Harris School of Public Policy
PPHA 346 Program Evaluation
Winter 2018

Assignment 3

Due date: Feb. 19, 2018, by 11:59pm via Canvas.

Format: For this assignment, provide **a write-up in pdf or html format** where you answer the questions below. Selectively cut and paste output and be concise in your write-up; excess wordiness will be penalized. Also submit your **code script** that includes commands for your entire analysis.

Data: You will find data eitc_1993 on Canvas.

Non-Stata or -R Languages:
For this assignment, you are recommended to use either **STATA** or **R** because answer keys and script will be in STATA and R. When you get wrong answers, suggested solutions probably won't be helpful if you use other software. Please keep in mind:

- Your grade will be based entirely on whether you got the right numeric answer and whether your written explanations/interpretations are correct. STATA and R users who get wrong numeric answers due to a minor mistake in their code might still get partial credit if their reasoning and interpretation are right, but we might not be able to identify mistakes in code for other languages.
- Regardless of what language you use, you are always required to show your code after written explanations and before outputs, as indicated above.
- **Regardless of what language you use, PLEASE submit both of your write-up and code script on Canvas.**

Late Submissions:
Please turn in your assignment on time. Per syllabus policy, two percentage points of your grade (10 percentage points max) will be discounted due to late submission.

Academic Honesty

You can discuss this assignment with your classmates, but you should not share your code or your write-up. Academic dishonesty will not be tolerated. If you commit plagiarism, your homework will be penalized and it might result in further severe consequences.

Effect of the Earned Income Tax Credit on female employment

Part 1: Unconditional Difference-in-Difference Estimates of the Effect of the 1993 EITC Expansion on Employment of Single Women

1. To familiarize yourself with some basics of the dataset EITC_1993, please answer the following questions. When filing taxes, what is the population that is eligible for EITC in general? What is the population we are studying on in this sample data set? Who are the subjects that benefited from tax credits in this data set? Who are the subjects that didn’t receive EITC? What is the sample period covered by the data? How do you test the treatment effect and what is the dependent variable in the sample dataset?

2. Based on your answers above, define two dummy variables to indicate before/after and treatment/control groups. How many before and after observations for treatment group? For the control group? Verify your answers by reproducing the statistics displayed in Table 1. (**Note**: The EITC went into effect in the year 1994 and only those women with at least one child received EITC)

Table 1
Mean and Total Work by Treatment and Time indicators

Treatment Indicator Time Indicator	EITC Beneficiaries	EITC Non-Beneficiaries	Total
Pre-treatment	4247	3154	7401
Post-treatment	3572	2773	6345
Total	7819	5927	13746

3. Compute data needed in the Difference-in-Difference calculation and fill in the statistics displayed in Table 2.

Table 2
Mean Work by Treatment and Time indicators and Unconditional Difference-in-Differences Estimate of Effect of the Earned Income Tax Credit on Work

Treatment Indicator Time Indicator	Group affected by the policy change (treatment)	Group that is not affected by the policy change (control)	Difference	Difference-in-Difference
Pre-treatment				
Post-treatment				

4. Run a simple regression to estimate the difference-in-difference estimate of effect of Earned Income Tax Credit on Work. You will be able to come very close (4th decimal place) to matching the means, but won't be

able to reproduce them exactly. What is your regression equation (not your code)? Report and interpret the estimated coefficients. Why are your estimates from the regression not exactly the same as what you did manually above? (**Note**: use as few information you need as possible to run this regression)

5. The purpose of the program evaluation is to find a "good" estimate of treatment effect. Given the data that we have available, do you think the Difference-in-Difference estimate you got above is a “good” estimate? What is the problem of this simple regression? Explain your answer.

6. What data or information you would need to include in your analysis in order to improve your Difference-in-Difference estimate?

Part 2: Conditional Difference-in-Difference Estimates of the Effect of the 1993 EITC Expansion on Employment of Single Women

In this section, we will re-estimate this model including demographic characteristics.

1. Calculate the sample means of baseline characteristics for control group which includes single women with no child, and the sample means of baseline characteristics for treatment group, where single women had one or more children. Report your results in a table. Was treatment unrelated to outcomes at baseline? In other words, was allocation of treatment determined by outcome? Why would we want a balanced data? Explain your answer.

