Microeconometrics
Spring Semester 2019/20
Problem set 3: Solution topics

Exercise 1 (Panel data models)

For this exercise please create a do-file and corresponding log-file to hand in with your solution.

Consider a subset of the data used by Vella and Verbeek (1998) *wagepan.dta* to estimate the effects of unions on workers' wages. The dataset is comprised of 545 men who worked in every year from 1980 through 1987 in the United States.

Consider the following model:

$$lwage_{it} = a_i + \theta_t D_t + \beta x_{it} + \pi z_i + u_{it}$$

where i denotes the worker and t the year. The vector x_{it} is comprised of *exper* (labor market experience) and its square *expersq*, *married* equals 1 if the individual is married, and *union* equals 1 if the worker is unionized. The vector z_i includes the variables *black* equals 1 for blacks, *hisp* equals 1 for hispanic workers, and *educ* denotes the number of years of education.

(a) Explain which effects parameters θ_t and a_i are likely to capture.

The coefficients of the time dummies (δ_t) capture the aggregate shocks in a specific year that wages of all individuals in that particular year.

The term α_i captures the time-invariant characteristics of a worker i that affect wages and might also be correlated with the explanatory variables. Omitting this term may lead to an omitted variable bias

problem if the worker fixed effects are correlated with the explanatory variables.

(b) If unions are successful in their wage negotiations with employers, what should be the sign of β_{union}?

The estimated coefficient $\widehat{\beta}_{union}$ is expected to be positive. Unionized workers earn more on average than non-unionized workers, ceteris paribus.

(c) Estimate the equation by pooled OLS. Do you find any evidence for a union effect? Are the assumptions required for these estimates to be consistent plausible? If not, what would be the asymptotic bias you would expect in the union estimate?

According to the estimates, unionized workers earn on average approximately 18.25% more than non-unionized workers.

Stata command:
reg lwage educ black hisp exper expersq married union i.year, vce(robust)

The assumptions required for consistency are $E(u_{it}|x_{it}, a_i)$ and $Cov(x_{it}, a_i) = 0$, where $\mathbf{x}_{it}$ includes the explanatory variables in the estimated equation. In this case it is unlikely to hold. The classical example is worker's ability that affects wages and are likely to be correlated with the explanatory variables.

(d) Now estimate the model in first differences (FD). Can we estimate the returns to education in FD? Why? What about race effects and experience?

Stata command:
reg D.lwage D.educ D.black D.hisp D.exper D.expersq D.married D.union i.year, vce(robust)

The variables exper and educ are redundant because the first differenced variables ($exper_{i,t} - exper_{i,t-1}$, for example) assume the same range of values for all individuals in the sample. This happens due to the deterministic nature of the variables. The same is true for the race variables. Because they are time-invariant, their coefficients are not identified in the first differenced model.

(e) Comment what the FD results of the union effect suggest on the correlation between union and a_i.

According to the estimates, unionized workers earn on average approximately 4.11% more than non-unionized workers. The estimated coefficient is statistically significant at 5% significance level. Because the magnitude of the estimated coefficient on union is substantially lower than that of the pooled OLS, the results suggests the presence of correlation between union and a_i.

(f) Considering the time-varying variables, estimate the equation using the within estimator and the random effects estimator. What are the necessary assumptions for consistency of the random effects estimator?

Stata commands:
xtreg lwage exper expersq married union i.year, re vce(robust)
xtreg lwage exper expersq married union i.year, fe vce(robust)

The assumptions for consistency of the RE estimator are a_i i.i.d. (a, σ_a^2), i.e., $Cov(x_{it}, a_i) = 0$.

(g) Perform an Hausman test for fixed effects. What do you conclude? Does the random effects estimator or the pooled OLS provide consistent estimates? State clearly the null and alternative hypotheses, the test statistic, and the critical region.

The Hausman test compares two estimators that, under the null

hypothesis H_0, should be consistent for the same parameter. The null hypothesis can be written as:

$$E[a_i|x_{it}] = 0$$

meaning that, under the H_0, the individual effects are uncorrelated and both the FE and the RE estimators provide consistent estimates. Under the alternative hypothesis H_a only the FE estimator yields consistent estimates.

