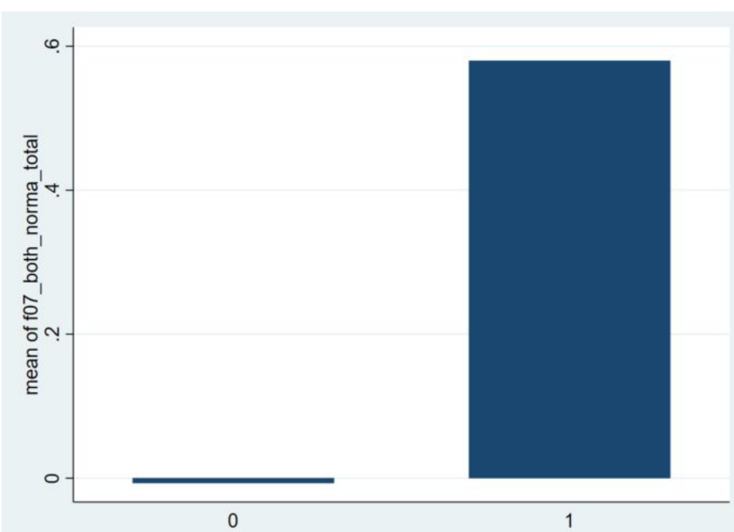
SI Yoke Hong - ECO372 Assignment 1
Student Nb: 1005815806

a. Explain briefly the research question, the treatment and outcomes that the researchers are investigating. (Max. 4 sentences.)

The research question of the paper is to estimate the 1-year effects of village-based schools in Afghanistan on the enrollment and academic performance of 1,490 primary school-age children. The treatment of this investigation is where the researchers use a randomized controlled trial to determine which villages the schools should be placed in. The outcomes of the investigation indicates that placing a school in a village improves academic participation and performance among all children, particularly for girls.

b.
(i) There are 40 levels of observation
(ii) 1728
(iii) treatment with label “Indicator set to one if village group assigned to treatment”
(iv) 892 participants were in a village with a school whereas 836 participants were not.
(v) f07_formal_school; f07_both_norma_total; s08_formal_school
(vi) f07_girl_cnt; s08_girls_cnt

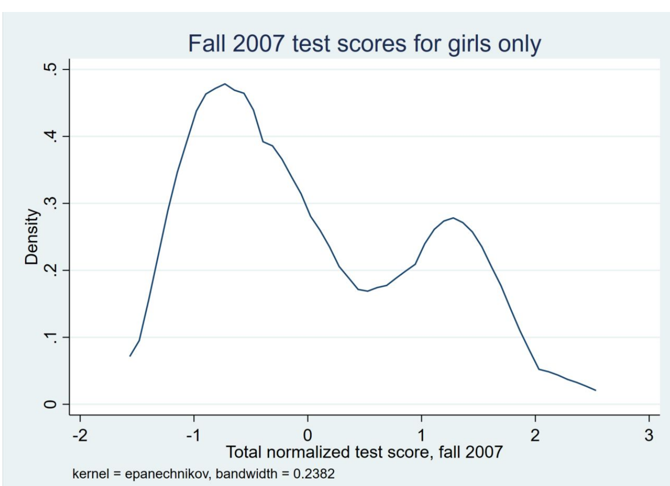
c.
Yes. Since it is a RCT, it satisfies the backdoor criterion and conditional independence assumption therefore, there is no selection bias and we could interpret the difference in test scores as the effect of placing the school.

d.


e.
The t statistic gives a value of -10.5403 which does not fall within the 95% confidence interval of [-1.96, 1.96]. Therefore, we can infer that there is a statistically significant difference in Fall 2007 test scores between the treatment and control and thus, we reject the null hypothesis.

f.
The numbers 0.691 and 0.424 refers to the increase in Fall 2007 test scores without demographic controls when provided treatment (villages with schools) for girls and boys respectively. Furthermore, it could be observed that the difference in the increase in test scores between the girls and boys is greater amongst the girls by (0.691 - 0.424) 26.7 percent. The magnitude of the increase in test scores can be viewed as statistically significant at the 1% level as indicated by *** in the data.

g.
From part e, our estimate of the mean increase in test scores from the treatment group is 0.5798561. Since the treatment group consists of both girls and boys, we would expect our estimate to be the average of the numbers in table 4, column 3 (0.691 and 0.424) as they represent the increase in test scores from girls and boys under the treatment group respectively. Hence, we would expect our estimate to lie between 0.691 and 0.424.

h.


We can observe a bimodal distribution from the “Fall 2007 test scores for girls only” graph. A reason for the observation is that since 373 girls receive the treatment (education) whereas 326 girls did not, we would expect the girls who do receive the treatment to get higher test scores (explaining the peak at around 1.2) and the girls who did not receive the treatment to get lower grades (explaining the peak at around -0.9).

i.

From the t-test, I obtained a difference of 0.7464044 between the girls with and without treatment in their fall 2007 test scores. The standard error obtained from the t-test is 0.0684148 which has a difference of around 0.06 with the value in the table of 0.130.

j.

Yes, this would affect the findings in questions c. and e. The findings in questions c and e would have overestimated the effect of the increase in test scores of children with schools placed in their villages. Since a higher education of the children's parents would tend to lead to better educated children, we can view the parents who are more educated as a backdoor causal path to the academic abilities of their children. Therefore, since there is a backdoor path, the conditional independence assumption will not hold and the experiment would no longer be considered as randomised. Thus, selection bias would be present in the experiment, overestimating the effects of the treatment (schooling) on the test scores.

k.

The t value obtained is -1.1490. Therefore, we fail to reject the null hypothesis at $\alpha = 0.05$ since the absolute value of the t value (1.1490) is less than the t value required to reject the null (1.96). Therefore, the difference in the level of education of the head of household (measured in Fall 2007) is not statistically significant between the treatment and the control. Hence, in our sample, the level of education of the head of household is 0.210431 (3.301911 - 3.09148) greater in the treatment group than the control group, but the difference is not sufficient to conclude a difference at the population level.

l.

There are 881 individuals in group0 (randvar<0.5) and 847 individuals in group1.

```
. // question l, generating variables
. generate randvar = runiform(0,1)

. generate randvar1 = 1 if randvar >= 0.5
(847 missing values generated)

. replace randvar1 = 0 if randvar < 0.5
(847 real changes made)

. count if randvar1 == 1
  881

. count if randvar1 == 0
  847
```

m.

The t value obtained is -0.7702. Therefore, we fail to reject the null hypothesis at $\alpha = 0.05$ since the absolute value of the t value (0.7702) is less than the t value required to reject the null (1.96). Therefore, the difference in the Fall 2007 test scores between the two groups, group-0 and group-1, is not statistically significant. Hence, in our sample, the Fall 2007 test score is 0.0446498 (0.3225308 - 0.277881) greater in group-1 than in group-0, but the difference is not sufficient to conclude a difference at the population level.

Yes, this is what I would expect as since the groups are divided using a random process, it would satisfy the backdoor criterion and conditional independence assumption, preventing any confounders or colliders from affecting the results of the experiment therefore, eliminating any selection bias.

n.

We would expect to find a statistically significant difference in Fall 2017 test scores from ($180 \times 5\%$) 9 students. A significance level of 0.05 indicates a 5% probability of rejecting the null when it is true. The 0.05 significance level also implies a 95% confidence interval where 95% of samples will have a confidence interval containing the true mean of 0. Hence, the remaining 5% of the samples will have confidence intervals which are unable to capture the true mean and these samples would be in the rejection region which are the 9 students.