ECON7310: Elements of Econometrics

Research Project 2

Fu Ouyang

May 15, 2023

Instruction

Answer all questions following a similar format of the answers to your tutorial questions. When you use R to conduct empirical analysis, you should show your R script(s) and outputs (e.g., screenshots for commands, tables, and figures, etc.). You will lose *2 points* whenever you fail to provide R commands and outputs. When you are asked to explain or discuss something, your response should be brief and compact. To facilitate our grading work, please clearly label all your answers. You should upload your research report (in PDF or Word format) via the "Turnitin" submission link (in the "Research Project 1" folder under "Assessment") by 11:59 AM on the due date May 22, 2023. Do not hand in a hard copy. You are allowed to work on this assignment in groups; that is, you can discuss how to answer these questions with your group members. However, this is *not* a group assignment, which means that you must answer all the questions in your own words and submit your report separately. The marking system will check the similarity, and UQ's student integrity and misconduct policies on plagiarism apply.

Panel Data (35 points)

Background

DiTella and Schargrodsky (2004)[1] examine how the street presence of police officers reduces car theft. Rational crime models predict that the presence of an observable police force will reduce crime rates (at least locally) due to deterrence. The causal effect is difficult to measure, however, as police forces are not allocated exogenously but rather are allocated in anticipation of need (i.e., reverse causality). The innovation in DiTella and Schargrodsky (2004) was to use the police response to a terrorist attack as an exogenous variation.[2]

In July 1994, there was a horrific terrorist attack on the main Jewish center in Buenos Aires, Argentina. Within two weeks, the federal government provided police protection to all Jewish and Muslim buildings in the country. DiTella and Schargrodsky (2004) hypothesized that their presence, while allocated to deter a terror or reprisal attack, would also deter other street crimes, such as automobile theft. The authors collected detailed information on car thefts in selected neighborhoods of Buenos Aires from April-December 1994, resulting in a panel for 876 city blocks. They hypothesized that the terrorist attack and the government's response were exogenous to auto thievery, thus a valid treatment. They postulated that the deterrence effect would be strongest for any city block which contained a Jewish institution (and thus police protection). Potential car thieves would be deterred from a burglary due to the threat of being caught. The deterrence effect was expected to weaken as the distance from the protected

[1] Di Tella, R. and Schargrodsky, E., 2004. Do police reduce crime? Estimates using the allocation of police forces after a terrorist attack. American Economic Review, 94(1), pp.115-133.

[2] DiTella and Schargrodsky (2004) is a very nice example for estimating causal effects with the difference-in-differences approach. We can obtain the same empirical results by using two-way fixed effects regressions.

sites increased. Their sample has 37 blocks with Jewish institutions (the treatment sample) and 839 blocks without an institution (the control sample).

Questions

Use the data set `DS2004.csv` to estimate the following regression model:

$$\texttt{thefts}_{it} = \beta_0 + \beta_1 D_{it} + u_{it}, \tag{1}$$

where the subscripts i and t label city blocks and months respectively, $D_{it} = \texttt{sameblock}_i \times \texttt{post-attack}_t$, and $\texttt{post-attack}_t$ is a binary variable indicating months in the data after the terrorist attack; i.e., $\texttt{post-attack}_t = 1$ if $\texttt{month} \geq 8$, and 0, otherwise. See the definitions for variables `thefts` and `samebloack` in the data description. For all the questions below, exclude observations for July.

(a) (3 points) Is this a balanced panel? Hint: Use the `is.pbalanced()` function in the `plm` package.

(b) (7 points) Estimate β_1 in (1) with OLS and compute the cluster-robust standard error (SE) (3 points). Why is it important to use clustered standard errors for the regression (2 points)? Do the results change if you just use heteroskedasticity-robust standard errors for cross-section model (2 points)?

(c) (5 points) Control time (month) fixed effects in model (1) and test if there are significant time fixed effects.

(d) (10 points) Extend model (1) to estimate the deterrence effect of the street presence of police officers using a difference-in-differences (DID) approach and compute the cluster-robust SE (5 points). Give your estimation result a causal interpretation (5 points).