The Hausman test statistic can be formalized as:

$$H = (\widehat{\beta}_{FE} - \widehat{\beta}_{RE})'\big(\widehat{V}(\widehat{\beta}_{FE} - \widehat{\beta}_{RE})\big)^{-1}(\widehat{\beta}_{FE} - \widehat{\beta}_{RE}) \sim \chi^2(k)$$

Stata:
xtreg lwage exper expersq married union i.year, re
est store re
xtreg lwage exper expersq married union i.year, fe
est store fe
hausman fe re, sigmamore

The $\chi^2(5)$ is equal to 23.54 and the associated p-value is equal to 0.0003 meaning that we reject the H_0 in favor of the H_a and the RE or the pooled OLS estimators are inconsistent.

Exercise 2 (Differences in differences)

The goal of this exercise is to apply some basic difference-in-difference estimations.

The specific intervention we will be analyzing is the construction of new secondary schools in Wonderland. In Wonderland, there were some communities that had new secondary schools, and other communities

that did not- this leads to the variation necessary to apply difference-in-difference. In short, we want to see if individuals who lived in areas with new secondary schools completed more school than individuals who lived in areas without secondary schools.

We will be looking at individuals who live in communities, so there are some individual level variables refer to individual characteristics, such as gender, age, education and some community level variables refer to characteristics of the community, such as access to clean water, electricity, and paved roads. Communities where a new secondary school was built will be known as "treatment" communities. Communities where no secondary school was built will be known as "control" communities. We will also look at two cohort groups. Young cohorts (aged 6-16 in 1985) and old cohorts (aged 21- 41) in both treatment and control communities. The idea is that new secondary schools should only affect young people who are still in school. If you have completed your studies, a new secondary school in your community will not change how much education you get. The idea is that the "treatment" or new secondary schools should only affect the young cohort living in treatment communities. This generates the difference- in-difference design.

1. Present individual summary statistics for the study sample for treatment communities vs. control communities. Do a t-test to see whether differences in age, education, and gender are statistically different between the treatment and control group. Do you see any differences? Are you concerned by any of the differences? How could they affect the analysis? (hint use: `ttest (variable), by(treat)`

   Stata commands:

   summary statistics:

   ```
   bysort treat:  sum sex age falive94 malive94 bothfmdead94
   educ secschol distschl motoroad roadquality electric pipwater
   ```

```
bar distcapital primary
```

ttest:
```
foreach x in educ age sex{
display "`x'"
ttest `x' if ycohort == 1, by(treat)
}
```

```
. bysort treat: sum sex age  falive94 malive94 bothfmdead94 educ  secschol distschl motoroad roadquality electric pipwater bar distcapital primary

-> treat = 0

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
         sex |       953    .4428122    .4969796         0          1
         age |       953    24.67996    9.704207        15         50
    falive94 |       855    .5450292    .4982597         0          1
    malive94 |       840    .7071429    .4553446         0          1
bothfmdead94 |       686    .2142857    .4106253         0          1
-------------+--------------------------------------------------------
        educ |       873    5.917526    2.466301         0         20
    secschol |       953    1.946485    .2251767         1          2
    distschl |       953    20.05467    20.46226        .1         80
    motoroad |       953    .9496327     .218165         0          1
roadquality |       953     .544596    .4982687         0          1
-------------+--------------------------------------------------------
    electric |       953    .2402938    .4274862         0          1
    pipwater |       953    .1867786    .3899383         0          1
         bar |       953    .6222455    .4850803         0          1
distcapital |       953    71.13905    73.96054         0     217.79
     primary |       953    .6789087    .4671409         0          1

-> treat = 1

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
         sex |       246    .4918699    .5009531         0          1
         age |       246     24.1748    9.324859        15         50
    falive94 |       211    .5118483    .5010483         0          1
    malive94 |       201    .7164179    .4518618         0          1
bothfmdead94 |       169    .1775148    .3832393         0          1
-------------+--------------------------------------------------------
        educ |       236    6.491525    2.187958         0         17
    secschol |       246    1.780488     .41476         1          2
    distschl |       246    2.336992    1.601343        .1          5
    motoroad |       246           1           0         1          1
roadquality |       246    .5650407    .4967624         0          1
-------------+--------------------------------------------------------
    electric |       246    .3902439    .4887994         0          1
    pipwater |       246     .300813    .4595465         0          1
         bar |       246    .5569106    .4977634         0          1
distcapital |       246    50.95829    28.10944     12.49      93.78
     primary |       246    .7804878     .41476         0          1
```

