EC338: Assignment 1

General Information

- The assignment has 5 sections. Each section carries a different weight, totalling 100.
- Do your best to answer all questions in each section.
- Submit two files: (1) a **pdf** containing your answers; (2) a do-file (or R-script) containing your code. I am happy for R users to include a single R-markdown file that contains both answers and code.
- Mathematical proofs may be hand-written, but must be legible. If you write the proof by hand, take a photo and past it in the document with your solutions. Of course, typed proofs are more likely to be legible.
- When asked to produce a figure or table of results, please include the output in the same document as your answers.
- For parts C, D, and E, I strongly recommend using either STATA or R.

DUE: Thursday, 17 November 2022, 2pm

Section A (15 marks) - Proof of Variance

In the lectures we d[illegible] ple difference-in-means estimator ($\hat{\tau}$) is an unbiased estimator of the τ^{fs}_{AT} [illegible] ment is assigned according to a completely random experiment. In addi[illegible] ed how the OLS estimator from a simple linear regression (of outcom[illegible] status) gave us the same estimator and is therefore unbias too,

$$[illegible] = \bar{Y}_t^{obs} - \bar{Y}_c^{obs} = \hat{\tau}$$

Assuming *homogeneous treatment effects*, we said that,

$$\hat{V}^{const} = s^2 \cdot \left(\frac{1}{N_c} + \frac{1}{N_t} \right)$$

was the most precise estimate of $Var(\hat{\tau})$. In addition, we know that with under *homoskedasticity* the estimator for $Var(\hat{\beta}^{OLS})$ is given by,

$$\hat{V}^{homosk} = \frac{\hat{\sigma}^2_{\varepsilon}}{\sum_{i=1}^{N}(W_i - \bar{W})^2} == \frac{\frac{1}{N-2}\sum_{i=1}^{N}\hat{\varepsilon}_i^2}{\sum_{i=1}^{N}(W_i - \bar{W})^2}$$

QUESTIONS:

1. Show that,

$$\hat{V}^{const} = \hat{V}^{homosk}$$

2. Next, show that,

$$\hat{V}^{neyman} \approx \hat{V}^{hetero}$$

   That is, show that,

$$\hat{V}^{hetero} = \frac{\sum_{i=1}^{N}\hat{\varepsilon}_i^2 \cdot (W_i - \bar{W})^2}{\left(\sum_{i=1}^{N}(W_i - \bar{W})^2\right)^2} = \frac{\tilde{s}_c^2}{N_c} + \frac{\tilde{s}_t^2}{N_t}$$

   where,

$$\tilde{s}_c^2 = \frac{1}{N_c}\sum_{i:W_i=0}^{N}(Y_i^{obs} - \bar{Y}_c^{obs})^2$$

$$\text{and} \quad \tilde{s}_t^2 = \frac{1}{N_t}\sum_{i:W_i=1}^{N}(Y_i^{obs} - \bar{Y}_t^{obs})^2$$

3. Discuss formally the relationship between homoskedasticity (heteroskedasticity) and homogeneous (heterogeneous) treatment effects.

Note: For part 1, you should use the definition of s^2 in the lecture notes. For part 2, a formal discussion … definition of the error term in the linear regression model, with and …us treatment effects.

Hint: First show … the lecture notes,

$$\ldots \bar{W})^2 = N\bar{W}(1-\bar{W})$$

Section B (10 marks) - Power Calculation

The following quote [illegible] BER paper by Gennetian et al. (2022, p.8) included with the assignment [illegible]

> "Striking a [illegible] statistical power and project costs, 40% of the recruited sam[illegible] was randomized to receive \$333 monthly cash gifts and 60% [illegible] monthly gifts. With an enrolled sample of n=1,000 mother-infant dyads, and accounting for a predicted 20% attrition over longer-term follow-ups, the anticipated sample size of 800 dyads during subsequent waves of data collection is estimated to provide 80% power to detect a .207 standard deviation impact at p<.05 in a two-tailed test on cognitive functioning and family process outcomes."

QUESTION:

1. Demonstrate how the authors arrived at this power calculation: "80% power to detect a .207 standard deviation impact at p<.05". That is, demonstrate that the power of the relevant test is 80% for under the alternative hypothesis that the treatment effect is 0.207 of a standard deviation.

> Note: Please treat this question like a proof where you show your working. **You should be able to show that power is $\gtrsim$82% using either a standard normal or t-distribution approximation of the test-statistic distribution.** This can be done without any additional information from the paper. Demonstrating this is sufficient for full marks. You need to assume that attrition is independent of treatment status. The relevant null hypothesis is one of no effect; i.e. $H_0 : \tau = 0$.

