

Computer assignment 6a – Econometrics I – due October 10, 12.45pm, 2017

THIS IS A GROUP EXERCISE. PLEASE SUBMIT PDF THROUGH BLACKBOARD TIMELY. PART II & III OF THE ASSIGNMENT ONLY. PDF SHOULD SHOW COMMANDS + OUTPUT + (IF ASKED) EXPLANATION, NO DO-FILES. PLEASE INCLUDE YOUR NAMES.

Goal of the assignment

In this computer assignment, you learn how to deal with treatment heterogeneity. This should prepare you to do the same thing on your own. The data are taken from a field experiment conducted in the city of Tilburg during August-November 2015. The data are downloadable from Blackboard (heterogeneity.dta). Goal of the treatment was to stimulate households to separate their waste. It was preceded by a similar treatment (which we analyzed in Computer Assignment 5a). We look at the effect of repeating the treatment, and analyze whether the marginal effect is conditional on the time between the first and the second treatment. Main outcome variable is the weight of residual waste collected per garbage collection route per calendar week (the Department of Sanitation defines its own neighborhoods and calls them garbage collection routes). Each of the routes counts some 1,000 households. Similar to the timing of the first treatment, the timing of the repeated treatment was randomized at the level of 5 or 10 garbage collection routes. The treatment itself ran for two weeks. We have many calendar weeks of data after the end of the second treatment.

I. Preparing the data file for analysis: creating your do file

(a) Open STATA and then open the STATA do file editor. Give a short description in the first line of the do file, for instance:

```
* Computer Assignment 6a 'Treatment heterogeneity', Oct 2017
```


then save the do file under a name of your choosing.

The next line in the do file should tell STATA to open the data from the folder where you saved your data, for instance:

```
use "C:\Users\Nick\Data\heterogeneity.dta", clear
```

and the following command tells STATA the structure of your panel data set:

```
xtset route calendar_week
```

Run these first two commands of your do file (using the appropriate path) as follows: select the line of code and then click on:  (Execute selection). This loads the data into memory.

II. Descriptive statistics

Before you run any regressions, you should first conduct an exploratory data analysis. This gives you the opportunity to check your data for errors, and also to get a feeling for the data (every paper includes a section 'descriptive statistics').

The intervention involves a two-week campaign (`RepTreatmentOngoing` is 1 during the treatment and 0 otherwise). The variable `RepTreatmentCompleted` is 1 for the post-treatment period, and 0 otherwise. The variable `RepTreatment` is a combination of these two. The number of weeks between the first and second treatment is denoted by `TimeElapsed`.

(a) What is the mean value of the main outcome variable in week 34?

(b) What is the min and max value of the time in between treatments?

Let us have a look at the trends in the raw data.

(c) Create a scatter graph of the outcome variable by calendar week, pre and post the treatment, for the 16 routes that have the highest `TimeElapsed`:

```
graph twoway (scatter residual_weight calendar_week if
RepTreatmentOngoing==0&RepTreatmentCompleted==0) (scatter
residual_weight calendar_week if RepTreatmentOngoing==1 |
RepTreatmentCompleted==1) if TimeElapsed>25, by(route)
```

Do the raw data indicate that anything is going on at all as of the date of the repeated treatment?

III. Reporting and interpreting heterogeneity in the treatment effect

Now you are asked to analyze treatment heterogeneity, and to present the results in an easy-to-understand graph.

(a) First, estimate the average treatment effect (ATE), as follows:

```
xtreg residual_weight RepTreatment i.calendar_week, fe i(route)
cluster(route)
```

Remember that the `cluster` option clusters the standard errors by route, which is required to get the right standard errors in a difference-in-differences design. To check: run the same command, but without clustering the standard errors, and report the standard error of the treatment variable only (you do not need to include the full estimation output for the regression without clustering).

To be able to interpret the average treatment effects in percentage terms, execute the following command after running your regression:

```
margins, eydx(RepTreatment)
```

(b) Second, estimate the fully interacted model, as follows:

```
reg residual_weight i.RepTreatment##c.TimeElapsed  
i.calendar_week i.route, cluster(route)
```

In this regression, the terms of the fully interacted model are generated by including two hashes ## between two variables names. The `i .` tells STATA that the first variable is an indicator variable; the `c .` that the second variable is a continuous variable. The post-estimation command that we are going to use next does not work after `xtreg`, which is why we use `reg` instead and include the route-fixed effects in the regression (`i . route`). To verify that we run the same regression, estimate the average marginal effect of the treatment and compare it to your findings under (a):

```
margins, dydx(RepTreatment)
```

Obviously, the marginal effect of the treatment is now dependent on the value of `TimeElapsed`. To see how the marginal effect varies by `TimeElapsed`, run the following command:

```
margins, over(TimeElapsed) dydx(RepTreatment)
```

Do the conditional marginal effects of the treatment have a causal interpretation?

(c) To create a marginal effect plot, run the following command directly after the last `margins` command:

```
marginsplot
```

What does the plot suggest about the marginal effect of the treatment by time since the last treatment has ended? Why do the marginal effects line up so nicely?

(d) Let us now derive marginal effects *separately* for a range of values of `TimeElapsed` – and see how the results differs from above. First, create indicator variables for four ranges of values of `TimeElapsed`:

```
gen quick=(TimeElapsed<=17)  
gen medium1=(TimeElapsed>17&TimeElapsed<=23)  
gen medium2=(TimeElapsed>23&TimeElapsed<=27)  
gen slow=(TimeElapsed>27)
```

Then create the separate interaction terms:

```
gen RepTreatment_quick=RepTreatment*quick  
gen RepTreatment_medium1=RepTreatment*medium1  
gen RepTreatment_medium2=RepTreatment*medium2  
gen RepTreatment_slow=RepTreatment*slow
```

Run the regression with the interaction terms:¹

```
xtreg residual_weight RepTreatment_quick RepTreatment_medium1  
RepTreatment_medium2 RepTreatment_slow i.calendar_week, fe  
i(route) cluster(route)
```

(e) For an easy way to create a marginal effect plot from these results, create a variable that takes on the values of the marginal effects for different values of TimeElapsed:

```
gen coeff=.  
  
replace coeff=_b[RepTreatment_quick] if TimeElapsed==13  
  
replace coeff=_b[RepTreatment_medium1] if TimeElapsed==19  
  
replace coeff=_b[RepTreatment_medium2] if TimeElapsed==25  
  
replace coeff=_b[RepTreatment_slow] if TimeElapsed==29
```

Then plot the estimated coefficients by TimeElapsed (after sorting, this is necessary given the line plot that we use as well):

```
sort coeff  
  
graph twoway (scatter coeff TimeElapsed) (line coeff TimeElapsed)
```

Note that this marginal effect plot looks very different from the one generated under (c). Explain this difference.

Say that you are a policymaker, and you have to decide how much time you should leave between the first treatment and the repeated treatment. How does the above marginal effect plot help you to make this decision?

¹ In the lecture, I argued that you should always include all constitutive terms of an interaction. Why are the indicator variables `quick`, `medium1`, `medium2` and `slow` not included in the above regression? If you are not sure, run the regression again, but now with these terms.