Name: _____

Date: Wednesday, December 9th, 2020

Empirical Analysis 1

Final Exam

1. Please keep your cameras on throughout the exam.
2. Exams should be handwritten on paper, a tablet or a similar device.
3. You will be given 20 minutes at the end of the exam to scan/save any answers and upload them to Canvas.
4. The exam is open book and open notes, but not open internet in that the use of other materials is prohibited.
5. No calculators are allowed.
6. If you have any questions during the exam please message the TAs privately using Zoom.
7. There are a total of 115 possible points, of which 15 are labeled "EXTRA CREDIT."
8. Answer as many questions as you can. You do not need to answer the questions in order. Try to answer the later parts of a question even if you have difficulty with earlier parts.
9. Please clearly label your final answers where appropriate.
10. Any students caught cheating will fail the course. The Dean of Students will be notified as well.
11. Good luck!

1 True or False? (15 points)

Indicate whether each statement is true of false. If the statement is true, please provide a proof. If the statement is false, please provide a counterexample.

(a) Tightness (5 points) Suppose that the sequences (X_n) and (Y_n) are tight. Then so is the sequence (Z_n) defined by

$$Z_n = \begin{cases} X_n & \text{if } n \text{ is even} \\ Y_n & \text{if } n \text{ is odd} \end{cases}.$$

(b) Mean-Independence (5 points) Suppose X, Y are real-valued random variables. If X is mean-independent of Y and Y is mean-independent of X, then X and Y are independent.

(c) Convergence in Probability (5 points) Suppose $(X_i)_{i\in\mathbb{N}}$ are i.i.d. non-negative and random variables such that $\mathbb{P}[X_1 \in (a, b)] = 1$ for some $0 < a < b < \infty$. Then

$$\sqrt[n]{\prod_{i=1}^{n} X_i}$$

converges in probability as n tends to infinity.

2 BLP and Measurement Error (28 points)

(a) (10 points) Let (Y, X_1, X_2, X_3, U) be a random vector such that

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + U$$

where $E[U] = E[X_1 U] = E[X_2 U] = E[X_3 U] = 0$ by the BLP first order condition, and $\text{Cov}(X_1, X_3) \neq 0$, $\text{Cov}(X_2, X_3) \neq 0$.
Now suppose you only have data on Y, X_1 and X_2, and run a regression

$$Y = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + V$$

where $E[V] = E[X_1 V] = E[X_2 V] = 0$ by the BLP first order condition. Is $\alpha_1 = \beta_1$? Justify your answer by providing a proof or counterexample. If the answer is no, can you write a sufficient condition so that $\alpha_1 = \beta_1$? (Justify that too.)

(b) (5 points) Now suppose we are interested in the

$$Y = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + V$$

regression, particularly the parameter α_1, where $E[V] = E[X_1 V] = E[X_2 V] = 0$ by the BLP first order condition. Suppose that X_1 is measured with error

$$X_1^{obs} = X_1 + \epsilon$$

where X_1 is the true value, X_1^{obs} is the observed value and the error ϵ is independent of Y,X_1,X_2. How does the coefficient γ_1 from the regression

$$Y = \gamma_0 + \gamma_1 X_1^{obs} + \gamma_2 X_2 + W$$

relate to α_1? As before, assume $E[W] = E[X_1^{obs} W] = E[X_2 W] = 0$ by the BLP first order condition. For simplicity, assume all coefficients are positive numbers. You may appeal to results derived in class with appropriate reference to answer this question.

(c) (10 points) Suppose X_1 is still measured with error

$$X_1^{obs} = X_1 + \epsilon$$

where X_1 is the true value, X_1^{obs} is the observed value and the error ϵ is independent of Y,X_1,X_2. Suppose we run a reverse regression of X_1^{obs} on Y.

$$X_1^{obs} = \delta_0 + \delta_1 Y + \delta_2 X_2 + R \ ,$$

where $E[R] = E[YR] = E[X_2 R] = 0$ by the BLP first order condition. How does the coefficient δ_1 relate to α_1? As before, assume all coefficients are positive numbers for simplicity.

(d) (3 points) How could you use the parameters γ_1 an δ_1 obtained from the two regressions in part (b) and (c) to obtain bounds for the true coefficient α_1? What makes this interval more informative?

