Final exam

EC 607

Due *before* midnight on Friday, 12 June 2020

Instructions

DUE Your solutions to the final are due *before* 11:59pm on Friday, 12 June 2020 on Canvas.

Your **solutions must be typed** with R code beneath your responses. *E.g.*, `knitr` and `R Markdown`. If you have problems knitting/rendering your solutions, we will accept a Word document (or similar) with an attached R script. All figures and tables must be in the write-up document.

OBJECTIVE This final has three purposes: (1) test your econometric understanding and intuition (especially regarding causality and inference); (2) reinforce/extend topics we discussed in class; (3) build your R toolset.

IMPORTANT **You may not work with anyone else**. You may not communicate with anyone about this exam except Colleen and Ed. Any evidence of a violation of this rule will result in a grade of zero.

IMPORTANT$_2$ **All materials are allowed**—notes, previous assignments/solutions, Google, *etc.*.

IMPORTANT$_3$ You may use any functions/packages you would like on this exam.

IMPORTANT$_4$ Thank you for a great term.

Section 1: Selection and effects

For each of the following questions, provide **two answers:**

1. **Which treatment** effect will the method estimate (ATE, TOT, ITT, LATE, *etc.*)?
2. Will the proposed method be **consistent** for estimating a treatment effect (*i.e.*, did we avoid selection bias)?

For example, your answer to one of these questions may be:

> ATE; consistent (no selection bias).

There is no partial credit: do not justify/explain your answers.

NOTE Do not assume anything beyond what is described in the problem—for example, do not assume the treatment effect is homogeneous unless the problem explicitly states it.

`1.01` To estimate the effect of a college degree on earnings, you randomly distribute a set of scholarships to high-school seniors. Then, using the earnings the individuals (who either received or did not receive the scholarships), you estimate the returns to education as the ratio **A**/**B** where

- **A** = the (estimated) effect of winning a scholarship on earnings
- **B** = the (estimated) effect of winning a scholarship on the probability of receiving a college degree

`1.02` Repeat **`1.01`** but now assume a constant treatment effect.

`1.03` To estimate consumers' own-price elasticity of demand, you regress $\log(q_i)$ (the log of individual i's quantity purchased) on $\log(p_i)$ (the log of the marginal price facing individual i) and an intercept.

`1.04` Repeat **`1.03`** but now assume a constant treatment effect.

`1.05`–`1.12` To estimate the effect of access to a credit card on consumption, you focus on applicants to a specific credit-card company that does not give cards to individuals with credit scores below 670. If an applicant has a credit-card score above 670, then she *may* be approved, depending upon a bunch of other factors (account balances, employment status, *etc.*). Consumers do not know where the cutoff is.

`1.05` You regress individual i's consumption on an intercept and an indicator for whether the individual's credit score exceeds 670, *i.e.*,

$$\text{Consumption}_i = \beta_0 + \beta_1 \mathbb{I}(\text{Score}_i > 670) + u_i$$

`1.06` Repeat **`1.05`** but assume a homogeneous treatment effect.

`1.07` Repeat **`1.05`** but assume that you only use data on individuals with credit scores between 650 and 690.

`1.08` Instead of the strategies described in **`1.05`–`1.07`**, you estimate 2SLS, where your outcome is consumption, your variable of interest is credit-card approval, and your instrument is an indicator for whether the individual's credit score is above 670. You use only individuals with credit scores between 650 and 690.

`1.09` Repeat **`1.08`** but assume a homogeneous treatment effect.

`1.10` Instead of the previous strategies, you use credit scores and your rich set of observed variables (demographics, job, banking data, *etc.*) to match each approved individual to her most similar rejected individual. You then take the average of the differences between the approved individuals and their matches. Assume your matches are very "good" in terms of proximity but that you do not have all of the data that the credit card company has.

`Continued on the next page`

Continuing ...

1.11 Repeat **1.10** but assume (1) homogeneous treatment effects and (2) you have all of the data that the credit-card company has.

1.12 Instead of the previous strategies, you use all of the variables to which the credit-card company has access to flexibly estimate (using logistic regression) the probability the individual is approved for a credit card. The credit-card company confirms that these probabilities are essentially correct and that any additional variation in credit-card acceptances was arbitrary. You then ensure your accepted and rejected groups have similar distributions in these probabilities. Finally, you use these estimated probabilities as a control when regressing consumption on an intercept and credit-card approval.

1.13–1.15 To estimate the causal effect of a treatment D_i on your outcome Y_i, you randomly select 25% of your sample to receive treatment $(\mathrm{D}_i = 1)$ and 75% to receive control $(\mathrm{D}_i = 0)$.

1.13 You estimate the effect of treatment by taking the difference between (A) the mean of the treated group and (B) the mean of the control group.

1.14 Repeat **1.13** but assume a homogeneous treatment effect.

1.15 For each treated individual i, you calculate $\hat{\tau}_i = \mathrm{Y}_i - \mathrm{Y}_{j(i)}$, where individual $j(i)$ is a randomly selected person from the control group (sampled with replacement). You then estimate the causal effect of treatment as the mean of the $\hat{\tau}_i$'s.

The next section begins on the following page.

