Lab Report 4: Regression

Schools in Afghanistan

DUE: November 13 at 5 PM

Background

You work for the United Nations. Given the political and social upheavals in Afghanistan, your boss wants to know about the effects of schooling on children's education. Primary school participation rates in Afghanistan are very low, particularly for girls. In 2007, only 37 percent of primary school-age children attended school. In rural areas, the gender gap in enrollment was 17 percentage points (United Nations Development Programme and The Center for Policy and Human Development at Kabul University 2007).

The Afghan government and donor countries have prioritized the creation of schools to address the problem. 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 (United Nations (UN) 2008, AL-Qudsi 2003, Adele 2008). However, others argue that low demand for education and conservative beliefs may be more important impediments (for example, Adele 2008). Children often perform household chores and help with farming and animal husbandry. Boys and girls perform different tasks, and as a result, girls may face higher opportunity costs for their time (Sutton 1998). Early marriage, the lack of labor force opportunities, wage discrimination, and the fact that girls typically join a husband's household at marriage may all differentially reduce the returns to the education of girls (UN 2008). If these factors are sufficiently strong, they could prevent girls from going to school, regardless of the number of schools available.

Your boss has given you a dataset from an experiment that was conducted in Afghanistan. The researchers a randomized controlled trial in northwest Afghanistan to identify the effect of placing a school within a village. Specifically, within a sample of 31 villages, the researchers randomly assign village-based schools to 13 villages to estimate the 1-year effects of these schools on the enrollment and academic performance (math and language skills) of 1,490 primary school-age children. The remaining villages received schools the subsequent year[1].

Your Task

Your boss wants to understand the relationship between children's test scores (dependent variable) and whether they attended school. She also wants to know if schooling affects male and female students differently. She wants you to run regression analyses and summarize the results in a brief memo. As part of this task, you will:

[1]The description of this lab, as well as the data for this lab, are (in some cases directly) from the following paper: Burde, Dana, and Leigh L. Linden. 2013. Bringing Education to Afghan Girls: A Randomized Controlled Trial of Village-Based Schools. *American Economic Journal: Applied Economics* 5 (3): 27–40. You may find it interesting to read this paper after you complete the lab.

1. Analyze the data using bivariate and multivariate regression.
2. Use and interpret an interaction term.
3. Report your regression results in a professional table using `stargazer()`.
4. Calculate predictions from your linear model(s).

This assignment will walk you through how to do these analyses with space to take notes on the results from each step, but you will submit only the memo summarizing your results, including the results table you will make at the end.

Step 1: Read in the Data

Read in the dataset: `afghan_schools.csv`. The dataset is available on CANVAS in the Lab Reports folder or it can be imported using the following code:

```
data <- read.csv("https://raw.githubusercontent.com/ilaydaonder/POLS209/Lab-Report-4/afghanschools.csv")
```

The unit of analysis is the individual child. These data have many variables coded as follows:

- **testscores**: the score the child earned on a standardized exam (this is your dependent variable)
- **treatment**: whether a child attended school (1 means they did, 0 means they didn't)
- **girl**: coded as 1 if the child is female
- **age**: the age of the child
- **sheep**: the number of sheep the family owns (a measure of socioeconomic status)
- **duration_village**: the number of years the family has lived in the village
- **education_head**: the number of years of education for the head of the household
- **number_ppl_hh**: the number of people living in the household
- **distance_nearest_school**: the distance to the nearest school

Step 2: Bivariate Regression

Run a simple bivariate linear regression to evaluate whether going to school (binary independent variable) affects the child's test scores. Use the `lm()` function and save your model as an object called "Model1." You can view the results using `summary(Model1)`.

What can you conclude about this relationship? Interpret the coefficient on the treatment variable. How much of the variation in test scores are you explaining?

Step 3: Multivariate Regression 1

Now, control for the following variables: age, sheep, duration_village, education_head, number_ppl_hh, distance_nearest_school. Store this model as an object called "Model2."

Did the effect of your independent variable change after controlling for these other factors? Are any of the other factors important predictors of test scores? Has the model fit changed?

Step 4: Interaction Term

Your boss wants to know if schooling affects the test scores of girls differently than boys. In the previous model, you accounted for both things independently, but to determine if any improvements in test scores due to attending school is different among boys and girls, we need to include an interaction term between our treatment variable and girl variable.

Include an interaction term between the treatment variable and the girl variable by re-running the code for Model2 but replacing "treatment + girl" with "treatment*girl." Include all other variables. Store this model as "Model3."

*Step 5: Predictions from Model 3 Now let's use our model to predict test scores for:

1. A boy who did not attend school
2. A girl who did not attend school
3. A boy who did attend school
4. A girl who did attend school

To do so, we can use the `predict()` function. I'll show you how to get the first predicted test score, and you can do the other three.

