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. I think you'll be time-constrained, but answer as much as you can.

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! Use the force!

Honor Code

I, Answer Key, 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 results. 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 and upstanding 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 that the maximum penalty for being found guilty of honor code violations by the Honor Code Board is expulsion from SAIS.

_____
Signature

2 pts

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.

$Y = \beta X + \varepsilon$

$\varepsilon = Y - \beta X$

$$\min (\varepsilon)^2 = \varepsilon'\varepsilon$$

$$\varepsilon'\varepsilon = (Y - X\beta)'(Y - X\beta)$$

$$\varepsilon'\varepsilon = (Y' - \beta'X')(Y - X\beta)$$

$$\varepsilon'\varepsilon = Y'Y - Y'X\beta - \beta'X'Y + \beta'X'X\beta$$

$$\varepsilon'\varepsilon = Y'Y - 2Y'X\beta + \beta'X'X\beta$$

$$\frac{\partial}{\partial \beta} = 0 - 2X'Y + 2X'X\beta$$

$$0 = -2X'Y + 2X'X\beta$$

$$X'Y = X'X\beta$$

$$(X'X)^{-1}X'Y = (X'X)^{-1}X'X\beta$$

$$\hat{\beta} = (X'X)^{-1}X'Y$$

2 pts

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

We minimize the sum of squared residuals in order to produce a best fit line. When our OLS assumptions hold, minimizing $\varepsilon'\varepsilon$ gives us the best linear unbiased estimator since we are minimizing the variance. We square the residuals in order to account for the positive and negative values of ε_i.

2 pts

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

1.) $E(\varepsilon_i | x_{ik}) = 0$ for all k: the residuals are independent of the observables

→ 0 conditional mean assumption

2.) X has full column rank: no x_i is a linear combination of another x_i.

We care because we want our estimate of β to be centered around the true β.

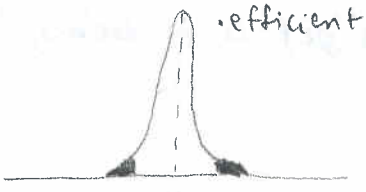
* note: the $\varepsilon_i \sim iid(0, \sigma^2)$ is the assumption about efficiency.

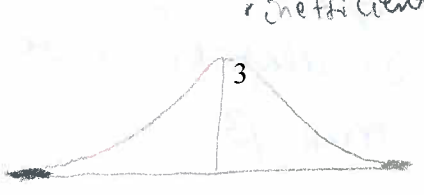
2 pts

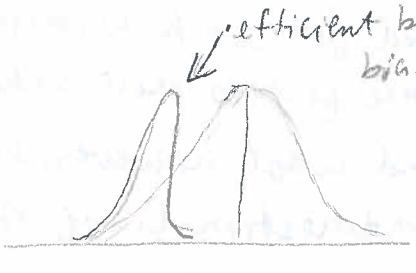
d. What are the assumptions for $\hat{\beta}_{OLS}$ to be efficient? Why do we care about efficiency?

1.) $\varepsilon_i \sim iid(0, \sigma^2)$: the expected value of the residuals is 0 and the residuals have constant variance.

We care about efficiency because with it, we'll get a more accurate estimate of the true β. With higher efficiency, (as $N \to \infty$ and $\varepsilon_i \sim iid(0, \sigma^2)$), we are able to get an estimate of β that is more tightly centered around the true β (assuming unbiasedness as well). If these assumptions are not met, we'll have a large variance and therefore large SE's, and will therefore have difficulty in achieving significance (rejecting H_0 in favor of H_A; i.e. finding our estimate of β is significantly different than 0). We should use the estimates from the higher variance, such as in the case of HIV testing.







	$Cov(S_i, A_i) > 0$	$Cov(S_i, A_i) < 0$
$\theta \oplus$	over	under
$\theta \ominus$	under	over

6 pts

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.

$\hat{\beta}_{OLS}$ may be biased because we have one explanatory variable with no controls: there could be other omitted variables that are also driving wages that we are not controlling for. So, we'll attribute too much "credit" to schooling (if there is a $\oplus$ correlation between the omitted variable and schooling). Our omitted variables are embedded in the error term in our original equation.