Section 2: 2SLS, the simulation

In this 2SLS simulation, we have two valid instruments, Z_1 and Z_2 for the same endogenous treatment indicator (D). As usual, our outcome is called Y.

Individuals fall into five groups. Treatment effects are **constant within group** but **differ across groups**.

Group	Treatment determination	Group's trt. effect
G_1	$\mathrm{D}_i = 1$ regardless of Z_{1i} and Z_{2i}	$\tau_i = 0$ for i in G_1
G_2	$\mathrm{D}_i = 1$ if $\mathrm{Z}_{1i} - w_i > 0.5$. (Independent of Z_{2i})	$\tau_i = 2$ for i in G_2
G_3	$\mathrm{D}_i = 1$ if $\mathrm{Z}_{2i} - w_i > 0.5$. (Independent of Z_{1i})	$\tau_i = 0$ for i in G_3
G_4	$\mathrm{D}_i = 1$ if $\mathrm{Z}_{1i} - w_i > 0.5$ **and** $\mathrm{Z}_{2i} - w_i > 0.5$	$\tau_i = 10$ for i in G_4
G_5	$\mathrm{D}_i = 0$ regardless of Z_{1i} and Z_{2i}	$\tau_i = 0$ for i in G_5

Assume it is equally likely an individual belongs to any of the five groups.

Along with the definition of treatment (D_i) above (remember that it is separately defined for each group), define our data-generating process (DGP) as

$$
\begin{aligned}
u_i &\sim \text{Uniform}(-1,\, 1) \\
w_i &\sim \text{Uniform}(-4,\, 4) \\
\mathrm{Z}_{1i} &\sim \text{Bernoulli}(0.5),\ i.e.,\ \in \{0, 1\} \text{ with equal probability} \\
\mathrm{Z}_{2i} &\sim \text{Bernoulli}(0.5),\ i.e.,\ \in \{0, 1\} \text{ with equal probability} \\
\mathrm{Y}_{0i} &= u_i + w_i \\
\mathrm{Y}_{1i} &= \mathrm{Y}_{0i} + \tau_i \\
\mathrm{D}_i &= \text{Defined above} \\
\mathrm{Y}_i &= (1 - \mathrm{D}_i)\,\mathrm{Y}_{0i} + \mathrm{D}_i \mathrm{Y}_{1i}
\end{aligned}
$$

`2.01` Step 1: Generate a 100,000-individual sample that matches the defined DGP. Make sure your groups are approximately equally sized. *Hint:* The `rep()` function could be helpful here.

`2.02` Should we expect a regression of Y_i on D_i to produce a consistent estimate for the causal effect of treatment? Explain your answer. In other words: Using the defined DGP, explain why the treatment is endogenous.

`2.03` Run a regression that demonstrates the endogeneity of the treatment. Explain how this regression demonstrates the endogeneity and report your results.

`2.04` Bonus (Optional) Derive the selection bias inherent in the DGP.

`2.05` Calculate the selection bias in your generated sample. *Note:* This question is a calculation that should generate an actual number (the derivation is in **`2.04`**).

`2.06` Using the defined DGP, explain how you know the instruments are *exogenous*.

`2.07` Using the defined DGP, explain how you know the instruments are *relevant*.

`2.08` Using the defined DGP, explain why the instruments do not violate *monotonicity*.

`Continued on the next page`

2.09 Who are the compliers for each instrument? *Hint:* Be careful here.

2.10 Instrument D_i with Z_{1i}. What is your estimate for the average effect of treatment? Explain why the coefficient is "reasonable" based upon who the compliers are for this instrument.

2.11 Instrument D_i with Z_{2i}. What is your estimate for the average effect of treatment? Explain why the coefficient is "reasonable" based upon who the compliers are for this instrument.

2.12 Instrument D_i with Z_{1i} **and** Z_{2i}. What is your estimate for the average effect of treatment? Explain why the coefficient is "reasonable" based upon who the compliers are for this instrument.

2.13 Why should we expect to get different point estimates in **2.10**, **2.11**, and **2.12**?

2.14 True/False: If we instrument D_i with $\mathrm{Z}_{1i} \times \mathrm{Z}_{2i}$, the only compliers are members of G_4. Explain your answer.

2.15 Conduct an actual simulation with at least 500 iterations, each of size 100,000. In each iteration, estimate the three point estimates in **2.10–2.12 and also** an estimate that instruments D_i with Z_{1i}, Z_{2i}, **and** $\mathrm{Z}_{1i} \times \mathrm{Z}_{2i}$. Produce a well-labeled, aesthetically pleasing figure of the distributions of these four estimation strategies.

Note: If it is taking too long to run on your computer, you can make your sample size smaller (maybe 1,000?).

2.16 Do the centers of the distributions in **2.15** match your expectations for the estimates of the treatment effect? Explain.

2.17 Change the DGP for G_4's treatment to $\mathrm{D}_i = 1$ if $\mathrm{Z}_{1i} - w_i < 0.5$ **and** $\mathrm{Z}_{2i} - w_i < 0.5$. Notice that the signs have flipped from "$>$" to "$<$". Why does 2SLS no longer consistently estimate the treatment effect?