Section C (25 marks) - Simulation Exercise

We have discussed [illegible] estimating treatment effects under the Conditional Independence Assum[illegible]tional data. In this section you will design and execute a simulation th[illegible] compare of the mean and variance of five potential estimators.

You should be able t[illegible]nar 1's STATA do-file (or R-script). However, the set up has been adapted in a number of ways. Please follow the steps carefully and report back any ambiguities.

- Set the number of observations to 1000, as observational datasets tend to be larger.
- Generate the data according to the data generating process

$$Y_i(0) = \gamma_0 + \gamma_1 age_i + \gamma_2 female_i + \gamma_3 age_i \cdot female_i + \varepsilon_i$$

  where,

$$[\gamma_0, \gamma_1, \gamma_2, \gamma_3] = [1.2, 0.015, -0.02, -0.01]$$

  Let,

  - Generate $\varepsilon_i \sim N(0, 0.55^2)$
  - Generate age_i as a random *integer* that is uniformly distributed between $[20, 65]$.
  - Generate $female_i$ as a dummy variable where (female) labour force participation declines rapidly during years of childbirth,

$$female_i \sim B(1, \rho(age_i)) \quad \text{where} \quad \rho(age_i) = 0.5 - 0.25 \cdot \frac{ln(age_i - 19)}{ln(46)}$$

- Allow for heterogeneous treatment effects by age,

$$\tau_i(age_i) \sim N(\mu(age_i), 0.01) \quad \text{where} \quad \mu(age_i) = 0.02 + 0.06 \cdot \mathbf{1}\{age_i > 43\}$$

  Since age is uniformly distributed, $\tau_{ATE} \approxeq 0.05$.

- Simulate unconfoundedness **on age**, alone. Assign treatment status in such a way that the average level of treatment increases with age, but is independent of $Y_i(0)$ conditional on age. Use the binomial distribution where the probability of success depends on age in the following way,

$$W_i \sim B(1, \rho(age_i)) \quad \text{where} \quad \rho(age_i) = 0.25 + 0.5 \cdot \frac{ln(age_i - 19)}{ln(46)}$$

  The probability of treatment should be between $[0.25, 0.75]$.

- Estimate the following five models 1,000 times. Report the mean and standard deviation of the simulated samples of $\hat{\beta}_1$. Provide a plot of the kernel density distribution of each estimator.

- **Model 1:** Estimate a linear regression model <u>without</u> covariates,

$$Y_i^{obs} = \beta_{01} + \beta_{11} D_i + \upsilon_i$$

- **Model [illegible]** [illegible] regression model that matches the CEF of Y_i^{obs},

$$Y_i^{o[illegible]} [illegible] \gamma_{12} age_i + \gamma_{22} female_i + \gamma_{32} age_i \cdot female_i + \varepsilon_i$$

- **Model [illegible]** [illegible]ted linear regression model,

$$Y_i^{obs} = \beta_{13} D_i + \sum_{j=20}^{65} \gamma_{j3} \mathbf{1}\{age_i = j\} + \epsilon_i$$

- **Model 4:** Estimate Model 1, applying inverse probability weights,

$$Y_i^{obs} = \beta_{04} + \beta_{14} D_i + \upsilon_i$$

where the estimated weights are based on the estimated propensity scores,

$$\lambda_i = \frac{1}{\hat{e}(X_i)^{D_i} \cdot (1 - \hat{e}(X_i))^{1-D_i}}$$

derived from a logit model,

$$e(X_i) = Pr(D_i = 1 | age_i) = \Lambda\left(\psi_{04} + \psi_{14} age_i\right)$$

- **Model 5:** Repeat model 4, but use a saturated logit model,

$$e(X_i) = Pr(D_i = 1 | age_i) = \Lambda\left(\sum_{j=20}^{65} \psi_{j5} \mathbf{1}\{age_i = j\}\right)$$

DISCUSSION:

