Econometrics I (Second Half)
Problem Set 2

Instructions (Read Carefully)

Please submit all of the files related to your solution to this problem set via email to me (andrea.flores@fgv.br) and your TA Taric (tariclatif@gmail.com) by **Thursday, December 14th at 11:59pm**.

Your solution to this problem set should consist of (1) a pdf with your responses to the items in each of the questions as if it was a report (for items that require coding with no discussion, include a reference to a particular part of your code), and (2) the files containing the code used in each question that requires estimation. **Submit all of the relevant files in a zip folder.**

Read each question carefully – make sure you address all points raised. Go as far as you can. Even if you don't manage to complete all items in a question, partial credit will be applied generously if you clearly describe the issues you faced in approaching that particular question and intuitively explain how you think these issues could be addressed.

Good Luck!

Question 1: Event Study

In this problem, you are going to quantify the dynamic impact of COVID-19 on married women's quarterly labor market rates. Use the data contained in the csv file `enoe_married_female`. The file includes the variables *newid* which captures a women's identifier in the survey, *time* which captures the quarter of survey, *eda* capturing age, $dent2 - dent32$ denoting state-specific dummy variables, $dchild2\_12$ denoting an indicator for the presence of children in the household younger than 12, *edu* which contains information on educational attainment, *inact* being an indicator of inactivity, *unemp* an indicator of unemployment, $formal\_new$ an indicator of formal employment and $informal\_new$ an indicator of informal employment.

$$Y_{ist} = \sum_{j \in [-3,-2] \cup [0,7]} \alpha_j D^j_{it} + \boldsymbol{\beta} \mathbf{X}_{it} + \eta_s + \epsilon_{ist} \qquad (1)$$

where X_{it} includes a constant, $age, edu, edusq, dchild2\_12$. η_s denotes state-specific fixed effects.

(a) Letting $time = 4$ denote the quarter capturing the onset of COVID-19, define an event-time variable centered around this quarter so that $event = time - 4$. To check that this has been defined correctly, compute and report the mean of the variable *event* by *time*. Before implementing the event study design in 1, generate a set of dummies specific to each event time (for example, $Dm3 = 1$ if $event = -3$ and 0 otherwise, $D1 = 1$ if $event = 1$ and 0 otherwise, and so on).

(b) Using the specification described in 1, run an event study of

- Unemployment and inactivity on pandemic quarters
- Formal employment and informal employment on pandemic quarters

As a normalization, leave out the dummy variable corresponding to the quarter just before the onset of COVID-19. Make a plot of the coefficients associated with the event dummies with confidence intervals. Interpret your results

(c) Notice that the specification described in 1 does not include quarter-specific fixed effects. Explain why adding quarter-specific fixed effects is not necessary in this application.

Question 2: Policy Evaluation

Throughout this question, you will refer to the csv file `progresa.csv`. Notice that the dataset is formatted as a wide panel. The outcome variable of interest is y which captures school enrollment status of children aged 6-16. In the dataset, you can find the variable called *cut*, which captures the cutoff values c (it is okay to see variation in the values of these cutoffs since these vary by municipality). *yycali_97* contains the values of the wealth index used by the *Progresa* administration to assess eligibility such that households with *yycali_97* above the cutoff were assigned to the control group and households such that *yycali_97* are at or below the corresponding cutoff are treated (i.e. receive the cash transfer) [thus, *yycali_97* would be denoted as the running variable within a RD approach].

(a) Suppose that you trust that randomization was properly implemented by the program administration such that you can expect a simple comparison of outcomes after the receipt of the *Progresa* cash transfer would suffice to identify the effect of the program on children's school enrollment. For this, estimate the following:

$$s_i = \alpha + \beta D_i + \epsilon_i$$

where D_i denotes the treatment status of child in household i. Report and interpret your results. Do your results change once you add controls relating the characteristics of children at baseline? Explain.

(b) You now suspect that there might be some issues with the randomization implemented by the program and that potential selection of gains might not be fully controlled for in the specification estimated in **(a)**. Thus, estimate the following exploiting the fact that you observe outcomes for children both before and after the rollout of *Progresa* to capture the effect of the program on children's school attendance:

$$s_{i,t} = \alpha + \beta_1 D_i + \beta_2 Post_t + \beta^{DID}(Post_t \times D_i) + \epsilon_{i,t} \qquad (2)$$

For this, report and interpret the results from implementing the specification described in 2. Do your results change once you add controls relating the characteristics of children at baseline? Explain.

(c) Re-do part **(b)** using a Matching DID estimator by estimating equation (2) on a sample in which the non-treated households will be re-weighted to assign higher weights to those non-treated households that are more observably similar to the treated households. To prepare the matched sample, implement a propensity-score based matching procedure that uses the Epanechnikov kernel to match treated and non-treated units.

(d) We now want to exploit our knowledge on non-parametric methods to estimate β^{RDD} using kernel-based local linear regression – letting $X_i = yycali\_97_i$ and using children's school enrollment after the rollout of *Progresa* as the outcome variable ($Y_i = y\_post_i$) – implement the following steps:

1. Define $\tilde{X}_i = \begin{bmatrix} 1 \\ X_i - c \end{bmatrix}$
2. For observations such that $D_i = 1$, compute

$$\widehat{\beta}_1 = \left(\sum_{i=1}^{n} K\left(\frac{X_i - c}{h}\right) \tilde{X}_i \tilde{X}_i' \mathbb{1}\{X_i - c \leq 0\} \right)^{-1} \left(\sum_{i=1}^{n} K\left(\frac{X_i - c}{h}\right) \tilde{X}_i Y_i \mathbb{1}\{X_i - c \leq 0\} \right)$$

3. For observations such that $D_i = 0$, compute

$$\widehat{\beta}_0 = \left(\sum_{i=1}^{n} K\left(\frac{X_i - c}{h}\right) \tilde{X}_i \tilde{X}_i' (1 - \mathbb{1}\{X_i - c \leq 0\}) \right)^{-1} \left(\sum_{i=1}^{n} K\left(\frac{X_i - c}{h}\right) \tilde{X}_i Y_i (1 - \mathbb{1}\{X_i - c \leq 0\}) \right)$$

4. Let $\widehat{\beta}^{RDD}$ be the first element of $\widehat{\beta}_1 - \widehat{\beta}_0$.

Use the Epanechnikov kernel function. Compute the bandwidth using Silverman's plug-in estimator for the Epanechnikov kernel (check lecture slides). Report your results in a table. Test the sensitivity of your results for β^{RDD} to (i) an increase in h and (ii) a decrease in h. Interpret your findings **and** compare with your estimates of β obtained in part **(a)**, of β^{DID} obtained in part **(b)**, and of β^{MDID} obtained in part **(c)**.