André Filipe Silva 26005
João Seixo 40510
Márcia Serra 41221

MICROECONOMETRICS PROBLEM SET 3

EXERCISE 1

a)

θ_t is the coefficient of the time dummies D_t. It captures the aggregate shocks for a given year in the sample – shocks that affect all workers' wages in that particular year.

The parameter a_i, on the other hand, captures time-invariant characteristics of workers that affect wages and that possibly correlate with explanatory variables. It is important to include this term to try to avoid omitted variable bias problems.

b)

The sign of β_{union} should be positive – this would mean that being unionized leads to an increase in the worker's wage, on average, all else constant.

c)

We want to estimate the following equation:

$$\begin{aligned}\text{lwage}_{\text{it}} = {} & \text{a}_{\text{i}} + \theta_1 \text{d81} + \theta_2 \text{d82} + \theta_3 \text{d83} + \theta_4 \text{d84} + \theta_5 \text{d85} + \theta_6 \text{d86} + \theta_7 \text{d87} + \beta_1 \text{exper}_{\text{it}} \\ & + \beta_2 \text{expersq}_{\text{it}} + \beta_3 \text{married}_{\text{it}} + \beta_4 \text{union}_{\text{it}} + \pi_1 \text{black}_{\text{i}} + \pi_2 \text{hisp}_{\text{i}} + \pi_3 \text{educ}_{\text{i}} \\ & + \mu_{\text{it}}\end{aligned}$$

Where the variables d81 through d87 are dummy variables that take the value 1 if the observation reports to that year (t=1981, t=1982, …, t=1987). Example: if the observation reports to 1984 (t=1984), d84 = 1, and all the other dummies will equal 0. The base category is t = 1980, occurring when d81 through d87 are zero.

And our coefficient of interest is β_4.

Computing a t-test for individual significance of the coefficient:

$H_0: \beta_4 = 0$

$H_1: \beta_4 \neq 0$

The test statistic: $\text{t} = \frac{\widehat{\beta_4} - \beta_4}{\widehat{Var_{\beta_4}}} \sim t_{(4360-14)}$

The critical value for t is +- 1.960 (this is a two-tailed test).

We get, from our regression output, t= 6.65 > |1.960|, so we reject the null hypothesis and conclude for statistical significance at a 5% level. This could also be directly concluded from looking at p-value for this variable (since the p-value is 0.00 [smaller than 0.05] we reject the null).

Given this, there is statistically significant evidence for a positive union effect – being part of a union has a positive impact on the worker's wage.

The assumptions required for the Pooled OLS model to be consistent are two. Given

$$\mu_{it} = c_i + \epsilon_{it}$$

1. Weak exogeneity: $E[x_{it}\epsilon_{it}] = 0$
- This assumption seems plausible. We are essentially saying that there is no correlation between the variables of the vector x_{it} and the transitory (or random) part of the error term.
2. Uncorrelated unobserved effect: $E[x_{it}c_i] = 0$
- It is harder to find this assumption plausible. For instance, we can easily have endogeneity due to an unobserved effect of ability on labor market experience, which would simultaneously correlate with c_i , thus violating this assumption. This can lead to an asymptotic underestimation of the union effect.
- We should also consider that observations for the same individual are likely correlated with each other – so it is not optimal to weigh them equally. This also points towards Pooled OLS inconsistency.

d)

We want to estimate the following regression:

$$\begin{aligned} lwage_{it} - lwage_{it-1} &= a_i - a_i + \theta_t(D_t - D_{t-1}) + \beta_1(exper_{it} - exper_{it-1}) \\ &+ \beta_2(expersq_{it} - expersq_{it-1}) + \beta_3(married_{it} - married_{it-1}) \\ &+ \beta 4(union_{it} - union_{it-1}) + \pi_1(black_i - black_i) + \pi_2(hisp_i - hisp_i) \\ &+ \pi_3(educ_i - educ_i) + \mu_{it} - \mu_{it-1} \end{aligned}$$

With $\mu_{it} = c_i + \epsilon_{it}$

This simplifies to:

$$\begin{aligned} \triangle \text{lwage}_{\text{it}} = \theta_t \triangle D_t + \beta_1 \triangle \text{exper}_{\text{it}} + \beta_2 \triangle \text{expersq}_{\text{it}} + \beta_3 \triangle \text{married}_{\text{it}} + \beta_4 \triangle \text{union}_{\text{it}} + \\ \triangle \epsilon_{it} \end{aligned}$$

The first thing we should observe, is that with our First Differences estimation, we lose one time period. We had 8 time periods initially (1980 to 1987), and we now have only 7 time periods (1981 to 1987). We excluded one time-dummy (d81) to avoid perfect collinearity.

We can also see by the model that all time-invariant variables disappear – here they correspond to all the variables from the vector z_i. As such, we can conclude that it is not possible to estimate, using First Differences, the returns on education or race.