3 Heterogeneous Treatment Effects (27 points)

You are interested in examining the effects of education of earnings. You observe an iid sample $\{Y_i, D_i\}_{1 \leq i \leq n}$ where Y_i denotes the observed income of individual i and D_i denotes whether individual i attended college. Let $Y_i(1)$ and $Y_i(0)$ denote the potential outcomes with and without treatment, which can be linked to observed outcomes through the usual switching regression:

$$Y_i = Y_i(0) + D_i \left[Y_i(1) - Y_i(0) \right] \ .$$

(a) (2 points) Consider the following regression:

$$Y_i = \beta_0 + \beta_1 D_i + \nu_i.$$

Provide a sufficient condition in terms of the potential outcomes such that β_1 recovers $E[Y(1) - Y(0)]$. Do you think this assumption is plausible in this setting? Why / why not?

As a robustness exercise you consider controlling for the age of each individual using the following specification:

$$Y_i = \beta^* D_i + \sum_{a=1}^{A} \gamma_k \mathbb{1}\{X_i = a\} + \epsilon_i \ ,$$

where X_i takes on values in the finite set $\{1, \ldots, A\}$ and denotes the age (in years) of individual i, and $\mathbb{1}\{X_i = a\}$ is an indicator variable which takes the value 1 if individual i is a years old. You are only interested in the coefficient β^*, and so to simplify your analysis you work with the residualized regression equation

$$\tilde{Y}_i = \beta^* \tilde{D}_i + \tilde{\epsilon}_i,$$

where $\tilde{D}_i = D_i - BLP\left(D_i | \mathbb{1}\{X_i = 1\}, ..., \mathbb{1}\{X_i = A\}\right)$, and similarly for Y_i.

(b) (5 points) Is it true that $BLP\left(D_i | \mathbb{1}\{X_i = 1\}, \ldots, \mathbb{1}\{X_i = A\}\right) = E[D_i | X_i]$ in this setting? If yes, verify your answer using the orthogonality condition used in the definition of conditional expectations. If not, provide a counterexample.

(c) (14 points) Show that β^* can be written in the following form

$$E\left[w(X)\left(E[Y|D=1, X] - E[Y|D=0, X]\right)\right]$$

where $w(X) > 0$ and $E[w(X)] = 1$. Provide an expression for the weights $w(x)$.

(d) (3 points) What additional conditions do you need to impose to guarantee that $\beta^* = E[Y(1) - Y(0)]$? Are they stronger or weaker than those in (a)? [Hint: You may answer this question using the result in part (c) even if you were unable to do it.]

As an additional specification you wish to allow for the possibility that some individuals attended college, but dropped out without completing their degree. Let $Y_i(0)$, $Y_i(1)$ and $Y_i(2)$ denote these potential outcomes

without college, with some college and with completed college respectively. You consider the following specification:

$$Y_i = \beta_1 \mathbb{1}\{D_i = 1\} + \beta_2 \mathbb{1}\{D_i = 2\} + \sum_{a=1}^{A} \gamma_k \mathbb{1}\{X_i = a\} + \epsilon_i,$$

where X_i denotes the age of individual i, and $\mathbb{1}\{X_i = a\}$ is an indicator variable which takes the value 1 if individual i is a years old as before.

(e) (3 points) Derive expressions for:

i. $\text{Var}(\mathbb{1}\{D_i = 1\} \mid X)$

ii. $\text{Var}(\mathbb{1}\{D_i = 2\} \mid X)$

iii. $\text{Cov}(\mathbb{1}\{D_i = 1\}, \mathbb{1}\{D_i = 2\} \mid X)$

in terms of the probabilities $p_0(X) \equiv \mathbb{P}[D_i = 1|X]$, $p_1(X) \equiv \mathbb{P}[D_i = 1|X]$ and $p_2(X) \equiv \mathbb{P}[D_i = 2|X]$.

(f) (EXTRA CREDIT +15 points) Derive an expression for β_1 of the form

$$\beta_1 = E[w_1(X)\,(E[Y|D = 1, X] - E[Y|D = 0, X])] + E[w_2(X)\,(E[Y|D = 2, X] - E[Y|D = 0, X])]$$

where $E[w_1(X)] = 1$ and $E[w_2(X)] = 0$.