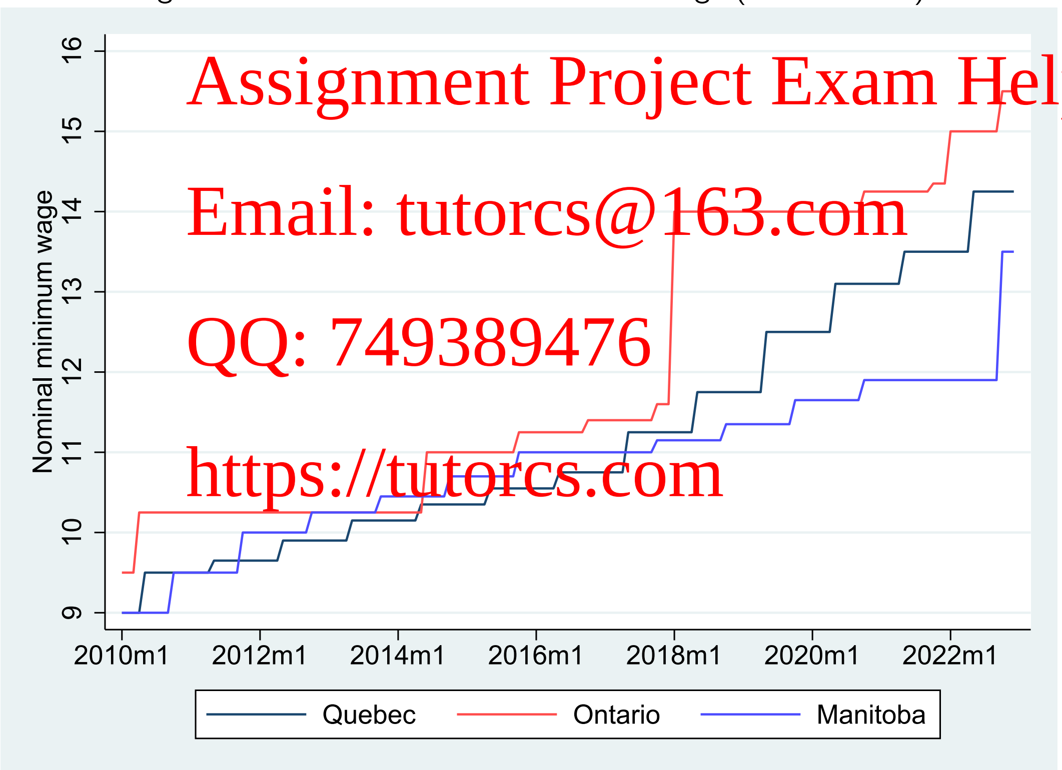
1. In addition to reporting on the distributions of these five estimators, provide a discussion of the simulation results. In particular, discuss how important model specification appears to be relative to omitted variable bias (or selection on unobservables).

Section D (40 marks) - Differences-in-differences

In this section you w[illegible]ange in the Ontario minimum wage using a difference-in-difference model. [illegible]ominal minimum wage for Ontario alongside its two neighbouring provin[illegible]uebec. Canadian provincial minimum wages tend to be pegged to inflati[illegible]ases at regular intervals.[1]

In January 2018 the [illegible]nt enacted a non-standard increase in the minimum wage from \$11.60 to \$14, just 3 months after its annual inflation adjustment in October 2017.[2] This is an increase of \$2.40, or 20%, during a period of low inflation.[3]

Figure 1: Provincial Nominal Minimum Wage (Canadian \$'s)



For this question I have provided you with a 20% sample of the publicly available Canadian Labour Force Survey from 2010-2019. The survey is a repeated cross-section and contains a select set of variables. The sample includes all working-aged adults. It is a large file and contains some variables which are not observed in all years.[4] As with all survey data, there may be missing data for certain variables.

[1] You can examine the provincial minimum wages in other provinces around this time using this link.

[2] The policy appears to have been announced in June 2017 (see link). One might interpret the change in minimum wage as an attempt to appease voters by the incumbent Liberal government of Ontario, given that provincial elections were to take place in June 2018. The Liberals lost the election to the Progressive Conservative Party, which is possibly why the nominal minimum wage remain fixed for almost three years after this unprecedented increase.

[3] Ontario has a different minimum wage for those under 18 and liquor servers. These too were increased by \$2.25 and \$2.10 respectively.

[4] Beware, this may affect your sample if included in a your estimating equation.

QUESTIONS:

1. Discuss why t[illegible]ght make for a good *or bad* 'natural' experiment.

2. Create two tir[illegible]onstrating the employment rate (proportion of population emplo[illegible]ged 15-24 for the period 2010 to 2019 in Ontario, Manitoba, anc[illegible]should depict the *annual* employment rate, while the second graph [illegible] denominated by calendar months, not years. Comment on the trend and seasonality of youth employment. Does it 'look' like the parallel trend assumption holds between Ontario and its neighbouring provinces.