As for experience, even though it is a time-variant variable, we should be cautious as in this equation we do not have the variable in levels, but in differences. When the variable experience is first-differenced, it becomes a vector of 1s – and this is not very informative. It is more appropriate to get information from $expersq_{it}$, and refrain from interpreting the estimates on the experience variable itself.

In fact, if we included a constant in the model, $exper_{it}$ would be perfectly collinear and disappear. This is one more reason to refrain from interpreting estimates on it from our equation.

e)

We get an estimate result for the union effect of 0.0411497. However, as can be read from the p-value (0.061), the union effect is not statistically significant at a 5% significance level.

So, we can see that the union effect is not statistically significant here, as opposed to the results found when estimating through Pooled OLS. What this suggests is that there probably exists a significant correlation between $union_{it}$ and a_i, indicating the relevance of unobserved time-invariant heterogeneity that may be correlated with the explanatory variables.

f)

The time-varying variables are the ones associated to vector x_{it}.

For the within estimator (fixed effects), we need to estimate the following equation:

$$\begin{aligned}\text{lwage}_{it} - \overline{lwage}_i &= \theta_t(D_t - \overline{D}) + \beta_1(exper_{it} - \overline{exper}_i) + \beta_2(expersq_{it} - \overline{expersq}_i) \\ &+ \beta_3\left(married_{it} - \overline{married}_i\right) + \beta_4(union_{it} - \overline{union}_i) + \epsilon_{it} - \bar{\epsilon}_i\end{aligned}$$

$$\Leftrightarrow \widetilde{lwage}_{it} = \beta_1\widetilde{exper}_{it} + \beta_2\widetilde{expersq}_{it} + \beta_3\widetilde{married}_{it} + \beta_4\widetilde{union}_{it} + \tilde{\epsilon}_{it}$$

The Random Effects estimation equation is the following:

(Note: for simplicity, from now on we are writing only the vector x_{it} and not decomposing it in its different variables, as we have already done so before and repeating it is time-consuming)

$$\left(lwage_{it} - \lambda\,\overline{lwage_i}\right) = \theta_t(D_t - \lambda\,\overline{D}) + \beta(x_{it} - \lambda\,\overline{x}_i) + \left(\mu_{it} - \lambda\,\overline{\mu}_i\right)$$

With $\lambda = 1 - \left(\frac{\sigma_\epsilon^2}{(T\sigma_c^2+\sigma_\epsilon^2)}\right)^{(1/2)}$, T= 1, … , 8

And $\mu_{it} = c_i + \epsilon_{it}$

(Note: The parameter here designed by λ is usually designed θ. However, since we already have a θ in our equation, we chose λ to avoid confusion).

Necessary assumptions for consistency of the Random Effects estimator:

1. Strict Exogeneity
   $E[\epsilon_{it}|x_{i1}, \dots, x_{iT}, c_i] = 0$
2. Uncorrelated unobserved effects
   $E[x_{it}c_i] = 0$

g)

The Hausman test for this specific case will compare Fixed Effects and Random Effects estimators, and if the null hypothesis holds, both provide consistent estimates. Since Pooled OLS is also consistent under the Random Effects model assumptions, if the Random Effects is consistent, so is the Pooled OLS model.

$H_0: E[\alpha_i|x_{it}] = 0$
$H_1: E[\alpha_i|x_{it}] \neq 0$

This means that, under H_0, the individual effects α_i are uncorrelated and the Random Effects estimator as well as the Fixed Effects estimator give us consistent estimates.
Under H_1 , only the Fixed Effects estimator gives us consistent estimates.

Hausman Test Statistic:

$$H = \left(\widehat{\beta_{FE}} - \widehat{\beta_{RE}}\right)' \left(\hat{V}\left(\widehat{\beta_{FE}} - \widehat{\beta_{RE}}\right)\right)^{-1} \left(\widehat{\beta_{FE}} - \widehat{\beta_{RE}}\right) \sim \chi^2(k)$$

The critical value for $\chi^2(4)$ is 0.711, and our test statistic equals 77. The associated p-value is 0.00. Given this, we reject the null hypothesis H_0 at a 1% significance level.

This means that the Random Effects estimator is inconsistent. This means that the Pooled OLS estimator is also inconsistent. The most appropriate estimator is the Fixed Effects estimator.