Results from the t-test on age and gender: p-values> 0.1*, hence we fail to reject the null hypothesis* $H_0 : \beta_j = 0$*. The average age and gender composition is not statistically different from the average age and gender in the control group.*
Regarding education, the t-test shows that on average individuals in the treatment communities have half a year more of schooling. In the dataset, it is not clear where this is an outcome or a baseline variable. If we do ***tab educ primary*** *we notice that primary (our outcome variable) and the variable named education do not have a clear correspondence, so it does not seem to be an outcome variable (if it was we would have observed primary=1 for education>6). Most likely it is a baseline variable, those who did not have primary*

education after the treatment did not have more than 6 years of education before the treatment (with some measurement error, that is why it is not clear).
The relevant point here was to notice that if it was an outcome variable, we expected the averages to be different as part of that difference would capture the effect of the program. Regarding baseline characteristics, our identification strategy is not affected by differences in levels, actually we would expect some differences in the main variables as the selection of the communities was not random. However, we would have expected that treatment communities had on average less years of eduction than the control communities and not the reverse, as the data seems to suggest.
Our identification strategy would be compromised if each community follows different time trends, as we can only eliminate characteristics that follow common trends.

2. Do a *t*-test to see whether differences in access to electricity (electric), piped water (pipwater), and distance from the capital (distcapital) are statistically different between the treatment and control group at the community level . Do you see any differences?

   Stata commands:
   ttest: `foreach x in electric pipwater distcapital{`
   `display "‘x’"`
   `ttest ‘x’ if` $ycohort == 1$`, by(treat)`
   }
   The results of the t-test show that the communities are statistically different, however and especially in the short run, we don't expect these characteristics to change over time. Therefore, it does not compromise our identification strategy.

3. Based on your answers to questions 1 and 2, do you think that treatment communities are different from control communities?

Yes, they are.

4. Run the following regression $Primary_i = \alpha + \beta T_i + u_i$, only for the young cohort (ycohort==1). Based on your estimates, can we say that constructing new secondary schools has a direct and causal impact on primary school completion? Why?

```
. reg primary treat if ycohort==1, cluster(cluster)

Linear regression                               Number of obs =     873
                                                F(  1,    50) =    5.57
                                                Prob > F      =  0.0222
                                                R-squared     =  0.0128
                                                Root MSE      =  .38485

                                  (Std. Err. adjusted for 51 clusters in cluster)
------------------------------------------------------------------------------
             |               Robust
     primary |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       treat |   .1074364   .0455247     2.36   0.022     .0159974    .1988755
       _cons |   .7942029   .0245633    32.33   0.000      .744866    .8435398
------------------------------------------------------------------------------
```

No, we can say that constructing new secondary schools is positively correlated with completing primary school, but we have learnt that the treatment and the control groups are different in some baseline characteristics and therefore we cannot make any inference regarding causal treatment effects.

5. Do the regression again but add additional controls. Explain any difference with previous results.

This is an open question, you could have chosen any covariates to include in the regression. The best way to proceed is to run a general ttest and identify those characteristics that are significantly different across treatment groups, and then include them in the regression. What we observe is that once we control for individual and community characteristics the estimate on treatment is not statistically different from zero. This result confirms that on

the previous regression our results were driven by omitted variables bias.

6. We are now just going to get the mean value for each cohort group and complete the following table:

	Treatment commun.	Control commun.	1st Diff
Young Cohort	.9016393	.7942029	0.1074364
Old Cohort	.4285714	.3764259	0.0521455
1st Difference	0.4730679	0.417777	0.0552909

Compute sample means and complete the table.