4 MLE (18 points)

Let $(Y_i, X_i), i = 1, \ldots, n$ be an i.i.d. sequence of observed random variables, where $Y_i \in \{0, 1\}$ denotes whether a voter chose to vote for Donald Trump in the 2020 election and $X_i \in \mathbb{R}^d$ are the demographic characteristics that determine the voters' latent preference for Mr. Trump. Suppose that each voter chooses whether to vote for Mr. Trump according to this rule:

$$Y_i = \mathbb{1}\{X_i'\beta_i \geq U_i\}$$

where $(\beta_i, U_i), i = 1, \ldots, n$ are i.i.d. unobserved random variables with $\beta_i \in \mathbb{R}^d$ and $U_i \in \mathbb{R}$. Assume further that X_i, β_i and U_i are mutually independent, $U_i \sim N(0, 1)$ and $\beta_i \sim N(\beta, \Sigma)$, where $\beta \in \mathbb{R}^d$ is unknown and Σ is the known covariance matrix of β_i.

(a) (5 points) What is the distribution of $X_i'(\beta - \beta_i) + U_i$ conditional on X_i ?

(b) (8 points) Use your answer to to construct the log likelihood for estimating β. [Hint: Rewrite the Y_i equation in terms of the standard normal CDF of the observables.]

(c) (5 points) Compute the limiting variance for the (appropriately centered and normalized) MLE estimator under the assumption that the model is correctly specified and any additional conditions necessary for asymptotic normality are satisfied.

5 Instrumental Variables (12 points)

You are interested in estimating β in the following regression:

$$Y_i = \beta_0 + \beta_1 X_{i,1} + \epsilon_i \ . \tag{1}$$

We observe an i.i.d. sample of size n of (Y_i, X_i), where $X_i = (1, X_{i,1})'$, that satisfies $E[X_i X_i'] < \infty$. We wish to interpret (1) as a causal model of Y_i, but do not believe that $E[X_{i1}\epsilon_i] = 0$. Let $Z_i = (1, Z_{i,1})'$ be an additional random vector satisfying $E[Z_i Z_i'] < \infty$, $\text{Var}[Z_{i,1}] > 0$, $E[Z_i X_i'] < \infty$, $\text{Cov}[Z_{i,1}, X_{i,1}] \neq 0$ and $E[Z_i \epsilon_i] = 0$.

(a) (4 points) Suppose we stack our observations so that $\mathbb{Y} = (Y_1, ..., Y_n)'$, $\mathbb{X} = (X_1, ..., X_n)'$ and $\mathbb{Z} = (X_1, ..., X_n)'$ so that the TSLS estimator can be written as

$$\hat{\beta}_{2SLS} = (\mathbb{X}'\mathbb{P}_Z\mathbb{X})^{-1}(\mathbb{X}'\mathbb{P}_Z\mathbb{Y}) \ ,$$

where $\mathbb{P}_Z$ is some n-by-n matrix. Provide an expression for $\mathbb{P}_Z$ and use it to verify that $\hat{\beta}_{2SLS} = \hat{\beta}_{IV} = (\mathbb{Z}'\mathbb{X})^{-1}(\mathbb{Z}'\mathbb{Y})$.

(b) (8 points) Suppose instead that we first run the following regression of X_{i1} on Z_i

$$X_{i1} = \gamma_0 + \gamma_1 Z_{1i} + U_i \ ,$$

where we then estimate $\hat{\gamma}_{OLS} = (\mathbb{Z}'\mathbb{Z})^{-1}(\mathbb{Z}'\mathbb{X}_1)$ in order to construct the residuals $\hat{U}_i = X_{i1} - \hat{\gamma}_0 - \hat{\gamma}_1 Z_{1i}$. Finally we run the regression

$$Y_i = \beta_0 + \beta_1 X_{i1} + \gamma \hat{U}_i + \nu_i \ .$$

Show that in this model $\hat{\beta}$ can be written as $\hat{\beta} = (\mathbb{X}'\mathbb{M}_U\mathbb{X})^{-1}(\mathbb{X}'\mathbb{M}_U\mathbb{Y})$ for some n-by-n matrix $\mathbb{M}_U$ and use it to prove the new estimator is identical to the one you found in (a).