EXERCISE 2

Exercise 2.1

Here are some summary statistics, separated between control (treat==0) and treatment (treat==1) groups, on our dataset:

Summary statistics: N mean sd min max by(treat)
treat: 0

	N	mean	sd	min	max
cluster	953	27.475	12.84	4	51
hh	953	10.123	6.741	1	44
id	953	4.303	3.529	1	35
sex	953	.443	.497	0	1
age	953	24.68	9.704	15	50
fhere94	757	.262	.44	0	1
falive94	855	.545	.498	0	1
mhere94	717	.314	.464	0	1
malive94	840	.707	.455	0	1
	686	.214	.411	0	1
bothfmdead94					
educ	873	5.918	2.466	0	20
primnum	953	1.231	.422	1	2
secschol	953	1.946	.225	1	2
distschl	953	20.055	20.462	.1	80
distschl5km	953	.257	.437	0	1
num2schols	926	1.243	.633	1	3
public	926	.692	.462	0	1
private	926	.71	.454	0	1
motoroad	953	.95	.219	0	1
roadquality	953	.545	.498	0	1
electric	953	.24	.427	0	1
pipwater	953	.187	.39	0	1
bar	953	.622	.485	0	1
distcapital	953	71.139	73.961	0	217.79
primary	953	.679	.467	0	1
ocohort	953	.276	.447	0	1
ycohort	953	.724	.447	0	1
treat	953	0	0	0	0
schl2	953	0	0	0	0
ycohortxtreat	953	0	0	0	0
ycohort2xschl	953	0	0	0	0
ycohortxschl2	953	0	0	0	0
ocohortxtreat	953	0	0	0	0
treat: 1					
cluster	246	21.659	19.321	1	50
hh	246	10.492	5.983	1	26
id	246	4.297	3.043	1	18
sex	246	.492	.501	0	1
age	246	24.175	9.325	15	50
fhere94	193	.233	.424	0	1
falive94	211	.512	.501	0	1
mhere94	178	.382	.487	0	1
malive94	201	.716	.452	0	1
	169	.178	.383	0	1
bothfmdead94					
educ	236	6.492	2.188	0	17
primnum	246	1.577	.803	1	3
secschol	246	1.78	.415	1	2
distschl	246	2.337	1.601	.1	5

distschl5km	246	1	0	1	1
num2schols	246	1	0	1	1
public	246	.622	.486	0	1
private	246	.711	.454	0	1
motoroad	246	1	0	1	1
roadquality	246	.565	.497	0	1
electric	246	.39	.489	0	1
pipwater	246	.301	.46	0	1
bar	246	.557	.498	0	1
distcapital	246	50.958	28.109	12.49	93.78
primary	246	.78	.415	0	1
ocohort	246	.256	.437	0	1
ycohort	246	.744	.437	0	1
treat	246	1	0	1	1
schl2	246	.711	.454	0	1
ycohortxtreat	246	.744	.437	0	1
ycohort2xschl	246	.488	.501	0	1
ycohortxschl2	246	.528	.5	0	1
ocohortxtreat	246	.256	.437	0	1

First, we are going to test if differences in **age** in control and treatment group are significative, by performing a t-test. The group 0 is the control and the group 1 is the treatment.

$$H_0: Mean(control\ group) - Mean(treatment\ group) = 0$$

$$H_1: Mean(control\ group) - Mean(treatment\ group) \neq 0$$

Two-sample t test with equal variances

	obs1	obs2	Mean1	Mean2	dif	St_Err	t_value	p_value
age by treat: 0 1	953	246	24.68	24.175	.505	.689	.75	.464

As we can clearly observe in the table, with a p-value of 0.464, we do not reject the null hypothesis at a 5% significance level. We can state that there is no significant difference in the groups mean. Thus, we know that the groups are similar in gender structure. As the groups are similar, we know that there is no gender bias in the selection sample. This gender sample selection bias, if existing, could be problematic depending on whether the new school built was female or male only, or even for both genders.

Now, we are going to use the same approach to test the differences in **education** for both groups.

$$H_0: Mean(control\ group) - Mean(treatment\ group) = 0$$

$$H_1: Mean(control\ group) - Mean(treatment\ group) \neq 0$$

Two-sample t test with equal variances

	obs1	obs2	Mean1	Mean2	dif	St_Err	t_value	p_value
educ by treat: 0 1	873	236	5.918	6.492	-.574	.177	-3.25	.001

In this case, with a p-value of 0.001, we reject the null hypothesis at a 5% significance level. This means that there are significant differences between control and treatment group in education. The treatment group shows a higher value for years of education. However, this is not worrying because if we are testing whether building a secondary school leads to increases in primary school completion rates, then its expectable to find differences in educational attainment between control and treatment groups.

Regarding differences in gender, we have that:

$$H_0: Mean(control\ group) - Mean(treatment\ group) = 0$$

$$H_1: Mean(control\ group) - Mean(treatment\ group) \neq 0$$

	obs1	obs2	Mean1	Mean2	dif	St_Err	t_value	p_value
sex by treat: 0 1	953	246	.443	.492	-.049	.035	-1.4	.169

With a p-value of 0.169, we do not reject the null hypothesis at a 5% significance level. This tells us that the treatment and control groups are similar in age, which is fundamental in order for us to compare if indeed the building of a secondary school does in fact lead to an increase in primary school completion rate.