If $Corr(S, A) > 0$ and $\theta > 0 \Rightarrow \oplus$ bias
If $Corr(S, A) < 0$ and $\theta < 0 \Rightarrow \oplus$ bias
If $Corr(S, A) < 0$ and $\theta > 0 \Rightarrow \ominus$ bias
If $Corr(S, A) > 0$ and $\theta < 0 \Rightarrow \ominus$ bias

Estimated equation:

$$W = \alpha + \tilde{\beta}_1 S + \varepsilon$$

True model:

$$W = \gamma + \beta_1 S + \theta A + \eta$$

← ability; causes our estimate of $\hat{\beta}_1$ in original equation to be biased

$$E(\tilde{\beta}_1) = \beta_1 + \theta \frac{(Cov\, S, A)}{var\, S}$$

estimated $\tilde{\beta}_1$ | true form | bias

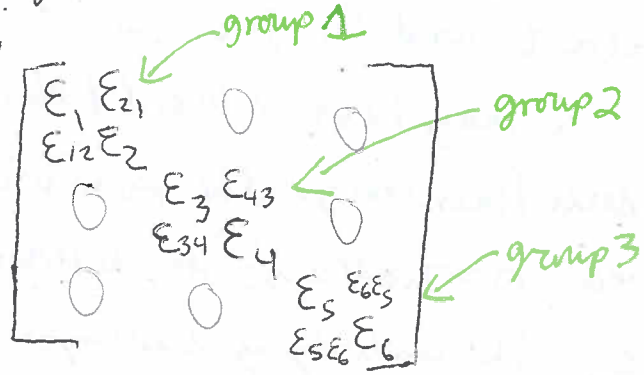
$\oplus$ Bias: If $\theta \overset{\oplus}{>} 0$ and $Cov(S, A) \overset{\oplus}{>} 0$; ability is likely to be positively related to schooling and log wage. We therefore have a $\oplus$ bias on our $\tilde{\beta}_1$ schooling variable and have overestimated its true effect.

$\ominus$ Bias: If $\theta \overset{\oplus}{>} 0$ and $Cov(\text{Schooling}, \text{work experience}) \overset{\ominus}{<} 0$: our estimation of the effect of schooling and wages could be an underestimation if we haven't controlled for work experience. Assuming that a person w/ greater work experience forgoes more schooling, schooling and WE are $\ominus$ correlated. WE and wage, however, are $\oplus$ related, so we'll get a $\ominus$ bias, or underestimate, of the true β.

6 pts

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?

- Two reasons that we incorrectly estimate the variance of $\hat{\beta}_{OLS}$ are:
  1.) heteroskedasticity: $\text{var}\,\sigma_i^2 \neq \sigma_j^2$ - our ability to predict our Y-value varies across our X-values.
  2.) When there is clustering of the errors - our samples may not be i.i.d. and partially correlated with each other within groups. If the correlation of $\varepsilon_i, \varepsilon_j > 0$, then we underestimate our variance; if $\text{Corr}\,\varepsilon_i, \varepsilon_j < 0$, we will overestimate the variance of $\hat{\beta}_{OLS}$ if we incorrectly assume i.i.d. This mostly happens when you have group effects, such as twins, families, classrooms.
- We can correct for heteroskedasticity by putting , robust at the end of the regression and to correct for clustering, we put , cluster at the end. , robust corrects heteroskedasticity by summing the variances along the variance-covariance matrix instead of taking a mean σ^2: $\text{Var}(\hat{\beta}_{OLS}) = (X'X)^{-1}X'\Sigma X(X'X)^{-1}$, where the diagonal of Σ is $[\varepsilon_1^2 \varepsilon_2^2 \ldots \varepsilon_n^2]$. For , cluster, we can cluster on groups:



note: If we don't have > 42 groups to cluster on, then bootstrap, which calculates the variance of β_{OLS} 400 times and takes the standard deviation instead of the standard error

- We don't want to underestimate the true variance because when we do, our standard errors are too small, leading us to incorrectly reject the null when we should in fact fail to. We could, for example, reject the null that a blood sample is HIV ⊕ and falsely conclude that it is actually HIV ⊖. We always choose the higher (more conservative) standard errors.

1 pt

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?

Panel data - many observations over time on many individuals

Cross-section - one observation for each individual (one time period)

4 pts

   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?