3. Estimate a 2-period 2-group difference-in-difference model using the 2017 and 2018 data.

$$Y_{itc} = \alpha + \psi D_c + \delta T_t + \beta D_c \cdot T_t + \varepsilon_{itc} \tag{1}$$

Where the outcome Y_{itc} is a dummy variable indicating that individual i is employed. In this application, the assignment-group (denoted by c) is the individual's province of residence: $D_c = \mathbf{1}\{Ontario\}$. **You need to justify your choice of control group and may choose a province other than Manitoba or Quebec. You may also use more than one province as a control group, assuming there is reason to do so.**

In a second specification, include month fixed-effects (λ_m) to account for seasonality.

$$Y_{itc} = \alpha + \psi D_c + \delta T_t + \beta D_c \cdot T_t + \lambda_m + \nu_{itc} \tag{2}$$

In a third, include a set of good covariates of your own choosing. Justify your choice.

$$Y_{itc} = \alpha + \psi D_c + \delta T_t + \beta D_c \cdot T_t + \lambda_m + X'_{itc}\gamma + \upsilon_{itc} \tag{3}$$

Present the results in a single table and comment on any important differences across the specifications.

4. Estimate a dynamic difference-in-differences specification that will allow you to test the parallel trends assumption in the pre-treatment period. You should include data from 2014 to 2019 in the model and normalize the results relative to 2017, the year before treatment. Try to present the estimates of $\hat{\beta}_j$ in a graph, along with 95% confidence intervals.

$$Y_{itc} = \alpha + \psi D_c + \delta_t + \sum_{j \neq 2017} \beta_j \mathbf{1}\{t = j\} \cdot D_c + \lambda_m + \epsilon_{itc} \tag{4}$$

Comment on whether you find support for the parallel trends assumption. In addition, comment on the dynamics of the treatment effect in the post-treatment period. Are the results consistent with your expectations of the policy impact?

5. As we have monthly data, we could estimate a dynamic model with monthly treatment effects. What additional assumptions would we need to make with respect to the CEF of $Y_i(0)$?

6. The policy was announced in June 2017. As such, we may expect pre-emptive adjustments to the demand for labour. Propose *and execute* a way of checking for pre-emptive behaviour usi[illegible]

7. The estimate[illegible]y be explained by differential trends in the aggregate labour market[illegible]nstrate how we might use the remaining sample to test this hypo[illegible]

8. Are you convi[illegible]tes? By in large, the literature on minimum wages finds that a *reasonable* increase in the minimum has no effect on employment levels (see review by Manning, 2021).

Section E (10 marks) - Synthetic Control

In this section you w[illegible] control estimator to examine the impact of the above minimum wage poli[illegible] he code needed to execute this estimator in STATA. R users should be ab[illegible] milar Synth package developed by Abadie, Diamond, and Hainmueller (se[illegible] velopers as the corresponding STATA package.

```
ssc install synt[illegible]
** Look at the syntax help synth

use lfs_2010_2019_ages1564_20per.dta, clear

keep if agegrp<=2
gen employed = empstat<=2 if empstat!=.
gen male = sex==1 if sex!=.
tab edugrp if edugrp!=., gen(edu)
recode efamtype (1 = 1) (2/4 = 2) (14 16 17 = 3) (5/10 = 4) (11/13 15 = 5) (18 =
6), gen(fam)
tab efamtype fam [illegible]
tab fam if fam!=., gen(fam)

collapse (mean) employed male edu1-edu3 fam1-fam6 [w=wgt], by(province year)

tsset province year

** 1.  Match over the outcome in the pre-period
synth employed , tru(6) trp(2018) fig
** 2.  Match over the outcome in the periods 2014-2016
synth employed employed, tru(6) trp(2018) fig mspeperiod(2014(1)2016)

** 3.  Match over covariates in the pre-period
synth employed male edu2 edu3 fam2 fam3 fam4 fam5 fam6, tru(6) trp(2018) fig
** 4.  Match over covariates in the periods 2014-2016
synth employed male edu2 edu3 fam2 fam3 fam4 fam5 fam6, tru(6) trp(2018) fig
mspeperiod(2014(1)2016)
```


DISCUSSION:

1. Comment on the appropriateness of the weights assigned to each of the control provinces in each instance. As a researcher, discuss whether you think it is more appropriate to choose a control group (as in Section D) or construct one using a method such as synthetic control.

Note: There is no one right answer to this question.