Exercise 2.2

In this exercise we have strong reasons to suspect of clustering effects. On the one hand, *electric* and *pipwater* are defined as electricity and piped water availability in the community. On the second hand, when we sort summary statistics by cluster for the *electric*, *pipwater* and *distcapital* variables, we see that the standard errors are zero within clusters. As such, it is reasonable to consider calculating the t-tests, considering the different clusters. For that reason, we shall provide two answers. On the first answer, we are considering the possibility of clustering effects, thus adjusting accordingly. On the second answer, we are discarding this possibility. As it shall be seen, this choice is not trivial as the effects will change dramatically our answer.

On both our answers, the test hypotheses are the following.

$$H_0: Mean(control\ group) - Mean(treatment\ group) = 0$$

$$H_1: Mean(control\ group) - Mean(treatment\ group) \neq 0$$

Once more the group 0 and 1 are respectively control and treatment. We will be presenting the tables with the results in the following order: ***electric***, ***pipwater*** and ***distcapital*.**

Using a t-test adjusted for clusters

Now, taking into account clusters, we shall use the package cltest, which allows us to perform a t-test on clustered data. Our dataset cluster's ID is cluster. The results and their interpretation are presented below:

electric:

```
t-test adjusted for clustering
electric by treat, clustered by cluster
------------------------------------------------------------------------------
 Intra-cluster correlation         =           1.0000
------------------------------------------------------------------------------
             N    Clusts    Mean          SE             95 % CI
treat=0     953     41      0.2403      0.0725       [  0.0938,  0.3868]
treat=1     246     10      0.3902      0.1524       [  0.0454,  0.7351]
------------------------------------------------------------------------------
Combined   1199     10      0.2711      0.0656       [  0.1393,  0.4028]
------------------------------------------------------------------------------
Diff(0-1)  1199     51     -0.1500      0.1688       [ -0.4892,  0.1893]

Degrees freedom:    49

                  Ho: mean(-) = mean(diff) = 0

 Ha: mean(diff) < 0          Ha: mean(diff) ~= 0          Ha: mean(diff) > 0
      t =  -0.8883                t =  -0.8883                t =  -0.8883
  P < t =   0.1893            P > |t| =   0.3787            P > t =   0.8107
```

With a p-value of 0.1893, at a 5% significance level, we do not reject the null hypothesis. There is no statistically significant difference between treatment and control, at the cluster level, of access to electricity. This is good because it shows that the both the treatment and control groups are similar. If access to electricity was different between cluster treatment and control groups, then we could expect to find some differences in education attainment related to access to electricity.

Households which do not have access to electricity are more likely poorer, which is usually correlated with lower educational attainment. At the same time, lack of electricity may negatively affect educational attainment as it could imply increased difficulties in studying. As such, it is important that both the treatment and control cluster groups have similar access to electricity.

pipwater

```
t-test adjusted for clustering
pipwater by treat, clustered by cluster
------------------------------------------------------------------------------
 Intra-cluster correlation         =           1.0000
------------------------------------------------------------------------------
             N    Clusts    Mean           SE              95 % CI
treat=0     953     41      0.1868      0.0666       [  0.0521,   0.3215]
treat=1     246     10      0.3008      0.1401       [ -0.0162,   0.6178]
------------------------------------------------------------------------------
Combined   1199     10      0.2102      0.0603       [  0.0891,   0.3313]
------------------------------------------------------------------------------
Diff(0-1)  1199     51     -0.1140      0.1552       [ -0.4259,   0.1978]

Degrees freedom:    49

                  Ho: mean(-) = mean(diff) = 0

 Ha: mean(diff) < 0          Ha: mean(diff) ~= 0          Ha: mean(diff) > 0
      t =  -0.7349                 t =  -0.7349                 t =  -0.7349
  P < t =   0.2330          P > |t| =   0.4659           P > t =   0.7670
```

With a p-value of 0.2330, at a 5% significance level, we do not reject the null hypothesis. There is no statistically significant difference between treatment and control, at the cluster level, of access to piped water. Once more this is good as it shows that both the treatment and control groups are similar.

Just like in access to electricity, households which do not have access to piped water are more likely poorer, which is usually correlated with lower educational attainment. At the same time, when households do not have access to piped water, it is reasonable to expect that they do not have access to safe drinking water sources. When that is the case, we know that the household's health is deteriorating, when compared to the health of households who have access to piped water. We also know that with deteriorated health, educational attainment tends to be lower as individuals face increased constraints in their learning process.