TO control for an individual's observed and unobserved characteristics that jointly determine both wages & schooling, we can use individual fixed effects, or first differencing. This means that there is no fundamental change in characteristics between t_1 and t_2 for individual i. With individual FE, we may have reduced / eliminated OVB because we reduce / minimize the possibility that observed or unobserved covariates jointly determine both our X and Y variables. The underlying assumption is that these characteristics are time-invariant.

Drawbacks

1.) We can't see the effects of certain characteristics that are time invariant

2.) exacerbate attenuation bias if measurement error is present because we'll have reduced signal without reducing noise

5 pts

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]?

- In a cross section of individuals, we estimate β off of the variation across individuals measured at a single point in time.
- When we control for individual fixed effects in panel data, we estimate β off of the variation within an individual over time.

* note that if there is no variation in schooling over time for individuals, we cannot estimate β with fixed effects.

5 pts

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?

- There should not be much variation in schooling because levels of education in the working age population does not change much over time because they are done going to school.
- This variation won't be random because of selection: for example, people who go back for more schooling are showing increased motivation, which will bias our results because motivation will be an OV in our error term. People may also return to school in response to some wage shock.

5 pts

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).

If schooling is measured with error, controlling for FE could exacerbate attenuation bias by reducing signal more than noise ("throwing the baby out with the bathwater").

The formula for the bias in the estimated returns to schooling is

$$\lambda = \frac{\sigma_u^2}{\sigma_{s^*}^2 + \sigma_u^2}$$

$$E(\hat{\gamma}_{OLS}) = \left(1 - \frac{\lambda}{1 - R^2_{S,X}}\right)\gamma$$

where: S^* is the true measure of schooling

u is the error in measurement of schooling

$R^2_{S,X}$ is the R^2 from regressing S_i on X_i →

measurement error causes attenuation bias towards 0 for $\hat{\gamma}_{OLS}$ since $\lambda > 0$ and $0 \leq R^2_{S,X} \leq 0$

2 pts

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?

No, quarter of birth would drop out since it doesn't change over time, as would distance to college.

3 pts

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

- We should cluster the SE at the individual level to account for any errors that are correlated within each individual over time. Clustering at the individual level will lead to an increase in our SE, assuming the errors are positively correlated over time.
- We could also cluster on time, depending on which clustering variable yields the more conservative (larger) SE.
- Note you can only cluster if there are more than 42 individuals.

7 pts [parts a and b]

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)?

The two conditions are:

① strong - the instruments, z_1 and z_2, are strongly correlated with the endogenous variable (schooling, S_i) : $Cov(S_i, z_i) \neq 0$

② valid - z_1 or z_2 must influence Y_i only through the channel of S_i : $Cov(\varepsilon_i, z_i) = 0$ $\quad z \rightarrow X \rightarrow Y$ (NO!)

To determine if they are satisfied:

- From the first stage regression, $S_i = \alpha + \pi_1 z_{i1} + \pi_2 z_{i2} + \gamma X_i + \varepsilon_i$

test H_0: $\pi_{10} = \pi_{11} = 0$
H_1: $\pi_{10} \neq$ or $\pi_{11} \neq 0$

and see if the F-stat of your instruments is ≥ 10 to test for strength. Since this equation is overidentified (more instruments than endogenous variables), the Hansen-Sargan test is used to check for validity. The hypotheses are:

H_0: instruments are not invalid
H_A: instruments are invalid
} and we need a p-value $> .1$ [but prefer a p-value $\geq .60$]

to fail to reject the null (i.e. a low χ^2 value), which is what we are secretly hoping for.

In this test, we predict $\hat{\varepsilon}_i$ from our second stage equation

$$Y = \alpha + \beta_1 \hat{S}_i + \beta_2 X_i + \varepsilon_i$$

$$\hookrightarrow \hat{S}_i = \underset{\sim}{\alpha} + \underset{\sim}{\beta_1} z_{i1} + \underset{\sim}{\beta_2} z_{i2} + \beta_3 X_i + \mu_i$$

and regress $\hat{\varepsilon}_i$ on z_{i1}, z_{i2}, and X_i: $\hat{\varepsilon}_i = \tilde{\alpha} + \tilde{\delta}_1 z_{i1} + \tilde{\delta}_2 z_{i2} + \tilde{\gamma} X_i + \mu_i$

and then calculate $NR^2 \sim \chi^2(z_{ik} - S_{ik})$ DOF

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?