7. Now, estimate the treatment effect using a standard diff-in-diff regression

$$Primary_i = \alpha + \beta_1 T_i + \beta_2 YC_i + \beta_3 (T^* YC) + u_i \qquad (1)$$

How does it compare your estimate to the treatment effect obtained in the table?

Running equation 1 we see that the coefficient of the interaction term (β_3) captures the treatment effect, while β_2 captures the time effect and β_1 the pre-treatment difference between treatment and control groups.

8. Based on your analysis do you think building new secondary schools is effective at increasing primary school completion rates? What are some potential problems with the above analysis?

The estimates suggest a positive correlation between building secondary schools and the average level of education at the community level. We know that the estimates from section 1.4 are biased, as we have seen that the control and treatment communities are

not balanced in the baseline characteristics. Once we control for time-invariant characteristics and common trends we find that the treatment effect is not statistically different from zero at 10% significance level. Therefore, we cannot claim that the construction of new secondary schools have a positive impact on finishing primary education.

Exercise 3 (Reviewing LATE)

Say children can enroll in child care or not, and denote enrollment as $D_i = 1$ and non-enrollment as $D_i = 0$. Several researchers and politicians argue that child care may have positive effects on children cognitive development, while others argue that it has negative effects, particularly on young children. Let y_i be child performance on a language test at age 7.

1. The difference-in-means estimator $\bar{d} = \bar{y_1} - \bar{y_0}$ compares the observed means in the subpopulations with $D_i = 1$, $\bar{y_1}$, and with $D_i = 0$, $\bar{y_0}$. Show that $\bar{d}$ is a biased estimator for the average treatment effect on the treated, and derive an expression for the bias. What sign do you think the bias takes in this application? Explain. (*Hint: Take expectations* $\rightarrow E(y_1 - y_0)$)

   Taking expectations,
   $E(y_1|D = 1) - E(y_0|D = 0) = E(y_1|D = 1) - E(y_0|D = 1) + E(y_0|D = 1) - E(y_0|D = 0) = TT + Selectionbias$
   In this application, the bias may be positive if ability is inherited and more able mothers are more likely to work and demand child care. It may be negative if low income mothers are more likely to demand child care in order to gain income.

2. Say you get access to a program that randomly allocates a subset of child care places to children on the waiting list. Let $Z_i = 1$ if

the child receives an offer, and $Z_i = 0$ if the child does not receive an offer.

(a) Who are the compliers in this application? Who are the always-takers and never-takers? Do you think there might be defiers?

The compliers are children of mothers that apply to the subset of child care that is randomly allocated, and that enroll if offered a place but not if they are not offered a place.
In contrast, never-takers are children of mothers that do not apply, and always-takers are children of mothers that will find and enroll their children even if they do not get a child care place in the random allocation.
Defiers are children of mothers who enroll their child in child care if they are not offered a place, but do not enroll their child if they are offered a place. It is hard to think of examples here, but in principle, some mothers may not realize until after the offer that child care places in the randomized subset are of inferior quality, and therefore drop the program after the offer, while they might have been offered an alternative place if they did not get a randomized offer which was ex-post preferred.

(b) Under what conditions is the local average treatment effect (LATE) for the complier group identified?

- *Random assignment* $Z_i \perp \{y_{i0}, y_{i1}, D_{0i}, D_{1i}\}$. *Is needed to break the correlation between the treatment and the potential outcomes of y.*
- *Exclusion:* $y_i(d, 1) = y_i(d, 0)$.*Z runs only through the treatment. Only then can we interpret the effect of the instrument as representing the impact of the treatment.*
- *Relevance (First stage):* $E[D_{1i} - D_{0i}] \underbrace{=}_{Ass.1} E[D_i|Z_i = 1] - E[D_i|Z_i = 0] \neq 0$. *The instrument has to have a significant effect on treatment.*

- *Monotonicity:*

$$\text{Either, } D_{i1} \geq D_{i0} \forall i$$

$$\text{or, } D_{i1} \leq D_{i0} \forall i$$

Is needed to scale the effects correctly, since non-monotonicity implies that the effects of treatments being turned on are offset against effects of treatments being turned off.