Then, since the access to piped water might affect educational attainment, it is important to keep this in mind and assure that both the treatment and control cluster groups, have similar characteristics in access to piped water, so that there is no sample bias.

distcapital

```
t-test adjusted for clustering
distcapital by treat, clustered by cluster
------------------------------------------------------------------------------
 Intra-cluster correlation         =           1.0000
------------------------------------------------------------------------------
             N    Clusts     Mean           SE             95 % CI
treat=0     953     41     71.1390     11.0484       [ 48.8095, 93.4686]
treat=1     246     10     50.9583     23.2338       [ -1.6002,103.5168]
------------------------------------------------------------------------------
Combined   1199     10     66.9985      9.9920       [ 46.9190, 87.0781]
------------------------------------------------------------------------------
Diff(0-1)  1199     51     20.1808     25.7270       [-31.5195, 71.8810]

Degrees freedom:    49

                  Ho: mean(-) = mean(diff) = 0

 Ha: mean(diff) < 0          Ha: mean(diff) ~= 0         Ha: mean(diff) > 0
      t =   0.7844                t =   0.7844                t =   0.7844
  P < t =   0.7817            P > |t| =   0.4366            P > t =   0.2183
```

With a p-value of 0.7817, at a 5% significance level, we do not reject the null hypothesis. There is no statistically significant difference between treatment and control, at the cluster level, of proximity to the capital. Once more this is good because it shows that the both the treatment and control groups are similar.

Now, if there was a difference, it would be problematic. That is so because we can consider that countries' capitals usually haver more schools, training facilities, universities, amongst others. As such, it is also reasonable to expect that households closer to the capital, would have more financial ease in accessing the educational opportunities presented by the capital.

This would be even more noticeable if we are comparing communities from small villages with no school in the vicinities. The village closer to the capital, would have a better chance at promoting the educational opportunities of its community. As such, to ensure that there is no sample bias, it is once again important to assure that both the treatment and control communities are at approximately the same distance from the capital.

Using a normal t-test

Now, we shall provide the answers for the normal t-test without considering possible cluster effects. It's important to state that the results are very different and, as such, will change our answers.

electric

Two-sample t test with equal variances

	obs1	obs2	Mean1	Mean2	dif	St_Err	t_value	p_value
electric by treat:~1	953	246	.24	.39	-.15	.032	-4.75	0

Here, we see that there are significant differences between treatment and control group in access to electricity. With a p-value of 0, we reject the null hypothesis of no difference between treatment and control group in access to electricity at a 5% significance level.

This can be problematic in the sense that it could possibly distort the results. This happens because it is reasonable to expect that households who have access to electricity are more likely to have completed primary school. This happens because usually households that do not have access to electricity tend to be poorer. At the same time, poorer households also tend to have lower educational achievements, making them less probable of completing primary school.

In this case, households in the treatment group have higher access to electricity, making the treatment and control group less comparable between each other.

pipwater

Two-sample t test with equal variances

	obs1	obs2	Mean1	Mean2	dif	St_Err	t_value	p_value
pipwater by treat:~1	953	246	.187	.301	-.114	.029	-3.95	0

Here, we see that there are significant differences between treatment and control group in access to piped water. With a p-value of 0, we reject the null hypothesis of no difference between treatment and control group in access to piped water at a 5% significance level.

This can be problematic in the sense that it could possibly distort the results. This happens because it is reasonable to expect that households who have access to piped water are more likely to have completed primary school. This can be so for two reasons. First, just like in the access with electricity, usually households that do not have access to piped water tend to be poorer. At the same time, poorer households also tend to have lower educational achievements, making them less probable of completing primary school. Adding to that, lack of access to piped water tends to imply decreased health conditions. Health conditions are also correlated with educational achievement. As such, it's quite reasonable to also expect lower probability of completing primary school for individuals in households which do not have access to piped water, as it implies a deterioration of their health and consequently their ability to function properly at school.

Once more, the treatment and control groups are not comparable on an area that might affect a priori the variable of interest.

distcapital

Two-sample t test with equal variances

	obs1	obs2	Mean1	Mean2	dif	St_Err	t_value	p_value
distcapital by tre~1	953	246	71.139	50.959	20.181	4.804	4.2	0

Here, we see that there are significant differences between treatment and control group in the distance to the capital. With a p-value of 0, we reject the null hypothesis of no difference between treatment and control group in the distance to the capital at a 5% significance level.

This, just like in the previous cases, is problematic. There are several reasons so. Firstly, a country's capital is the place where we can find more schools and opportunities. As such, households which are closer to the capital are most likely to complete primary school than household's who are further away from the capital. On the other hand, we could also be looking at spillover effects due to the proximity. Households in the control or treatment group could be benefiting from the proximity to the capital, thus leading to biased results of the effect of building a new secondary school in primary school completion.

For that matter, we should be extremely cautious when interpreting the results from the following regressions, to make sure that we are considering the risks of spill-over effects and heterogeneity between the treatment and control group.

Exercise 2.3.

Based on our answer to question 2, considering clustering effects, we see that treatment communities are similar to control communities. This is important because it assures us that both the control are treatment communities are comparable. As access to piped water, electricity and even the distance to the capital are similar, then none of these characteristics would have a differentiating effect on the completion of school between the treatment and control group. This would then allow us to compare whether communities who live in areas with new secondary schools (treatment) would actually complete more school than communities who live in areas in which there are no new secondary schools (control).