2. In order to test Parallel Trend Assumption, create a graph that plots the trend and visually inspect the trend of observations over many time points. Do treatment and control groups have Parallel Trends in outcomes in the absence of treatment? Why is this assumption the most critical among all assumptions and how would your answer affect your estimation of the causal effect? (**Note**: this question might be tricky but it is worth to spend a bit more time on it.)

3. Generate a new variable with earnings conditional on working and calculate its means by group as well. Include this new interaction term in your regression. Discuss your findings. (**Note**: missing values for variable work is considered as unemployed)

4. In order to re-estimate the treatment effect with a model including demographic characteristics, you need to add some additional baseline demographic variables in your regression. What are those additional variables you choose to add in your regression? You can also include polynomial terms if necessary. Write down the model in your write up (*not code of regression model). For example:

$$Y_i = \alpha + \beta T_i + \gamma t_i + \delta (T_i * t_i) + \omega X_i + \varepsilon_i$$

(**Note**: there is no standard solution for this question. Choose the variables that you consider important and be able to explain why these variables are important)

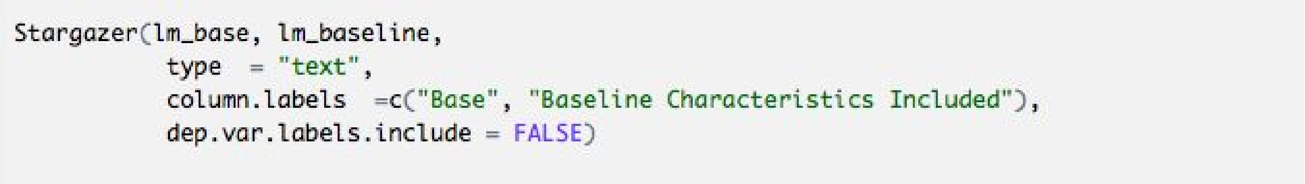
5. Run the regression in the software. Report and interpret each estimated coefficient. What is the difference-in-difference estimate in this regression? How do you interpret it? Did controlling for demographic characteristics change the difference-in-difference estimate in an expected manner?

6. Add the impact of state unemployment rate varied by treatment and control indicator (single women with no kid and single women with no less than 1 kid) in your regression. What is your regression equation? Run the regression in your software. Is there any improvement in your difference-in-difference estimate? Explain your answer. (**Hint**: you need to construct a new interaction term between unemployment rate and treatment dummy)

7. In your model, other than all of the terms you already have in your equation, add the impact of treatment to vary by single women with 1 or more than 1 children. Now you should have quite a few terms in your regression. What is your regression equation now?

8. Run the regression of question 7 in the software. In order to do so, first you will need to generate a new treatment dummy variable, which equals to 0 if the observation had exactly 1 child and to 1 if more than 1 children. Then you will need to construct a new interaction term by interacting the new dummy with before-after indicator. Finally, add the new treatment term and the new interaction term in your regression. What is your difference-in-difference estimates now? Is there any improvement in your estimate? Explain your answer.

9. Display the regression coefficients of all regression you have run above. Using the summary(.) function is standard. As a more elegant solution, stargazer package is recommended to produce nicely formatted output. It automatically displays the estimates for the same variable in the same row. This vastly improves the readability and comparability of the estimates across models. Here’s an example of how to use stargazer:



```
Stargazer(lm_base, lm_baseline,
          type  = "text",
          column.labels  =c("Base", "Baseline Characteristics Included"),
          dep.var.labels.include = FALSE)
```

We want to estimate a sequence of models with an increasing number of controls and compare the stability of key results across these models. Was it necessary to control for demographic characteristics, effect of earnings conditional on working, impact of state unemployment rate varied by treatment dummy, as well as the new treatment varying by single women with 1 or more than 1 children? Comment on the estimation results and summarize your findings after evaluating the effect of EITC program on female employment. Is there any way to improve your estimate if you can request more data? (**Note**: stargazer is not required as long as you can display your coefficient estimates properly)