ECON 21020: Final Examination

Allotted Time: 2 Hours

Problem 1

(20 Points.) You collect an *i.i.d.* sample $\{X_i\}_{i=1}^n$, where X_i denotes the average walking speed of a University of Chicago student. You assume these variables follow the CDF $F(x) = P(X_i \leq x)$.

(a) Show that $\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^n \mathbb{1}\{X_i \leq x\}$ is a consistent estimator for $F(x)$.

(b) Show that $\text{Var}(\mathbb{1}\{X_i \leq x\}) = F(x)[1 - F(x)]$.

(c) Write down the limiting distribution of $\sqrt{n}(\hat{F}_n(x) - F(x))$.

(d) Derive a consistent estimator for $\text{Var}(\mathbb{1}\{X_i \leq x\})$. Is your estimator unbiased? Explain.

(e) Suppose X_i is measured in miles per hour. Describe how to test H_0: $P(X_i \leq 2) = 0.5$ against the alternative H_1: $P(X_i \leq 2) \neq 0.5$. That is, write down T_n and c_n given level α.

Problem 2

(20 Points.) Let H be the average house price in a neighborhood, and let C be the number of upscale coffee shops per square mile. You want to study how coffee shops make neighborhoods more expensive, but your friend wants to study how more expensive neighborhoods tend to attract more coffee shops. Based on these research questions, you write down two models:

$$H = \alpha_0 + \alpha_1 C + U \quad (1)$$
$$C = \beta_0 + \beta_1 H + \varepsilon \quad (2)$$

Throughout this problem, you may assume that $E(U) = E(\varepsilon) = E(U\varepsilon) = 0$.

(a) Show that C is endogenous in model (1) and that H is endogenous in model (2).

(b) If you think that $\beta_1 > 0$ and $\alpha_1\beta_1 < 1$, will running an OLS regression on model (1) over- or under-estimate the effect of coffee shops on local house prices? Provide clear justification.

(c) Suppose you observe some external shock Z that only affects the number of coffee shops. Assume that $E(ZU) = 0$. Using Z as an instrument, show how to consistently estimate α_1.

(d) Can you also use Z as an instrument for H in model (2)? Explain.

Problem 3

(20 Points.) You want to study the effect of broadband internet on economic development. So, you write down the model $Y = \beta_0 + \beta_1 \text{WiFi} + X'\gamma + U$, where Y denotes income per capita, WiFi is local internet speed, X is some vector of exogenous control variables, and U captures all unobserved determinants of income. Throughout this problem, assume that $E(U) = 0$.

(a) Give a causal interpretation to β_1. Can you estimate this effect by running OLS? Explain.

(b) Let Z be some technological shock that reduces the cost of installing broadband internet. Given an *i.i.d.* sample $\{Y_i, \text{WiFi}_i, X_i, Z_i\}_{i=1}^n$, describe how to test for instrument relevance.

(c) Can you test whether $E(ZU) = 0$? Briefly explain.

(d) Assume Z is a valid instrument. Show how to estimate $\kappa = (\beta_0, \beta_1, \gamma)$ with the IV estimator.

(e) Suppose $\sqrt{n}(\hat{\kappa}_n - \kappa) \xrightarrow{d} N(0, \Omega)$. You compute $\hat{\kappa}_n$ and a consistent estimate $\hat{\Omega}_n$ for Ω. Explain how you would test whether β_1 is significantly different from zero at the 5% level (i.e. $\alpha = 0.05$). That is, state H_0, the appropriate test statistic T_n and critical value c_n.

(f) Suppose you have a second instrument (call it "Q"). Describe how to estimate κ via TSLS.

Problem 4

(20 Points.) Suppose you want to understand how neighborhoods affect children's future earnings. You collect an *i.i.d.* sample $\{Y_i, D_i\}_{i=1}^n$, where Y_i is a child's income in adulthood and D_i indicates whether that child grows up in a low-poverty neighborhood ($1 =$ *yes*, $0 =$ *no*).

(a) Define the ATE, ATT, and ATU. Can estimate them by regressing Y_i on D_i? Explain.

(b) Suppose you collect data about X_i, the parents' annual incomes when the child is young.

- (i) Under *selection-on-observables* (i.e. $Y_{d,i} \perp D_i | X_i$), explain how to estimate the ATE.
- (ii) Do you believe that *selection-on-observables* is likely to hold in this case? Explain.

(c) Suppose there exists a program that gives housing vouchers to underserved families with young children. Each voucher subsidizes the cost of moving to a low-poverty neighborhood. These vouchers are randomly assigned to families, and not everyone who is offered a voucher accepts one. Let Z_i indicate whether a child's family receives a voucher ($1 =$ *yes*, $0 =$ *no*). Assume that your population consists of children whose families are eligible for vouchers.

- (i) Define the *compliers*, *always-takers*, *never-takers*, and *defiers* in this setting.
- (ii) Argue whether Z_i is a valid instrument for D_i.
- (iii) Assuming Z_i is a valid instrument for D_i will running IV estimate the ATE? Explain.
- (iv) Describe how to consistently estimate the average treatment effect of Z_i on Y_i.

Problem 5

(20 Points.) Suppose you can see into the future. You collect an *i.i.d.* sample $\{Y_i, X_i\}_{i=1}^n$, where Y_i is the number of years that someone will live, and X_i is an individual's current age. You believe that the conditional PDF of Y_i given X_i is:

$$f_\theta(y|x) = \begin{cases} \frac{\theta x^\theta}{y^{\theta+1}} & \text{for } y \geq x \\ 0 & \text{otherwise} \end{cases}$$

(a) Write down the (conditional) likelihood and log-likelihood functions for θ.

(b) Compute the score $s(\theta|x)$ and the Fisher information matrix $\mathcal{I}(\theta|x)$.

(c) Show that the maximum likelihood estimator for θ is given by $\hat{\theta}_n = \left(\frac{1}{n}\sum_{i=1}^n \log(Y_i/X_i)\right)^{-1}$.

(d) Using the fact that $E[\log(Y/X)] = \theta^{-1}$, argue that $\hat{\theta}_n$ is a consistent estimator for θ.

(e) Is $\hat{\theta}_n$ an unbiased estimator for θ? Justify your answer.

(f) What is the limiting distribution of $\sqrt{n}(\hat{\theta}_n - \theta)$? What can you conclude about the asymptotic variance of the MLE in relation to all unbiased estimators for θ? Briefly explain.