However, if we do not consider those clustering effects, then we actually see that control and treatment communities are significantly different in terms of access to electricity, piped water and distance to the capital. These differences, as we have mentioned also in question 2, are very likely to affect the probably of completing primary school. As such, we should be extremely cautious when interpreting the impact of building a new secondary school, on the probability of completing primary school.

At the same time, in the individual level, we see no difference between the control and treatment groups. They are both similar in age and gender, which is important because it controls for situations in which different laws in the past or gender imbalances, could affect the school completion level. This would then pose significant problems for us to actually estimate the causal effect of building new secondary schools. However, since this is not the case, we are not worried about it.

The only difference at the individual level which we found was at the educational attainment level. However, this is expectable under the construction of our policy question. If we are expecting that the construction of a new secondary school would lead to higher school completion, then it is very reasonable to find significant differences between the treatment and control group in educational attainment. If this was not the case, then the construction of a new school would not lead to an increase in school completion.

Nonetheless, we should be cautious as we might need to study the differences between the treatment and control communities on other areas that might affect school completion. Some of these areas, which were not tested in our previous answers, include the distance to a school, the fact of whether that school is private or not, the fact of whether the closest relatives are alive (mother and father), amongst other factors that indirectly affect the school completion rate.

In that sense, considering what we studied so far, we can say that the treatment and control groups are similar, but only up to a certain point and depending on whether we are considering cluster groups or not. At any rate, we should be extremely cautious in that interpretation as other factors that we are not considering nor testing, may affect school completion rates and may differ between control and treatment groups.

Exercise 2.4.

Note: All regressions done from here on forward do not cluster, as was indicated in the instructions to solving this problem set.

Linear regression

primary	Coef.	St.Err.	t-value	p-value	[95% Conf	Interval]	Sig
treat	0.107	0.032	3.36	0.001	0.045	0.170	***
Constant	0.794	0.015	54.21	0.000	0.765	0.823	***

Mean dependent var	0.817	SD dependent var	0.387
R-squared	0.013	Number of obs	873.000
F-test	11.272	Prob > F	0.001
Akaike crit. (AIC)	812.226	Bayesian crit. (BIC)	821.770

**** p<0.01, ** p<0.05, * p<0.1*

$H_0: \beta = 0$
$H_1: \beta \neq 0$

With a p-value of 0.001, we reject the null hypothesis of no significance of the treat coefficient at a 5% level. Being in the treatment group leads to an increase in the probability of completing primary school of 10.7 p.p.

Nonetheless, like we have stated before, we need to be very careful before we can say whether building or not a new secondary school has a direct and causal impact on primary school completion. We should first control for extra covariates to test the robustness of our estimates. If, by adding extra controls, our estimates change significantly, then it means that there is no causal impact of building a new secondary school, on primary school completion.

Exercise 2.5.

When doing the previous regression with extra covariates, we should be careful. Firstly, we should not add the variable *educ* as it is educational attainment. Regressing the probability of completing primary school with educational attainment as a covariate, will confound the effect of other covariates. That is because larger educational attainments, by construction, imply larger probability of completing primary school. With this in mind, we will refrain from using this covariate.

We should refrain from inserting random covariates as that would only lead to overfitting. We must choose carefully, based on what we would expect to be the effect on primary school

attainment. We will refrain from using covariates for which there was not any statistically significant difference between treatment and control group.

From question 2, we shall include the *distcapital* and *pipwater* covariates. We will include the *distcapital* because it is safe to expect that differences between the household's distance to the capital might affect the primary school completion probability. We will also include the *pipwater* coefficient as it, like the electric covariate, will allow us to control for a possible measure of poverty and health-related issues that could affect primary school completion probability.

We will also be including the following covariates*: fhere94*, *mhere94*, to control for whether the father and mother are present or not. This allows us to control for the family stability, which is usually strongly correlated with educational attainment, thus the probability of completing primary school.

We will also add a final covariate: *num2schols*. This controls for the number of secondary schools in the area. We won't control for the distance to the schools, or the number of schools in a determined radius as we are already controlling for the distance to the capital (expected to have more secondary schools), and the number of schools existing.

$$Primary_i = \alpha + \beta_1 T_i + \gamma_1 distcapital_i + \gamma_2 pipwater_i + \gamma_3 fhere94_i + \gamma_4 mhere94_i + \gamma_5 num2schols_i + \mu_i$$

Our results are the following:

Linear regression

primary	Coef.	St.Err.	t-value	p-value	[95% Conf	Interval]	Sig
treat	0.095	0.039	2.44	0.015	0.018	0.171	**
distcapital	-0.001	0.000	-6.26	0.000	-0.002	-0.001	***
pipwater	0.006	0.037	0.17	0.868	-0.067	0.080	
fhere94	0.003	0.035	0.09	0.927	-0.065	0.071	
mhere94	0.077	0.034	2.27	0.023	0.010	0.144	**
num2schols	-0.012	0.029	-0.42	0.673	-0.069	0.045	
Constant	0.878	0.045	19.45	0.000	0.789	0.966	***

Mean dependent var	0.827	SD dependent var	0.379
R-squared	0.091	Number of obs	565.000
F-test	9.346	Prob > F	0.000
Akaike crit. (AIC)	465.866	Bayesian crit. (BIC)	496.223

**** p<0.01, ** p<0.05, * p<0.1*

$H_0: \beta_1 = 0$

$H_1: \beta_1 \neq 0$

With a p-value of 0.015, we reject the null hypothesis of no significance at a 5% level. When controlling for these extra covariates, for the subsample of the young cohorts, we see that being in the treatment group leads to an increase in the probability of primary school completion by 9.5 p.p., when compared to the control group, ceteris paribus.

