ECO 475–Homework 2
University of Toronto
Due: 05 April, 2024
Late assignments will not be accepted
For full credit, please show your work

# 1 Theoretical Problems

1. Dynamic Panel Data. Consider the model

$$y_{it} = \rho y_{it-1} + x_{it}\beta + a_i + \varepsilon_{it} \tag{1}$$

where $a_i$ is the fixed effect; $\rho \neq 0$ and $|\rho| < 1$; and $\{\varepsilon_{it}\}$ is i.i.d. with $E(\varepsilon_{it}) = 0$, $Var(\varepsilon_{it}) = \sigma_\varepsilon^2$, and $Cov(\varepsilon_{it}, a_i) = 0$. Assume (i) $Cov(y_{it-s}, \varepsilon_{it}) = 0$, for all $t$ and $s \geq 1$; (ii) $Cov(x_{it-s}, \varepsilon_{it}) = 0$, for all $t$ and $s$ (i.e., $x_{it}$ is strictly exogenous with respect to $\varepsilon_{it}$); and (iii) $Cov(x_{it}, a_i) \neq 0$, for all $t$.

   (a) Take the first-difference of (1) to eliminate the fixed effects. Explain why the first-difference estimator is not consistent.

   *Hint: remember that we need all the explanatory variables to be uncorrelated with the unobservables of the equation being estimated.*

   (b) Explain *intuitively* why the lagged dependent variable $y_{it-2}$ can be a good instrumental variable for the regressor $\Delta y_{it-1}$.

   *Hint: There is no need to use math to answer this question. Please just focus on the regressor $\Delta y_{it-1}$ and the instrument $y_{it-2}$. No need to find instruments for $\Delta x_{it}$, as it is already assumed to be exogenous (by the strict exogeneity assumption) – putting differently, $\Delta x_{it}$ is an instrument for itself.*

# 2 Computer Problems

1. **Fixed Effects**. Fryer and Levitt (2004, 2006) use the Early Childhood Longitudinal Study of 1998 (ECLS-1998) to examine racial achievement gaps among elementary school students. In this exercise, we will be extending their results to a more recent cohort of the data, ECLS-2011 (see https://nces.ed.gov/ecls/kindergarten2011.asp). ECLS2011Subset.dta contains a subset of this data relevant for this problem.

   (a) Using at most 2 sentences, briefly describe the data (i.e., the ECLS2011Subset.dta provided for this problem) in your own words.

(b) Standardize all math scores in each period to have mean 0 and standard deviation 1. Then, estimate the racial achievement gaps in Fall Kindergarten using the following regression

$$A_{ig} = \beta_0 + \beta_1 Black_{ig} + \beta_2 Hispanic_i + \beta_3 Asian_{ig} + \beta_4 Other_{ig} + \epsilon_{ig}, \quad (2)$$

where $A_{ig}$ is the standardized math test score for student $i$ in grade $g$. The first four variables are race indicators for Black, Hispanic, Asian, and other (native Americans, mixed race, and Hawaiians) students respectively with White students being the omitted category. Briefly interpret the results. (*Hint: use hetersokedastic-robust standard errors.*)

(c) Would it be appropriate to include either student or parent fixed effects in the above regression model?

(d) Fryer and Levitt (2004, 2006) compute the adjusted Black-White achievement gap by controlling for socioeconomic status and other student characteristics. Re-estimate the above model and control for the socioeconomic index, age at kindergarten entry, speaking non-english at home, and mother's age at birth. Briefly discuss how the Black-White and Hispanic-White achievement gaps change. (*Hint: use hetersokedastic-robust standard errors.*)

(e) Fryer and Levitt (2004, 2006) account for school quality by including school fixed effects into the regression model. Now estimate the model including school fixed effects. Briefly interpret the results (*Hint: use encode in Stata to convert school ID to numeric. Also, cluster standard errors at the school level.*)

(f) Consider an extreme scenario in which schools were racially segregated such that all white students went to schools with only white children. Would it be appropriate to include school fixed effects when estimating racial achievement gaps?

2. **Difference-in-differences**. Abadie, Diamond, and Hainmueller (2010) examine the effects of Proposition 99, a large-scale tobacco control program that California implemented in 1988. CigaretteSale.dta contains the data the authors used for this policy evaluation. We will replicate some of the authors findings in this problem.

(a) Given the policy was implemented in California in 1988. Describe why the structure of the data is appropriate, in principle, to use a difference-in-differences analysis. (*Hint: Discuss pre-post and control-treatment variation available in the data.*)

(b) We will replicate Figure 1 in Abadie, Diamond, and Hainmueller (2010). Plot the cigarette sales for California and the average of the remaining states over time, and include a vertical line for the introduction of the policy.

(c) What is the key underlying assumption for the difference-in-differences apporach? Describe using the above plot whether it seems plausible in this setting.

(d) The authors construct a comparable control group to California by appropriately weighting the remaining states. They only assign positive weight to 5 states as shown in their Table 2. Using these weights, construct a 'synthetic California'. Then, plot the cigarette sales for California and Synthetic California to replicate Figure 2.

*Hint: The weights are: Colorado, 0.164; Connecticut, 0.069; Montana, 0.199; Nevada, 0.234; and Utah, 0.334.*

(e) Why is constructing the synthetic California useful to carry out a differences-in-differences in this context? (*Hint: refer back to your answer in item c of this question.*)

(f) Using California and synthetic California, we will cary out a basic difference-in-difference estimation using the following regression:

$$CigSales_{st} = \beta_0 + \beta_1 California_s + \beta_2 Post1988_t + \beta_3 Post1988_t \times California_s + U_{st},$$

where $s$ indexes state, $t$ indexes year, $California_s$ is dummy for California, and $Post1988_t$ is dummy for post-treatment. Interpret your results. (*Hint: Cluster standard errors at the state-level.*)

3. **Tobit**. This problem will investigate the determinants of smoking. smoke.dta includes data on smokers and non-smokers.

(a) Plot and Describe the distribution for the number of cigarettes per day.

(b) How many smokers and non-smokers are in the data?

(c) Estimate the linear model using OLS:

$$cigs_i = \beta_0 + \beta_1 educ_i + \beta_2 income_i + \beta_3 cigpric_i + \beta_4 age_i + \epsilon_i$$

where $cigs_i$ is the number of cigarettes a person smokes per day, and $cigpric_i$ is the price paid (we are ignoring endogeneity of prices here). What is the impact of education on smoking? Describe at least one limitation of estimating the linear regression model with this type of data.

(d) Why might a Tobit model be appropriate in this context?

(e) Estimate a Tobit model and compute the average marginal effect of years of education on the amount of smoker. Compute also the impact of education on both the extensive margin (the probability of smoking) and on the intensive margin (the expected number of cigarettes given that already smoke). Briefly interpret your results.

(f) Estimate a Probit model where the endogenous variable is a binary indicator of whether the person smokes or not, and the exogenos regressors are the same as in the previous items. Compute the average marginal effect of years of education on the probability of smoking. Is your estimated impact similar to the average impact on extensive margin obtained in item e?

Provide your do file and log file as part of your submission.