**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 |
| bothfmdead94 | 686 | .214 | .411 | 0 | 1 |
| 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 |
| bothfmdead94 | 169 | .178 | .383 | 0 | 1 |
| 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.

**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.

**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:

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 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.

If the treatment and control group were significantly different in age, then we could not compare them because due to the difference in age, some groups could already be exposed to the building of previous schools or even different previous laws about mandatory school level attainment.

**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.

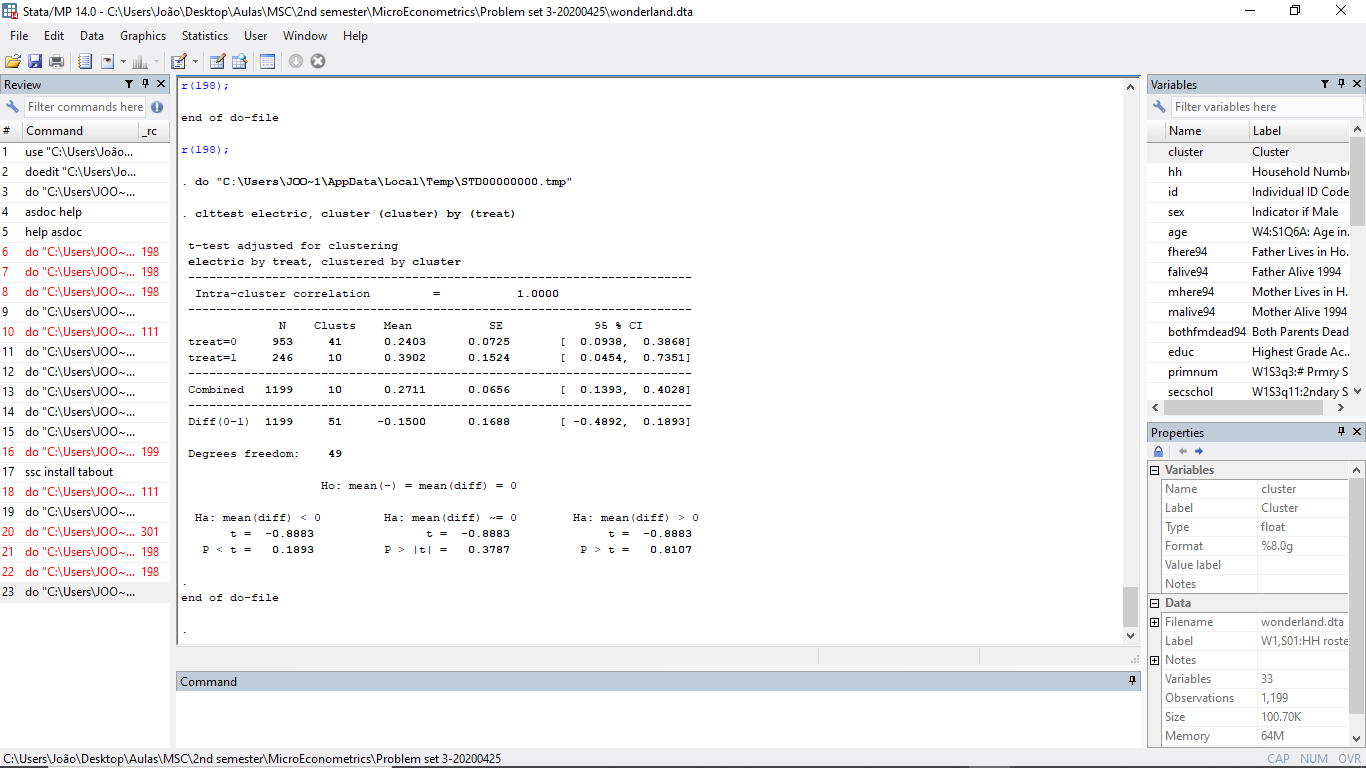
On both our answers, the test hypotheses are the following.

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:

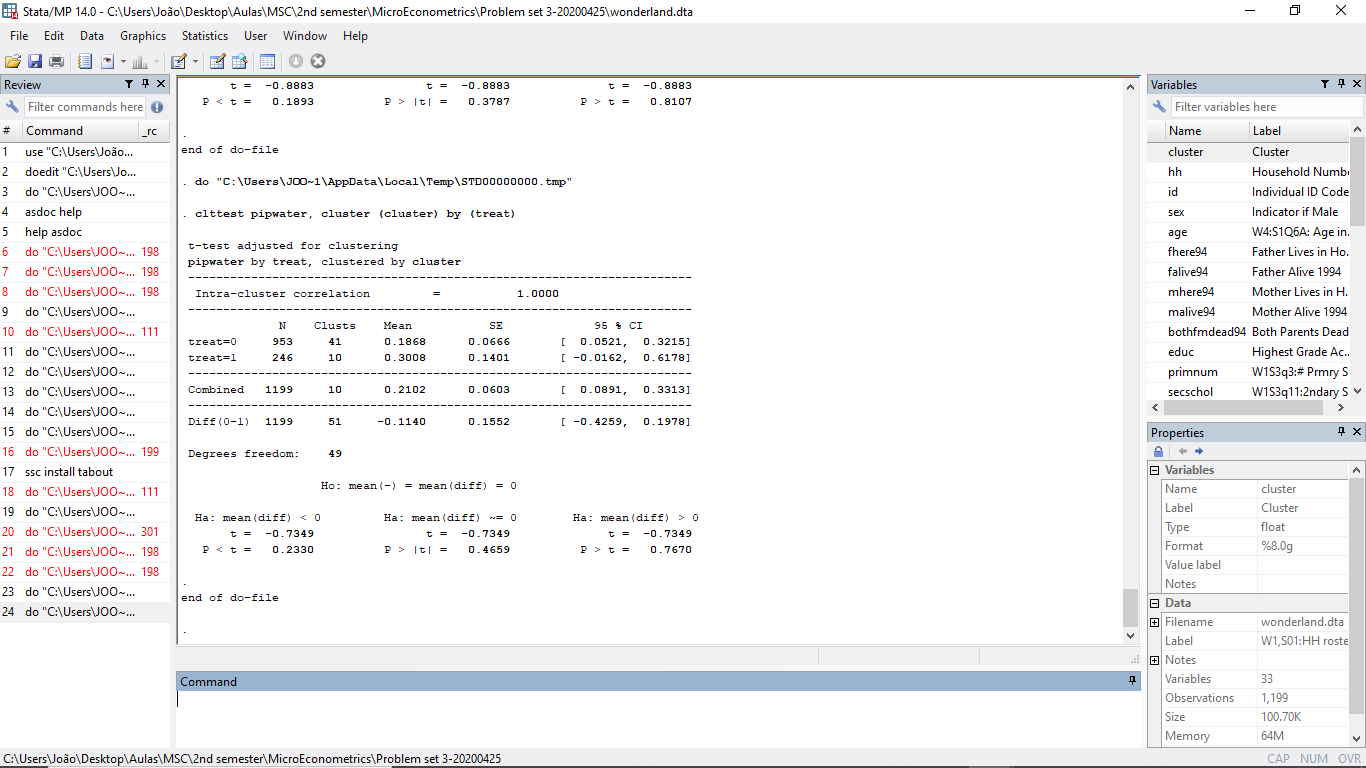
*electric:*



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*

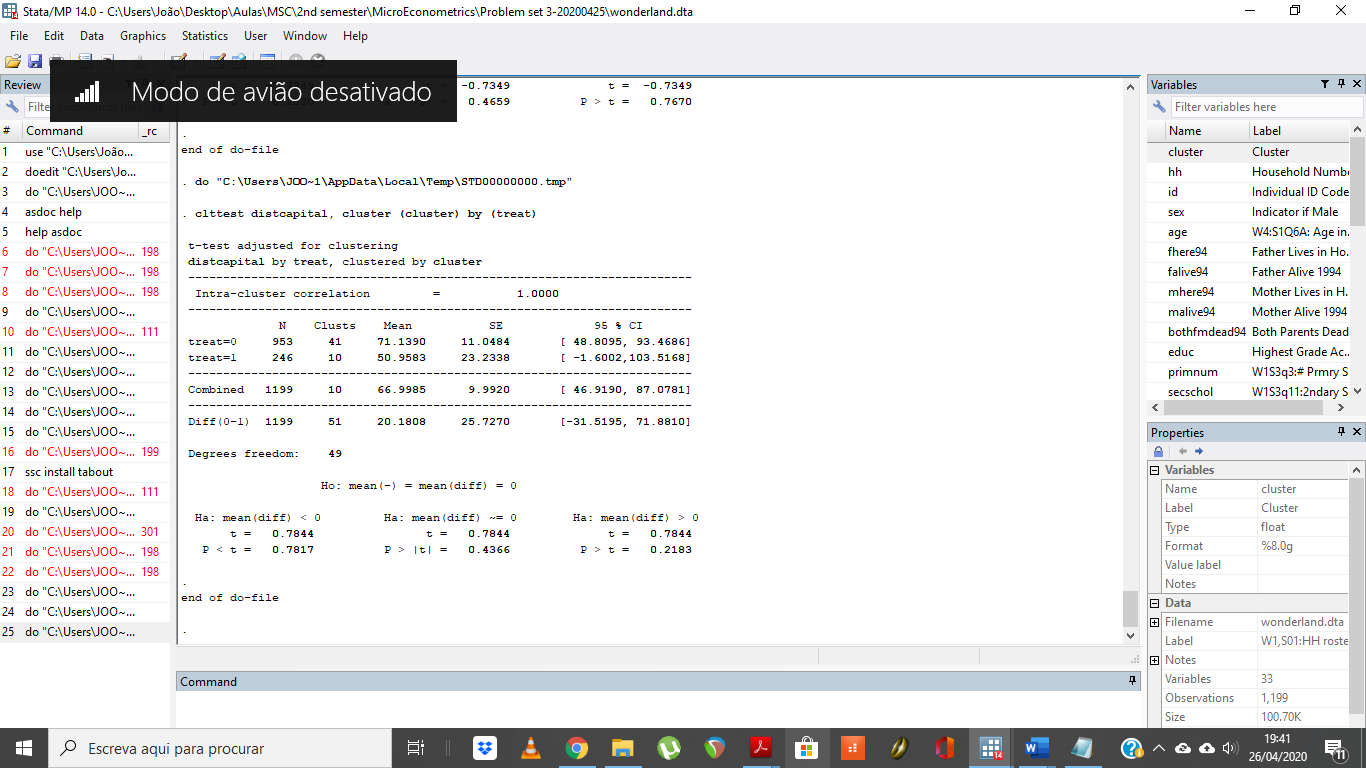


With a p-value of 0.2330, at a 5% significance level, we do not reject the null hypothesis. There is no statistically significant difference between treatment and control, at the cluster level, of access to piped water. Once more this is good as it shows that both the treatment and control groups are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



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* | | | | | | | | | |

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.

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* | | | | | | | | | |

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.