[see part a]

4 pts

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.
```

• We Cannot say immediately whether the IV procedure is legit. On one hand, there is a very high F-stat in the first stage of 179.57, which suggests that the instrumental variable is strong. However we cannot use the Hansen-Sargan test to test for validity because the model is just/exactly-identified. To use the H-S test, we need more IV's than endogenous variables.

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

dltotinc0604	Coef.	Robust 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

4 pts

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
```

• Not legit! The F-stat is strong (>10), but we cannot determine anything from Hansen Sargan. The assumption for H-S is that if one instrument is valid, can you test whether the other instrument is valid? "randomcrap" has no predictive power, so it cannot be used to calculate uncorrelated $\hat{\varepsilon}_i$'s in 2SLS. So we can't run

$$\hat{\varepsilon}_i = \alpha + \delta_1 Z_1 + \delta_2 Z_2 + \delta_3 X_i + \eta_i$$

required for the Hansen-Sargan test. By construction, Z_2 truly is "random crap", so R^2 should be pretty low anyway. We are actually no better off than we were in the previous question.

```
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
------------------------------------------------------------------------------
```

3 pts

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?

- A 1% increase in the share of the municipality's income in 2006 that comes from federal government conditional cash transfers is associated with a 1.105% increase in total income.
- So, a 20% increase in municipal income from government programs =

$$.2 \times 1.10536 = .221072$$

a 22.11% increase in municipal income.

3 pts

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?

We have to assume no omitted variable bias or reverse causality. A change in transfers is exogenous: No omitted variable is causing both incomes and transfers to increase. That is, nothing is causing a city to receive more transfers but is associated with slower growth (i.e. if only poor cities get transfers).

4 pts

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

The difference between the IV estimate and OLS estimate is 1.10536 - .112205 = 1.005. This means that something in our error term that we are not controlling for is biasing our OLS estimate of the effect of cash transfers on growth downwards.

Bc we have that $\hat{\gamma}$ = Sha~iDir2006

$$\hat{\gamma} = \gamma + \theta \frac{Cov(S,A)}{var(A)}$$

either $Cov(S,A) > 0, \theta < 0$ or $Cov(S,A) < 0, \theta > 0$

Bc we have a negative income shock, for example, this omitted variable will be positively correlated with federal transfers and negatively correlated with income.

By instrumenting with ag shock 03, we can predict the volume of cash transfers and thus the effect of cash transfers on income growth.

4 pts

4. Freebies: Regression Discontinuity [10 points]. 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?

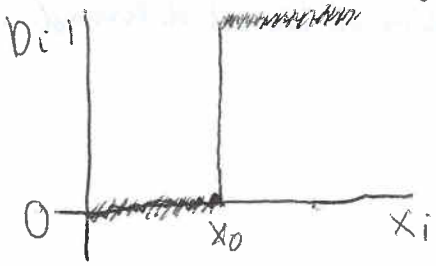
Regression discontinuity design is a technique that takes advantage of an arbitrary rule that determines assignment to treatment to estimate the effect of receiving that treatment. We use it to approximate random experiments as closely as possible. We can, for example, look within a narrow band around the discontinuity to see the effect of treatment on the treated group vs. the control group because we take advantage of the similarity in characteristics of those on each side of the discontinuity. In this way, we can control for OVB - it is just by random chance that 1 person/group was assigned to treatment and the other person/group was not.

We can thus use regression discontinuity when treatment is a deterministic and discontinuous function of a covariate, X_i.

$$D_i = \begin{cases} 1 & \text{if } X_i \geq X_0 \\ 0 & \text{if } X_i < X_0 \end{cases} \quad \text{where } X_0 = \text{Cutoff value}$$

"deterministic": whether you receive treatment completely depends on your value of variable X_i.

"discontinuous": no matter how close you are to X_0, you only get treated when you actually reach X_0



Sharp, not fuzzy discontinuity

3 pts

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?