It is interesting to note that only the *mhere94* and *distcapital* covariates had significant coefficients at a 5% level. A mother's presence in the household leads to an increase in the probability of completing the primary school by 7.7 p.p., compared to a situation in which a mother is not present, ceteris paribus. At the same time, it's interesting to note that the *num2schls* and *pipwater* had no statistically significant effect on the probability of completing primary school, suggesting that these are not relevant factors.

Nonetheless, we should still be cautious in our interpretation, and proceed with further robustness tests before we can say anything about causality.

Exercise 2.6.

In here we are calculating the difference in the primary school completion rate for young cohorts and old cohorts, between the treatment and control group. We then proceed to apply the differences between the young and old cohorts at both the treatment and control group and finalize by taking the differences of the differences. This result should be the actual effect of building a new secondary school, in the probability of completing the primary school.

	treatment	control	First diff
ycohort	0.9016393	0.7942029	0.1074364
Ocohort	0.4285714	0.3764259	0.0521456
First diff	0.4730679	0.417777	0.0552909

Note: The First diff figures are computed by simply subtracting the columns/rows.

Exercise 2.7.

Now, we will estimate once more the treatment effect, using a standard diff-in-diff regression which has an interaction term between the treatment and the young cohorts, which will be our coefficient of interest.

Linear regression

primary	Coef.	St.Err.	t-value	p-value	[95% Conf	Interval]	Sig
treat	0.052	0.058	0.90	0.371	-0.062	0.166	
ycohort	0.418	0.030	13.88	0.000	0.359	0.477	***
ycohortxtreat	0.055	0.068	0.82	0.414	-0.078	0.188	
Constant	0.376	0.026	14.70	0.000	0.326	0.427	***

Mean dependent var	0.700	SD dependent var	0.459
R-squared	0.182	Number of obs	1199.000
F-test	88.375	Prob > F	0.000
Akaike crit. (AIC)	1299.717	Bayesian crit. (BIC)	1320.074

*** $p<0.01$, ** $p<0.05$, * $p<0.1$

With a p-value of 0.414, our coefficient of interest is not significant at a 5% level. Nonetheless, its coefficient is the same as the diff in diff we got in table 6. However, as this interaction term is not significant, we can see that being in the young cohort and being a part of the treatment group, will not lead to any statistically significant difference in the probability of completing primary school, compared to those who are not in the young cohort treatment group.

Exercise 2.8.

Based on our estimates, we see that in fact building a new secondary school is not effective at increasing the primary school completion rates. This was easily shown above as the interaction term between young cohorts and control group had no statistically significant impact on the primary school completion rate.

Now, there are some key issues we should keep in mind with our former analysis. Firstly, when we were not considering the possibility of cluster communities, we saw that the treatment and control groups were statistically different in terms of access to electricity, piped water and distance to the capital. This raised several issues, that could affect primary school completion rate, which were addressed already on exercise 2.2.

However, we saw on exercise 2.5. that only distance to the capital had statistically significant impacts on the primary school completion rate, at a 5% significance level. Households further away from the capital had lower probability of completing the primary school.

On exercise 2.5., we also found out that family stability is a strong factor affecting the primary school completion rate. A mother's presence was found to be statistically significant at a 5% level, increasing the probability of completing primary school by 7.7 p.p.

This significant impact of distance to capital and family stability raised important questions regarding the comparability and heterogeneity of the treatment and control groups. As we found that, without using clusters, both the treatment and control groups were different in terms of distance to the capital, we must question ourselves whether they are similar enough to allow for comparability.

As such, for us to confirm, or not, if indeed a new secondary school has or not an effect in primary school completion rates, we should enrich our analysis. That enrichment could be achieved by changing the control group. As we have found that the treatment and control group are not similar, we should use other methods.

A possible method would be to use Propensity Score Matching to match the treatment and control groups. The choice of variables to be taken into account in the region of common support should be large enough to allow for us to encompass the heterogeneity inherent in the treatment and control groups.

With a new treatment group matched under a region of common support, we could be more certain that we were tackling the differences between the groups and, as such, we would be able to verify the robustness of our estimates in exercises 2.4, 2.5 and 2.7.

EXERCISE 3

Exercise 3.1.

