 2006 is 0.2 (so 20%), by how much does total income increase between 2004 and 2006?

f. In order for the increase you just found in part (e) to be causal, what assumptions do you have to make if you were estimating this in OLS?

g. Here's the OLS results of the estimations in parts (c) and (d). Why is the coefficient estimate on the independent variable of interest [Sha~iDir2006] so different than in the IV procedure? What does this suggest about the relationship between omitted variables and the dependent variable: the change in the natural log of total income in a municipality between 2004 and 2006, i.e., income growth in a municipality? How does instrumenting correct for this?

```
. reg dltotinc0604 ShareTransferMuniDir2006 dmedyrsofschooling0604 dlnpop0604 dshurban0604, robust

Linear regression                                      Number of obs =     817
                                                       F(  4,   812) =   39.16
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.1570
                                                       Root MSE      =  .21375
```

dltotinc0604	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
Sha~iDir2006	.1002205	.2952896	0.34	0.734	-.4794004	.6798415
dmedyrs~0604	.0420833	.008659	4.86	0.000	.0250867	.0590799
dlnpop0604	.6895863	.0735645	9.37	0.000	.5451874	.8339852
dshurban0604	-.4006078	.2812711	-1.42	0.155	-.9527121	.1514964
_cons	.225331	.0110086	20.47	0.000	.2037223	.2469397

4. <u>Freebies: Regression Discontinuity [10 points]</u>. These next questions are pretty easy. They're free points, basically, and a repeat of what you've seen before.
    a. What is regression discontinuity? When can you use it? Why do you use it?

b. In “Do Better Schools Matter”, Sandra Black uses a spatial regression discontinuity design to estimate willingness-to-pay for schools with better test scores. She is estimating willingness-to-pay based on housing price differences near borders (see figure). What are the assumptions that allow her to deduce that housing price differences are due to differences in school quality?

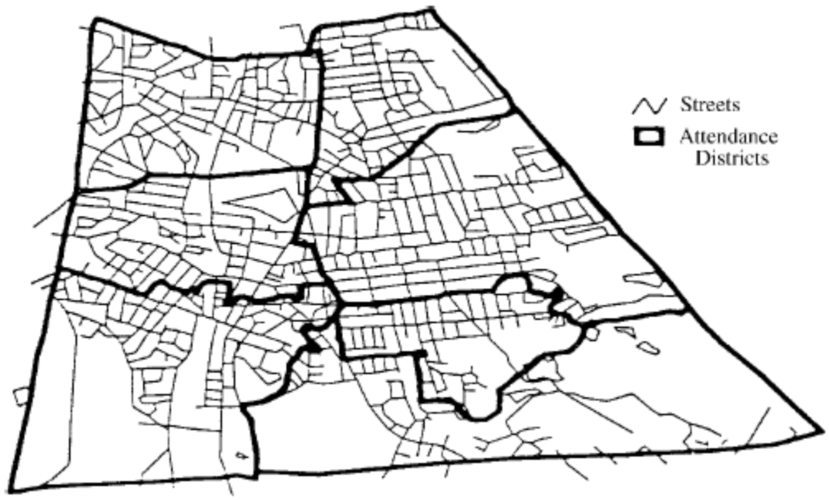

FIGURE I
Example of Data Collection for One City: Melrose Streets, and Attendance District Boundaries

c. Sandra Black's paper was heavily criticized. Perhaps your assumptions in part (b) were true when the attendance district boundaries were first introduced. But over time, sorting takes place. Those people who really value good schools for their children might move to another side of the boundary (i.e., assume these are somehow "better" people). Discuss how this would bias Sandra Black's results. Instead of estimating just willingness-to-pay for schooling, what might differences in housing prices now be capturing *in addition* to willingness-to-pay for schooling?

5. Freebies: Propensity Score Matching [10 points]. These next questions are again pretty easy.
   a. What is propensity score matching? When can you use it? Why do you use it?

b. What's the "algorithm" for estimating the propensity score?

c. What are weaknesses of the propensity score method?

6. Panel Data and Differences in Differences [10 points].

   a. You have the following empirical specification: $y_i = \alpha + \beta_1 Treat + \beta_2 Post + \beta_3 TreatXPost + \varepsilon_i$ where $Treat$ is a dummy equal to 1 for the treatment group, $Post$ is a dummy equal to 1 for the post-period, and $TreatXPost$ is an interaction of $Treat$ and $Post$. Describe what the coefficient estimates for α, β_1, β_2, and β_3 capture.

b. Now rewrite this empirical specification to include a fixed effect for each individual. What drops out and why?

c. If you use a random effect instead of a fixed effect, what assumptions are you making about how individual characteristics are correlated with y_i? What is the benefit of using random effects instead of fixed effects? How might your estimates be affected depending on whether your assumptions about the appropriateness of random effects are right or wrong?

[END OF EXAM. HAVE A GOOD BREAK.]