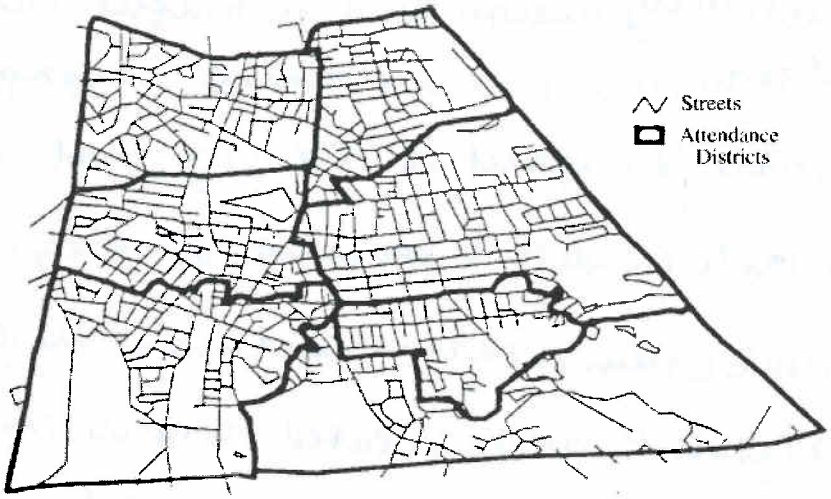


FIGURE I
Example of Data Collection for One City: Melrose Streets, and Attendance District Boundaries

Black assumes that houses within a narrow band around the discontinuity are very similar (same size yard, # of rooms, crime rate) and that the only difference in price is the amount that 1 house versus the other would pay for the better school (MWTP).

For this to be true, the boundaries that separate the two schools had to have been drawn arbitrarily: the city government did not just divide along already established "better" or "worse" neighborhoods. She also had to assume that there was no "funziness" - that those who were eligible for treatment received it and those who were ineligible did not receive it.

Finally, she would have to be willing to extrapolate her estimate of the local average treatment effect (LATE) around the discontinuity to the rest of the sample. There could be a variation of observable and unobservable characteristics beyond the narrow band around the discontinuity.

3 pts

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?

As time goes on, the externalities that result from having a good school in your district begin to undermine our assumption that the two houses/neighborhoods within the narrow band of the discontinuity are equal in all other X's other than assignment to treatment (quality of school). As time goes on and the "better" people move into the neighborhood, they could be more educated and thus more likely to have ↑ incomes, etc. and price out the "normal" people. Then, the neighborhood can become a safer, more desirable place to live with the new "better" group of people than with the original group that lived there when the boundary was first drawn. Thus, separating out people's marginal willingness to pay for better schools will be difficult to parcel out.

4 pts

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?

PSM is a procedure used to construct an index of the likelihood of receiving treatment based on a series of individual X_{iK}. You can use it when treatment is binary and when the X_{iK} are balanced across the sample so that we can match and then compare people based on their p-score. In this way, we compare individuals with the same propensity to smoke, for example, but by chance, 1 person starts smoking and the other does not (a way to look at a counterfactual reality) to estimate the effect of smoking (on birthweight). We can also focus on different ranges of p-scores (more in the middle [ATE] or more at the high end [TOT] to obtain different estimates of receiving treatment.

We use it to replicate a randomized experiment when it is not possible or unethical to do so. We also use it to solve the multidimensionality problem, giving us more efficiency. Instead of controlling for 1,000 characteristics, other X_i and still possibly suffering from OVB, we control for 1 variable, $P(X_i)$, saving us degrees of freedom. Also, there may be selection bias: we need to be comparing "apples with apples." Finally, we also do not need to worry about misspecification of correct functional form.

3 pts

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

1.) start w/ parsimonious logit → estimate $\hat{p}(X_i)$

2.) stratify data into 5 blocks of $\hat{p}(X_i)$

3.) test $\bar{X}_i = \bar{X}_o$ for all K within each block using t-test of significant differences in sample means

a.) if X_K's balance in each block, then STOP

b.) if X_K's are not balanced in some blocks, divide block into 2 blocks and reevaluate

c.) if X_K's are not balanced in <u>all</u> blocks, add interaction and/or polynomial of X_K's to logit and reevaluate

* repeat until you balance X_K's in treatment and control groups in each block. The big stopping rule is when you fail to reject $\bar{X}_{1K} = \bar{X}_{0K}$ for 90% of t-tests in over 90% of blocks (90-90 rule)

3 pts

c. What are weaknesses of the propensity score method?

1.) You can only use it if treatment is binary

2.) Your P-score can vary greatly based on how well you balance your blocks.

3.) You have to make sure that your T_i, X_i are not 100% collinear. This means that there is a high degree of similarity between treatment and control individuals across blocks on all covariates (X_{ik})

4.) It is hard to replicate the same results, making the process seem arbitrary.

