Applied Econometrics
Prof. Leo Feler
Quiz 7: Heckman Selection and Panel Data

Name: Key

1. Interpret the following output. The variable of interest is an indicator if a locality only has state-owned bank branches (DState). The dependent variable is lnloansfin (i.e., the natural log of all loans and financings made by banks in a locality). "imr" is the Inverse Mills Ratio. The other variables are just controls (so don't worry about them right now). The sample is localities with branches from only one type of bank (state-owned or private, but not both). Is there evidence of non-random selection into the sample? How do you know? Can we generalize the results from the sample to the overall population of localities, which includes localities with multiple types of bank branches? How do you interpret the coefficient on "DState"? How do you interpret the coefficient on "imr" in the output from the Heckman correction procedure: if the estimate were significant, what does it say about the types of localities that are in the sample?

This output shows results from a simple OLS regression:

```
Linear regression, absorbing indicators          Number of obs   =      57562
                                                 F(  9,   1077)  =       8.81
                                                 Prob > F        =     0.0000
                                                 R-squared       =     0.4149
                                                 Adj R-squared   =     0.3978
                                                 Root MSE        =     2.8658

                                 (Std. Err. adjusted for 1078 clusters in amc)
```

lnloansfin	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
DState	-.7861826	.2551891	-3.08	0.002	-1.286907	-.2854584
totbranches	1.478336	.3017359	4.90	0.000	.8862795	2.070393
lngmp	.5679066	.1624466	3.50	0.000	.2491588	.8866543
lntotpop	.1356452	.2187602	0.62	0.535	-.2935994	.5648898
shurban	1.701022	.7086902	2.40	0.017	.3104521	3.091592
shworking	-.4025278	1.551301	-0.26	0.795	-3.446443	2.641388
shworkershs	-.6578736	1.529667	-0.43	0.667	-3.659339	2.343592
shareag	1.884316	1.131294	1.67	0.096	-.3354735	4.104106
sharecommserv	.5958899	1.731609	0.34	0.731	-2.801819	3.993599
_cons	1.70966	2.402444	0.71	0.477	-3.004341	6.423661
statemonthy~r	absorbed					(1625 categories)

The following output is based on a Heckman correction procedure:

```
Linear regression, absorbing indicators          Number of obs  =     57562
                                                 F(  10,   1077) =      8.07
                                                 Prob > F       =    0.0000
                                                 R-squared      =    0.4152
                                                 Adj R-squared  =    0.3981
                                                 Root MSE       =    2.8650

                              (Std. Err. adjusted for 1078 clusters in amc)
```

lnloansfin	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
DState	-.7774262	.2551303	-3.05	0.002	-1.278035	-.2768175
totbranches	1.478935	.3011374	4.91	0.000	.8880523	2.069817
lngmp	.7115758	.2371965	3.00	0.003	.2461561	1.176995
lntotpop	.5512495	.5036442	1.09	0.274	-.4369856	1.539485
shurban	1.85191	.7323192	2.53	0.012	.4149754	3.288844
shworking	-.2015757	1.576764	-0.13	0.898	-3.295452	2.892301
shworkershs	-.4155295	1.598682	-0.26	0.795	-3.552413	2.721354
shareag	1.967182	1.140406	1.72	0.085	-.2704865	4.204851
sharecommserv	1.231435	1.867466	0.66	0.510	-2.432848	4.895719
imr	-.7568001	.7603886	-1.00	0.320	-2.248811	.7352108
_cons	-3.226677	6.290926	-0.51	0.608	-15.57054	9.117184
statemonthy~r	absorbed					(1625 categories)

The coefficient estimate for the Inverse Mills Ratio (IMR) is -.757 but not statistically significant. The coefficient on DState changes from -.786 to -.777 when controlling for selection. So no; no evidence of non-random selection.

Yes, we can generalize to the population, especially after we control for potential sample selection. After controlling (via the Inverse Mills Ratio), we have rescaled results from sample to population.

DState is a dummy variable. Localities with state branches have approximately

78% less lending than localities with private bank branches.

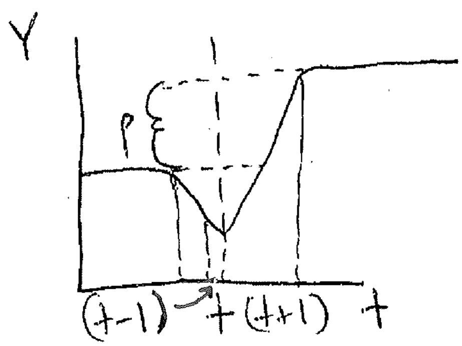
The coefficient on imr is -.76. Localities in the sample, given their characteristics, should have approximately 76% less lending than localities not in the sample. In other words, localities with a lower volume of lending are more likely to be in the sample.