(e) (5 points) Add time (month) fixed effects δ_t to (1) and write $u_{it} = \alpha_i + e_{it}$. Then model (1) extends to

$$\texttt{thefts}_{it} = \beta_0 + \beta_1 D_{it} + \alpha_i + \delta_t + e_{it}. \tag{2}$$

Treat α_i in (2) as entity (block) fixed effects, estimate β_1 with fixed effects (FE) method, and compute the cluster-robust SE.

(f) (5 points) The data has the dummy variable `oneblock` which indicates if the city block is one block away from a protected institution. Extend the FE regression in (c) by including one additional treatment variable–`oneblock` interacted with the post-attack dummy. Use this model to test if the deterrence effect extends beyond the same block?

Binary Choice Models (30 points)

You want to study female labor force participation using a sample of 872 women from Switzerland (`swiss.csv`). The dependent variable is `participation` (=1 if in labor force), which you regress on all further variables plus age squared; i.e., on `income`, `education` (years of schooling), `age`, $\texttt{age}^2$, numbers of younger and older children (`youngkids` and `oldkids`), and on the factor `foreign`, which indicates citizenship (=1 if not Swiss).

(a) (10 points) Run this regression using a linear probability model (LPM) and report the regression results (4 points). Test if `age` is a statistically significant determinant of female labor force participation (3 points). Is there evidence of a nonlinear effect of age on the probability of being employed (3 points)?

(b) (10 points) Repeat (a) using probit and logit regression models and report your results.[3]

(c) (5 points) Use the probit model to compute the predicted probability of being in the labor force for a Swiss female (A) with median income and age of the sample, 12 years of schooling, one young kid, and no old kid.

(d) (5 points) Keeping all other factors the same as in (c), consider another Swiss female (B) with the 75th percentile age of the sample. Compute the difference in the predicted probabilities of being in the labor force between A and B.

IV and TSLS (35 points)

Use the following regression model and dataset `cigbwght.csv` to estimate the effects of several variables, including cigarette smoking, on the weight of newborns:

$$\log(\texttt{bwght}) = \beta_0 + \beta_1 \texttt{male} + \beta_2 \texttt{parity} + \beta_3 \log(\texttt{faminc}) + \beta_4 \texttt{smoke} + u, \tag{3}$$

where `male` is a dummy variable equal to 1 if the child is male; `parity` is the birth order of this child; `faminc` is family income (in $1000); and `smoke` is a dummy variable equal to 1 if the mother smoked during pregnancy.

(a) (7 points) Estimate regression equation (3) using OLS and report regression results (3 points). Interpret the estimated coefficient on `smoke` (2 points) and test if the population coefficient β_4 is zero at the 1% significance level (2 points).

(b) (8 points) Some studies suggest that smoking during pregnancy may have different impacts on male and female babies. Modify the specification of the regression model (3) and test this hypothesis (4 points). In your modified model, does `smoke` still has significant (at 5% level) effects on the weight of newborns (2 points)? Explain your answer using test results (2 points). Hint: You don't need to report regression results here, but writing out your modified regression model may be helpful.

(c) (6 points) One of your classmates expresses her concern about the validity of your regression analysis and argues that there may be unobserved health factors correlated with smoking behavior that affect infant birth weight. For example, women who smoke during pregnancy may, on average, drink more coffee or alcohol, or eat less nutritious meals. If this is the case, do you think the OLS estimates you obtained in (a) are unbiased (consistent) (2 points)? Explain your answer (2 points). Is this a threat to your regression analysis's internal or external validity (2 points)?

(d) (4 points) You classmate then propose to use cigarette tax (`cigtax`) in each woman's state of residence as an instrumental variable (IV) for `smoke` and run a two-stage least squares (TSLS) regression. Take her suggestion and report your TSLS regression results.

(e) (10 points) Are coefficients of model (3) exactly identified, overidentified, or underidentified (2 points)? Does this TSLS regression suffer from the weak IV problem (2 points)? Why or why not (2 points)? Is it possible to test the exogeneity of `cigtax` as an IV for `smoke` (2 points)? Explain your answer (2 points).

[3] You don't need to compute robust SE here.