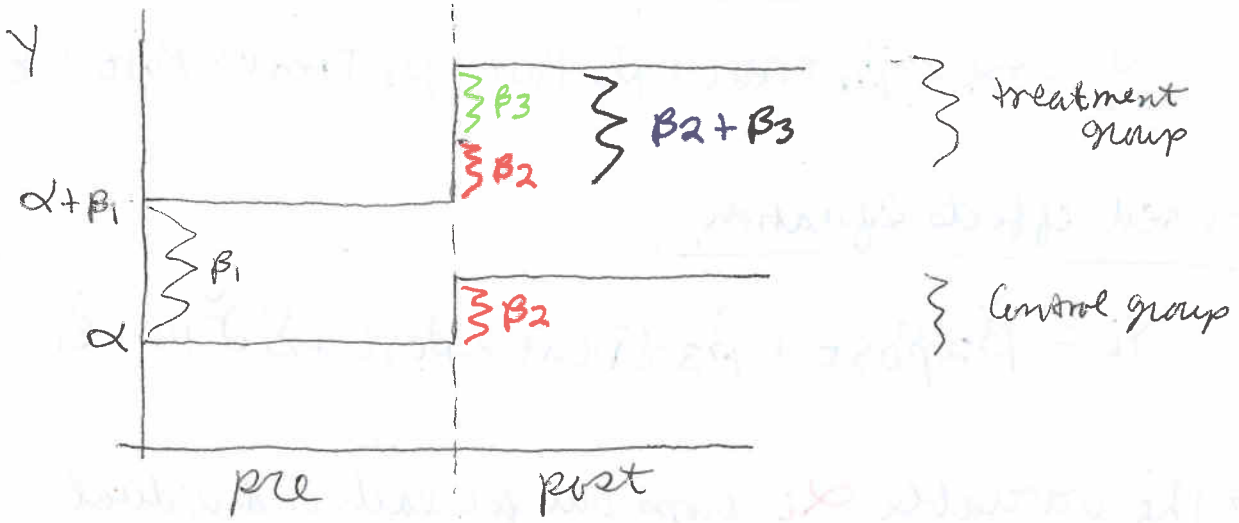
5.) Your different estimates of the effect of treatment can be overly sensitive to outliers and noise

6.) You may not get the sufficient overlap of the box plots you were hoping for, which means you can't proceed with the procedure

4 pts

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.



- α Captures the estimated Y_i for the control group in the pre-period.
- β_1 Captures the additional effect of being treated in the pre-period.
  → $\beta_1 + \alpha$ Captures the initial level of the treatment group
- β_2 Captures the effect of time across all individuals - the effect of being in the post-period.
- β_3 is the interaction of being treated and being observed in the post-period. β_3, therefore Captures the true causal effect of treatment.

3 pts

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

Original equation:

$$Y_i = \alpha + \beta_1 Treat + \beta_2 Post + \beta_3 Treat \times Post + \varepsilon_i$$

Fixed effects equation:

$$Y_i = \tilde{\beta}_2 Post + \tilde{\beta}_3 Treat \times Post + \sum \tilde{\gamma} p_i + \varepsilon_i$$

- The variable α_i drops out for each individual because the variable was constant for all individuals (treatment and control) through time (pre- and post-period): α is normalized to 0.
- Additionally, $\beta_1 Treat$ drops out because being treated was fixed for those individuals across time: An individual who was treated in the pre-period will have been treated in the post-period.

3 pts

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?

- With random effects, you assume that individual characteristics are not correlated with the Y_i. For example, if you are weird, we don't assume that being weird is correlated with your wage. With fixed effects, however, we assume that individual characteristics are correlated with Y_i and therefore need to be differenced out in order to produce a biased estimate.
- One benefit of doing RE is that we can see how characteristics such as education level, sex, and race (fixed) affect Y_i. If we were to use fixed effects, these variables would drop out because they do not change over time.
- Another benefit of doing random effects is that it won't exacerbate attenuation bias like fixed effects could if there is measurement error

$$Y_{it} = \alpha + \beta X_{it} + \rho D_{it} + \underbrace{\eta_{it}}_{\alpha_i + \varepsilon_{it}}$$

- If Corr$(\eta, D_{it}) \to 0$, then RE will look like FE
- If Corr$(\eta, D_{it}) \to 1$, then RE will look like OLS.

[END OF EXAM. HAVE A GOOD BREAK.]