2. Suppose schools are assigned to a program where they receive more resources based on having bad average test scores in a previous period. You are trying to figure out the effect of the program on test scores. You have information on test scores, treatment status, and socio-economic characteristics at the school-level (i.e., school averages, not information on individual students) for many periods before and after the program is put in place. Once a school is in the program, it remains in the program until the end of your sample. Write down the equation you would like to estimate. What are potential sources of bias that might affect your estimates of the program's effects? How would you obtain an upper and lower bound of the program's effects (and explain intuitively why these might be upper and lower bounds)? What kind of standard errors should you report (i.e., what do you need to do to ensure you don't underestimate your standard errors)?

$$Y_{it} = \alpha + \underbrace{\theta Y_{it-1}}_{\text{lagged scores}} + \underbrace{\beta X_{it}}_{\text{socio-economic controls}} + \underbrace{\rho D_{it}}_{\text{indicator for whether in program}} + \underbrace{\varepsilon_{it}}_{\text{error term}}$$

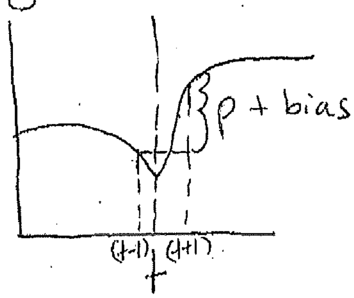
There might be fixed omitted variables that cause schools to have lower scores (or smaller improvements) and also cause it to be in the treatment group ($D_{it}=1$).

Estimate using a LDV (lagged dependent variable) specification:



Worse performance in t-1 is correlated with participation in treatment.

In any case, we would expect to see some mean reversion after poor performance if $0 < \theta < 1$. If we did not control for lagged test scores (or if we controlled for school fixed effects), then we would mistakenly ascribe all the change between $(t-1)$ and $(t+1)$ to the program, when in reality, some of this improvement would have occurred anyway simply due to mean reversion:



if we use LDV, this might underestimate the program effect (lower bound), but if we use school fixed effects, we might overestimate the program effect (upper bound).

Why might LDV underestimate the program effect?
if $Cov(A_i, D_{it}) < 0$ (i.e. $A_i \downarrow \rightarrow D_i \uparrow$), and $Cov(A_i, Y_{it}) > 0$

$$\rho = \rho_{true} + \overset{(+)}{\gamma} \left[\frac{\overset{(-)}{Cov(D_{it}, A_i)}}{Var(D_{it})} \right]$$

In this case, you would be underestimating the true ρ, i.e. schools with good principals are less likely to be in the program (good schools) and are more likely to have higher test scores.

Here, the omitted A_i is the unobserved quality of school principals.

You would want to report the clustered standard errors at the school level since observations of different periods within schools are not likely to be iid (they should be correlated with each other).

3. Suppose I have a program that offers deworming drugs to school students. Schools are randomly selected to receive deworming drugs. You have health information (weight-for-height, anemia, wormload, etc.) for both treatment and control school students for several years. Note that you have data at the student-level but the treatment is offered at the school-level. Write down and explain a difference-in-difference specification to measure the effect of deworming on health. Now suppose that treatment is only offered to males and not females within a school. Write down a triple difference specification to measure the effect of deworming on males versus females in treatment versus control schools before and after the deworming program is implemented. For the difference-in-difference and the triple difference specifications, what kind of standard errors should you report (i.e., what do you need to do to ensure you don't underestimate your standard errors)?

$$y_{it} = \alpha + \beta_1 treat_{it} + \beta_2 post_t + \beta_3 treat_i \times post_t + \varepsilon_{it}$$

β_1 captures how treated school students differ from control students in the pre-period.

β_2 captures how control school students experience changes between pre and post period (even though they are not treated).

β_3 is the difference-in-difference estimator. How treated school students differ from control school students after being compared to pre-treatment results is captured by this coefficient:

$\beta_3 = (T_1 - C_1) - (T_0 - C_0)$ where 0, 1 respectively index the pre- and post- distinctions.

continued on next page

Triple Difference:

$y_{it} = \alpha + \beta_1 treat_i + \beta_2 post_t + \beta_3 male_i + \beta_4 treat_i \times male_i + \beta_5 post_t \times male_i + \beta_6 treat_i + post_t + \beta_7 treat_i \times post_t \times male_i + \varepsilon_{it}$

β_1 captures the effect of being in a treated school vs. a control school for females

β_2 captures how control school females differ from pre- to post- (even though they are not treated).

β_3 captures how males differ from females in the pre-period.

β_4 captures by how much more do males in the treatment group (vs. the control) differ compared to females in the treatment group (vs. the control) in the pre-period

β_5 captures by how much more does y change for control school males between pre- and post- relative to control school females.

β_6 captures by how much more do treatment school students change between pre- and post- compared to control school students.

β_7 is the triple difference estimator: captures the difference between treatment and control groups, male vs. female students, and pre- vs. post- period.

Continued →

$$\beta_7 = [(M_{T_1} - M_{T_0}) - (F_{T_1} - F_{T_0}) - (M_{C_1} - M_{C_0}) - (F_{C_1} - F_{C_0})]$$

where M = male

F = female

T = treatment

C = control

0 = pre-period

1 = post-period

For the standard errors, you should report the clustered standard errors at the school level which should account for the correlation between students' errors and over time within schools.

Alternatively, you could cluster at the school × year level if there are not enough schools (<42), but this assumes students' errors within a school are not correlated over time.