Remember that to calculate our predictions, we can use the `predict()` function, specifying the model from which we are making predictions and giving hypothetical values for all variables included in the model. If we wanted the prediction for an average boy who did not attend school, we can calculate our prediction by assuming all other variables are at their median value:

```
predict(Model3,
        newdata = data.frame(
          treatment = 0,
          girl = 0,
          age = median(data$age, na.rm = T),
          sheep = median(data$sheep, na.rm = T),
          duration_village = median(data$duration_village, na.rm = T),
          education_head = median(data$education_head, na.rm = T),
          number_ppl_hh = median(data$number_ppl_hh, na.rm = T),
          distance_nearest_school = median(data$distance_nearest_school, na.rm = T)
          )
        )
```

Feel free to copy/paste the code above!). Notice that we told R to plug in the median values for all of the control variables, but we plugged in 0 for the treatment variable (so the hypothetical student did not attend school) and 0 for the girl variable (so the hypothetical student is a boy).

What is the predicted test score for a student of this profile, based on our model?

Now, do the same for the other three types of students: a girl who did not attend school, a boy who did attend school, and a girl who did attend school. Based on these values, does schooling seem to affect girls and boys differently? In other words, is the difference in test scores among girls that went to school and girls that didn't much larger/smaller than the difference in test scores among boys that went to school and those that didn't?

Step 6: Make a Professional Table using Stargazer()

Stargazer is an amazing tool to easily create professional regression tables. Let's use `stargazer()` to make a table that reports the results from all three of our models.

First, we need to install the `stargazer` package. Packages have built-in code to help us do things easier in R. Most of what we have done in class has relied only on standard functions in "base R" but other packages can add to our toolkit.

To install `stargazer`, click the "Packages" tab on the bottom right window in RStudio and then click "Install." In the "Packages" box in the window that pops up, type `stargazer`. Click "Install." R will do its thing... and that's it! You only need to install a new package once.

Now, to tell R we want to access the functions in a package we have installed, we have to load the package. This is easy. Just write the following code and run it:

```
library(stargazer)
```

Now we can use the functions in the `stargazer` package. Feel free to use the help menu to learn more.

The `stargazer()` function in the `stargazer` package is versatile in that we can specify a lot of different arguments, but we will focus on just a couple. At a minimum, we only need to supply the model results we want in our table and the format of the table output. I'll show you how to make a decently pretty table (`type= "text"`) and then a very pretty table (`type= "html"`). The simplest version solely requires the following:

```
stargazer(Model1, Model2, Model3, type= "text")
```

In other words, you just list the models you want in the table, specify "text" as the type, and you get a nice complete results table that you can copy/paste into a memo, paper, or poster! Notice two things though: there is no title, and the variable names could be improved to look more professional. You can either retype them yourself after copy/pasting this into a word document, or you can specify these things within the `stargazer()` command. Using the code below, replace `"Interesting Table Title"` with a title for the table (for example, `"Schools and Test Scores in Afghanistan"`) and all of the `"Variable"` names with better variable labels. For instance, `"Variable8"` is currently `"distance_nearest_school"`, so rename it `"Nearest School"`. Then you have a beautiful table for your memo!

```
stargazer(Model1, Model2, Model3,
          type = "text",
          title = "Interesting␣Table␣Title",
          covariate.labels = c("Variable1",
                               "Variable2",
                               "Variable3",
                               "Variable4",
                               "Variable5",
                               "Variable6",
                               "Variable7",
```

```
                          "Variable8",
                          "Variable9"),
      dep.var.labels = "Test Scores")
```

In case you're interested...

You can make this table even prettier if you change the `"type"` to `"html"` and then at the end add: `out= "AfghanTable.html"`. The table will be exported as an .html file into whatever your working directory is. So if you set your working directory to your POLS 209 folder, the table will pop up as a new file in that folder. For example:

```
stargazer(Model1, Model2, Model3,
      type= "html",
      title= "Interesting Table Title",
      covariate.labels = c("Variable1",
                          "Variable2",
                          "Variable3",
                          "Variable4",
                          "Variable5",
                          "Variable6",
                          "Variable7",
                          "Variable8",
                          "Variable9"),
      dep.var.labels = "Test Scores",
      out = "AfghanTable.html")
```

If you go to your folder, you can click the file and open it with a browser (e.g., Chrome, Firefox, etc.), and then copy/paste it into your document. The table looks even more professional. However, if you'd prefer to stick with the `"text"` version, that's fine too.

Good Work! Now report your results to your boss in a 1-2 page memo (including the table). In your memo:

- Provide your table.
- Summarize the results from the first two models.
- Describe and interpret your predicted values from the interaction term in Model 3.

You do not need to submit your code or any notes you've taken in this document, just the formal memo.