ECON 523: Program Evaluation for International Development
In-Class Activity 4

In this exercise, we're going to be replicating specifications from "Does a ban on informal health providers save lives? Evidence from Malawi" by Professor Susan Godlonton and Dr. Edward Okeke. The data set `E5-GodlontonOkeke-data.dta` contains information (from the 2010 Malawi Demographic and Health Survey) on 19,680 live births between July 2005 and September 2010. Each observation represents a birth. Before getting started, you will need to download the data set onto your laptop.

Before you begin, save a new do file containing the Stata code below.

```
clear all
set more off
set scheme s1mono
set seed 123456

cd "C:\Users\pj\Dropbox\ECON-523\topics\4-DD2\stata"
use "data\E4-GodlontonOkeke-data.dta"
```

Extend your do file as you answer the following questions:

1. To implement difference-in-differences, we need: a dummy variable for the post treatment period, a dummy for the treatment group, and an interaction between the two. The `post` variable is already present in the data set. What is the mean of the `post` variable? What fraction of the observations in the data set occur in the post-treatment period?

2. The `time` variable indicates the month and year in which a birth took place. If you type the command `desctime`, you'll see information about how the variable time is formatted. Notice that the `time` variable is formatted in Stata's date format: it is stored as a number, but appears as a month and year when you describe or tabulate it.

3. Use the `tab` command to see how Professor Godlonton and Dr. Okeke define the post-treatment time period in their analysis. What is the first treated month?

4. We need to define an indicator for the treatment group. Professor Godlonton and Dr. Okeke define the treatment group as DHS clusters (i.e. communities) that were at or above the 75^{th} percentile in terms of use of TBAs prior to the ban. Data on use of TBAs comes from the DHS question "Who assisted with the delivery?" Responses have been converted into a set of different variables representing the different types of attendants who might have been present at the birth. Tabulate (using the `tab` command) the `m3g` variable, which indicates whether a woman indicated that a TBA was present at a birth. What pattern of responses do you observe?

5. We want to generate a dummy variable that is equal to one if a TBA was present at a particular birth, equal to zero if a TBA was not present, and equal to missing if a woman did not answer the question about TBAs. There are several different ways to do this in Stata. One is to use the `recode` command (below). This generates a new variable, `tba`, that is the same as the `m3g` variable except that `tba` is equal to missing for all observations where `m3g` is equal to 9. (It is usually better to generate a new variable instead of modifying the variables in your raw data set, because you don't want to make mistakes that you cannot undo.)

```
recode m3g (9=.), gen(tba)
```

6. Confirm that your new variable, `tba`, is a dummy variable. Use the `tab` command to tabulate the observed values of `tba` (the `m` option tells Stata to tabulate the number of missing values in addition to the other values).

7. We want to generate a treatment dummy - an indicator for DHS clusters where use of TBAs was at or above the 75^{th} percentile prior to the ban. How should we do it? The variable `dhsclust` is an ID number for each DHS cluster. How many clusters are there in the data set?

8. We can use the **egen** command to generate a variable equal to the mean of another variable, and we can use **egen** with the `bysort` option to generate a variable equal to the mean within different groups:

```
bysort dhsclust:  egen meantba = mean(tba)
```

9. However, this tells us the mean use of TBAs within a DHS cluster over the entire sample period, but we only want a measure of the mean in the pre-ban period. How can we modify the code above to calculate the level of TBA use prior to the ban?

10. Summarize your `meantba` variable using the `detail` or `d` option after the `sum` command so that you can calculate the 75^{th} percentile of TBA use in the pre-ban period. As we've seen in earlier exercises, you can use the `returnlist` command to see which locals are saved when you run the `summarize` command. Define a local macro `cutoff` equal to the 75^{th} percentile of the variable `meantba`. Then immediately create a new variable `high_exposure` that is an indicator for DHS clusters where the level of TBA use prior to the ban exceeded the cutoff we just calculated.

11. At this point, `meantba` is only non-missing for births (i.e. observations) in the pre-treatment period. Modify the code so that you only define `high_exposure` for births where the `meantba` variable is non-missing. Then we need to replace the missing values of `high_exposure` in the post-treatment period with the correct ones (based on the values in the same cluster in the pre-treatment period), Here are three lines of code that will fix it:

```
bys dhsclust:  egen maxtreat = max(high_exposure)
replace high_exposure = maxtreat if high_exposure==. & post==1 & tba!=.
drop maxtreat
```

12. Tabulate your `high_exposure` variable to make sure that it is only missing for observations with the `tba` variable missing. What is the mean of `high_exposure`?

13. The last variable we need to conduct difference-in-differences analysis is an interaction between our treatment variable, `high_exposure`, and the `post` variable. Generate such a variable. I suggest calling it `highxpost`. You should also label your three variables: `high_exposure`, `post`, and `highxpost`.

14. Now you are ready to run a regression. Regress the `tba` dummy on `high_exposure`, `post`, and `highxpost`. What is the difference-in-differences estimate of the treatment effect of the TBA ban on use of informal birth attendants? How do your results compare to those in Table 5, Panel A, Column 1 of the paper?

15. Read the notes below Table 5. See if you can modify your regression command so that your results are precisely identical to those in the paper.