

ECON 379: Program Evaluation for International Development

Empirical Exercise #3

Due 3/4 by 11:30 AM

In this exercise, we will use the same data set as in Exercise 1, `E1-CohenEtAl-data.dta`, which is a subset of the data used in the paper “Price Subsidies, Diagnostic Tests, and Targeting of Malaria Treatment: Evidence from a Randomized Controlled Trial” by Jessica Cohen, Pascaline Dupas, and Simone Schaner, published in the *American Economic Review* in 2015.

In this exercise, we’ll look at the way beliefs about how malaria is transmitted differ across households. We’ll review Stata’s `generate` and `egen` commands, and we’ll also practice applying the concepts covered in lecture and the readings, testing the equality of means across two groups using several different methods.

Create a do file containing the code below, and then use/modify it to answer the questions beginning at the bottom of the page:

```
// LOAD DATA

clear all
set scheme s1mono
set more off
set seed 314159
version 16.1

** change working directory as appropriate to where you want to save
cd "C:\Users\pj\Dropbox\econ379-2021\exercises\E3-false-counterfactuals"

** load the data from the course website
webuse set https://pjakiela.github.io/ECON379/exercises/E1-intro/
webuse E1-CohenEtAl-data.dta

** save the data locally (and then stop and save the do file as well)
save E3-CohenEtAl-raw-data, replace
```

Use the code above to help you answer the following questions:

1. Summarize the `b_h.edu` variable using the `sum` command. What is the median level of education among household heads in the sample?
2. Generate a dummy variable `med.edu` for having at least the median level of education in the sample. What is the mean of this dummy variable?

3. The variable `b_knowledge_correct` is a dummy variable equal to one for if the respondent knows that malaria is transmitted by mosquitos. What proportion of respondents know how malaria is transmitted?
4. What is the mean value of `b_knowledge_correct` among respondents who have at least 6 years of education (ie at least the median level of education)?
5. The `sum` (or `summarize`) command tells you the standard deviation of the mean of the `b_knowledge_correct` variable in the set of observations with `med_edu==1`, and it also tells you the number of observations with `med_edu==1`. Use (only) these two pieces of information to calculate the standard error (not the standard deviation) of the mean of the `b_knowledge_correct` variable in the set of observations with `med_edu==1`. What is this standard error?
6. Now use the `ttest` command to test whether the mean of `b_knowledge_correct` is equal in the group with `med_edu==1` and `med_edu==0`. Confirm that the standard error that you calculated in Question 5 is correct. Does the mean value of `b_knowledge_correct` differ for respondents with above versus below median education? What is the t-statistic associated with this hypothesis test?
7. What is the standard error associated with the difference in means (of the `b_knowledge_correct` variable) between the `med_edu==1` group and the `med_edu==0` group? Calculate this by hand and convince yourself that you could derive the same answer (up to rounding error) as the `ttest` command.
8. Regress `b_knowledge_correct` on the continuous `b_h_edu` variable. What is the OLS coefficient on `b_h_edu`?
9. Generate a variable equal to the mean of `b_h_edu`. What is this mean?
10. Generate a variable equal to the difference between `b_h_edu` and its mean (so, the value of this variable will be different for different observations). Call this variable `ed_diff`. What is the mean of `ed_diff`?
11. Generate a variable `y_times_ed_diff` equal to `b_knowledge_correct` time `ed_diff`. What is the mean of this variable?
12. Generate another variable `ed_diff_square` equal to `ed_diff*ed_diff`. What is the mean of this variable?
13. What is the ratio of the mean of `y_times_ed_diff` to the mean of `ed_diff_square`? (Note that this is the same as the ratio of the `*sum*` of all the values of `y_times_ed_diff` to the `*sum*` of all the values of `ed_diff_square`).
14. This number was the answer to one of the first 12 questions – which one?