

8. Change the sample size to $n = 30$ and repeat 7. What do you observe? Explain your observation without providing a formula. (max 60 words)
9. Change the sample size back to $n = 500$ but now change to $\sigma_{x_1x_2} = 0$. Repeat 4 and 7. What do you observe? Explain without providing a formula. (max 100 words)
10. Change the sample size to $n = 30$ and repeat the previous question 9. What do you observe and which estimator would you use? Explain without providing a formula. (max 120 words)
11. Change back to $n = 500$ and $\sigma_{x_1x_2} = 0.5$. However, now change to $\sigma_{x_2z} = 0.1$.
 - (a) Repeat the algorithm 7. What do you observe? Does the value correspond to your expectation? Explain. (max 80 words)
 - (b) Repeat the estimation but for $\sigma_{x_1z} = 0.1$. Compare and explain any differences to the previous question. Do you obtain the expected value? Explain. (max 220 words)

Exercise 3 (40%)

In this exercise, we are going to use data from the study “Bringing Education to Afghan Girls: A Randomized Controlled Trial of Village-Based Schools” by Burde and Linden (2013).

According to Burde and Linden (2013), “primary school participation rates in Afghanistan are very low, particularly for girls. [...] Schools are often far (Sutton 1998), and when available, the lack of separate sanitation facilities, female teachers, and gender-segregated classrooms may also deter girls’ enrollment”.

The authors used a randomized controlled trial to assess the 1-year effect of village-based schools on childrens’ school enrollment and performance in rural northwestern Afghanistan.

The intervention is based on a sample of 31 villages which were grouped into 12 equally sized village groups based on political and cultural similarities (of which one dropped out). The authors randomly assign 5 groups to obtain village-based schools a year before the other groups starting in summer 2007. The final sample has data on 11 village groups, 31 villages and a total of 1,490 primary school-age children. The treatment group received a school whereas the control group did not during this year of the intervention.

The authors surveyed all available households in the fall of 2007 and in spring 2008. The survey collected basic demographic information, enrollment data and test scores of the children and the households. We are going to focus on the enrollment results from the fall 2007 survey. Only use observations observed at this time. Please be aware that in this homework, we do not exactly follow the authors methodologies which might explain differences in your results and the ones in the paper.

Please *individually* download the data from here: <https://doi.org/10.3886/E113861V1>. OpenICPSR is a data repository which hosts for example data from the AEA – in this case the data for this paper. You will have to make an account at OpenICPSR in order to download the data.

Set `seed = 1234`. In order to read the data into R, you can use the *haven* package together with the function `read_dta`. After loading the data, you can extract the labels using the following code (using the *dplyr* package):

```
labels = c()
for (i in colnames(dat)){
  temp = attr(dat[[i]], "label")
```

```

labels[i] = ifelse(is.null(temp), NA,temp)
}
labels = tibble(name = names(labels), label = labels)
View(labels)

```

For your convenience, please find here a summary of the first 10 variables from the raw dataset to make sure you have loaded your data correctly:

```

> summary(dat[,1:10])
 f07_hh_id      s08_heads_child_cnt s08_girls_cnt      s08_age_cnt      s08_duration_village_cnt
Length:1804    Min.      :0.0000      Min.      :0.0000    Min.      : 6.000    Min.      : 0.00
Class :character 1st Qu.:1.0000      1st Qu.:0.0000    1st Qu.: 7.000    1st Qu.:16.00
Mode  :character Median :1.0000      Median :0.0000    Median : 9.000    Median :30.00
              Mean  :0.9301      Mean  :0.4806    Mean  : 8.651    Mean  :30.97
              3rd Qu.:1.0000      3rd Qu.:1.0000    3rd Qu.:10.000   3rd Qu.:40.00
              Max.  :1.0000      Max.  :1.0000    Max.  :11.000   Max.  :80.00
              NA's   :260        NA's   :260      NA's   :261     NA's   :261
s08_age_head_cnt s08_yrs_ed_head_cnt s08_num_ppl_hh_cnt s08_jeribs_cnt s08_num_sheep_cnt
Min.      :14.00    Min.      : 0.000    Min.      : 2.000    Min.      : 0.000    Min.      : 0.000
1st Qu.:35.00      1st Qu.: 0.000      1st Qu.: 7.000      1st Qu.: 0.000      1st Qu.: 0.000
Median :40.00      Median : 3.000      Median : 8.000      Median : 1.000      Median : 5.000
Mean  :40.62      Mean  : 3.558      Mean  : 8.711      Mean  : 1.169      Mean  : 7.322
3rd Qu.:47.00      3rd Qu.: 5.000      3rd Qu.:10.000     3rd Qu.: 2.000     3rd Qu.:10.000
Max.  :80.00      Max.  :20.000      Max.  :33.000      Max.  :10.000      Max.  :150.000
NA's   :261       NA's   :262       NA's   :260       NA's   :260       NA's   :260

```

1. State the research question. Why did the authors randomize the treatment? (max 200 words)
2. Implement the following restrictions in your data:
 - The number of people in the household is smaller than 25
 - No one owns more than 10 jeribs of land
 - No household has more than 60 sheep and goats

How many observations did you exclude? What could be the reason for restricting the sample in this way? (max 150 words)

3. Estimate the effect of the treatment on girls' enrollment in a formal school without *any* additional controls. Write down the model, estimate and interpret the result. Use the traditional standard error. (max 140 words)
4. Run the regression using heteroskedasticity robust standard errors and report their standard errors. Do the results change? (max 65 words)
5. The authors cluster their standard error at the village group level.
 - (a) Explain the rationale behind this and explain the potential consequences if they were not to do that. Would it be possible to randomize at the individual level in this context? (max 180 words)
 - (b) Run the regression using Cluster Standard Errors (Liang and Zeger, 1986). Report your results. What changes? (max 60 words)

- (c) Run the regression using the group average approach to account for the clustering problem. Report your results. What changes? (max 70 words)
 - (d) Run the regression using the clustered bootstrap and 2000 bootstrap replications. Report your results. What changes? (max 70 words)
 - (e) Comment on the suitability and reliability of the standard errors from question 4, 5b, 5c given this setup. Are your results surprising? (max 200 words)
6. You are also interested in the treatment effect for boys. How would you obtain the treatment effect for boys and girls (the latter from question 3) from a single regression? Write this regression down and estimate using a suitable clustering method. Explain how to find the different effects using population expressions, calculate them and interpret. (max 400 words)

References

- BURDE, D. AND L. L. LINDEN (2013): “Bringing Education to Afghan Girls: A Randomized Controlled Trial of Village-Based Schools,” *American Economic Journal: Applied Economics*, 5, 27–40.