To calculate the Average Treatment Effect on the Treated (TT), this is, the effect of childcare on children that haven been exposed to it, we would need to have the following:

$$E(y_1|D=1) - E(y_0|D=1)$$

However, what we are given is:

$$E(\bar{d}) = E(y_1 - y_0) = E(y_1|D=1) - E(y_0|D=0)$$

It could be the case that they are the same, only if $E(y_0|D=1) = E(y_0|D=0)$. However, if that is not the case, we will be facing **selection bias.**

$$E(y_1|D=1) - E(y_0|D=0)$$
$$= E(y_1|D=1) - E(y_0|D=1) + E(y_0|D=1) - E(y_0|D=0)$$

The component written in red is the selection bias. This component reflects the fact that individuals subjected, and individuals not subjected to the treatment could have pre-existing differences regardless of the application of the treatment. If the treatment was not completely random and depends on factors that also affect the outcome, we would have Selection Bias.

In this specific case, to analyze whether we have the problem with selection bias, we need to think if the two groups in question were random selected. This is, the group of children that go and the group that doesn't go to childcare. In this case, we think that this decision is careful made by the parents, and, thus, not random at all.

Children aged six and seven years old that don't go to child care after school maybe stay with their grandparents, or perhaps one of the parents doesn't work or works only part time to be able to play an important role in children's education, besides, it might also be the case that children have a private teacher/tutor after school. Hence, gathering all the hypothesis, it might be the case that children that don't go to childcare are children whose parents are intrinsically more concerned about their studies and that put a bigger effort on their education. Maybe their kids are more likely to get better grades the language test. If this is the case, then:

$$E(y_0|D=0) > E(y_0|D=1)$$

Which means the bias would be negative and the effects would be underestimated.

However, it could also be the case that children don't go to school because parents are unemployed (not a decision to stay home helping with the studies but simply don't have a job).

We could think that parents who don't have jobs maybe studied less an are not that keen on helping and motivating kids to study. Hence, this kids that don't go to school are more likely to perform worse on the language test, not being a good control group. If that is the case, then:

$$E(y_0|D=1) > E(y_0|D=0)$$

Which means the bias would be positive and the effects would be overestimated.
So, in fact, the bias could go either way.

Exercise 3.2a)

Compliers: Children whose mothers always will enroll their kids when offered the chance to enter a childcare unit, (Z=1) (D=1), but who don't enroll their kids if they weren't randomly assigned an offer.

Always Takers: Children that would enroll and attend the childcare unit (D=1) even if they didn't got an offer to enter (Z=0).

Never Takers: Children that will not enter the center (D=0) even if they got an offer (Z=1).

Defiers: The group of defiers are those who always do the opposite. This, if they are given a offer to enter the childcare unit (Z=1) they would not enter (D=0) and they were not given the offer (Z=0), they would enter (D=1). In this case, it doesn't seem plausible that parents only enroll their kids when they have no offer but wouldn't enroll them if they have.

Exercise 3.2b)

In this case, we are intrinsically using the offer of a place in the waiting list as an IV for the actual enrollment, but as we know, the effects of offering a place in childcare will translate in different behaviors as we have seen in a). Hence, we cannot assume that offering a place will affect all parents the same way. In this case, the IV estimates will produce a **Local Average Treatment Effect (LATE).** In this framework, there are four assumptions that must hold for us to get consistent estimates:

1. **Independence of Z: it must be orthogonal to $\boldsymbol{y_{i0}, y_{i1}, D_{i0}, D_{i1}}$**

This means that the offer of entering in the childcare unit must not be correlated with neither potential outcome or potential to enter not in the unit. This means, that the instrument must truly be **randomly assigned.**

2. Exclusion Restriction

The only causality channel between y_i and Z_i is through D_i. This is, the potential of the grade in the language test cannot be correlated with the attribution of the offer.

3. The instrument Z must influence D, the relation it cannot be zero

In this case, the offer of entering the child unit must have effects in the enrollment. Thus, there should be more kids enrolling if they were offered a pace in the waiting list.

$$E(D_{i1} - D_{i0}) = E(D_i|Z = 1) - E(D_i|Z = 0) \neq 0$$

4. Monotonicity

Monotonicity implies the absence of defiers. People can indeed correspond differently to the treatments but they should respond in the same direction. This is,

$$D_{i1} > D_{i0} \forall i$$

$$D_{i1} < D_{i0} \forall i$$

In this case and, in general, we assume

$$D_{i1} > D_{i0} \forall i$$

If all these assumptions hold, we would get the Local Average Treatment Effect, the effect of the treatment on the complier's group.

$$LATE =$$

$$= \frac{\mathrm{E}(y_i|Z_i = 1) - \mathrm{E}(y_i|Z = 0)}{\mathrm{E}(D_i|Z = 1) - \mathrm{E}(D_i|Z_i = 0)}$$

$$= E(y_{i1} - y_{i0}|D_{i1} - D_{i0} > 0)$$