Intro to Econometrics **FINAL EXAM** Thursday 05/14/20

T. Christensen Time Allowed: 24 hours

Instructions

- This exam is not for distribution to anyone outside the class.
- This exam is open book. You may consult class notes but no other resources.
- Do not discuss the contents of this exam with anyone.
- Do not post questions on Piazza. If you have questions, please email me (Christensen).
- Justify your answers whenever possible to ensure full credit. Be clear and to the point.
- Upload your exam to NYU classes, preferably as a single pdf document.
- Please check that whatever you upload to NYU classes is clearly readable.
- There are 4 questions worth 70 points in total.

Honor Code

In submitting this exam, you are agreeing to be bound by the NYU College of Arts and Science Honor Code, the first part of which reads:

> As a student in the College, I pledge that I shall perform honestly all my academic obligations. I will not represent the words, works, or ideas of others as my own; will not cheat; and will not seek to mislead faculty or other academic officers in their evaluation of my course work or in any other academic affairs.

Any student found to have used online services for this exam, plagiarized, copied from or colluded with classmates about answers, or cheated in any other manner on this exam, will receive an F as their overall grade for the course and be referred to the Dean's office for further disciplinary action.

Intro to Econometrics **FINAL EXAM** Thursday 05/14/20

T. Christensen Time Allowed: 24 hours

Question 1. (25 points in total, each part is worth 5 points)

You wish to estimate the causal effect β_1 of X on Y:

$$Y_i = \beta_0 + \beta_1 X_i + u_i \,. \tag{1}$$

You are concerned endogeneity bias might lead to inconsistency of the OLS estimate of β_1. You have a control variable C_i, which is not binary. The control variable satisfies conditional mean independence:

$$\mathrm{E}[u_i|X_i, C_i] = \mathrm{E}[u_i|C_i] \,. \tag{2}$$

However, the conditional mean of u_i depends on C_i in a nonlinear fashion:

$$\mathrm{E}[u_i|C_i] = \delta_0 + \delta_1 C_i + \delta_2 C_i^2 \,, \tag{3}$$

where each of the δ coefficients is non-zero.

You have data on X_i, Y_i and C_i drawn i.i.d. from their joint distribution. You also know that each of Y_i and X_i has finite nonzero fourth moments and C_i has finite nonzero eighth moment.

Hint: conditioning on C_i is the same as conditioning on C_i and C_i^2. This is because C_i^2 contains no extra information beyond that contained in C_i. Therefore, $\mathrm{E}[u_i|C_i] = \mathrm{E}[u_i|C_i, C_i^2]$ and similarly for other conditional expectations.

(a) Propose an approach for consistently estimating β_1 from data on X_i, Y_i and C_i.

Be sure to clearly describe the model you would estimate. You should state what the dependent and explanatory variable/s are and the method you would use to estimate β_1.

(b) Write the model from (a) in BLP form. In answering, clearly relate the BLP coefficients to your parameter of interest β_1. Show your working to receive full credit.

(c) Show that the procedure you describe in part (a) will produce a consistent and unbiased estimate of β_1.

You do not need to provide a formal proof of consistency and unbiasedness, but you should be able to show whether or not the relevant key assumption is satisfied.

(d) Briefly explain and distinguish the concepts of consistency and unbiasedness. In answering, give an example of an estimator we've used this semester which is consistent but not unbiased.

(e) How, if at all, would your answer to (a) change if C_i was binary? Explain.

Question 2. (15 points in total, each part is worth 5 points)

You wish to investigate whether an individual's previous union membership status influences their current status. You have panel data on individuals' union membership over 4 years ($t = 1, 2, 3, 4$) on the variable M_{it}, which takes the value 1 if individual i was a union member in year t and 0 otherwise. You model individual i's utility from choosing to be a union member (U_1) or not (U_0) in year t as a function of previous membership status M_{it-1}, a fixed effect α_i, and random components $\varepsilon_{it,1}$ and $\varepsilon_{it,0}$:

$$U_1(M_{it-1}, \alpha_i, \varepsilon_{it,1}) = u_1(M_{it-1}, \alpha_i) + \varepsilon_{it,1}\,, \tag{4}$$

$$U_0(M_{it-1}, \alpha_i, \varepsilon_{it,0}) = u_0(M_{it-1}, \alpha_i) + \varepsilon_{it,0}\,. \tag{5}$$

The $\varepsilon_{it,0}$ and $\varepsilon_{it,1}$ terms represent the parts of individual i's utility from each choice in year t that are not explained by previous membership status and the fixed effect. These are drawn randomly each year whereas the fixed effect is constant over time. You assume

$$u_1(M_{it-1}, \alpha_i) - u_0(M_{it-1}, \alpha_i) = \beta_1 M_{it-1} + \alpha_i\,. \tag{6}$$

You also assume that, for each year t, the conditional distribution of $\varepsilon_{it,1} - \varepsilon_{it,0}$ given M_{it-1} and α_i is a logistic distribution:

$$(\varepsilon_{it,1} - \varepsilon_{it,0})|M_{it-1}, \alpha_i \text{ has cdf } \Lambda, \text{ where } \Lambda(u) = \frac{1}{1 + e^{-u}}\,. \tag{7}$$

(a) Derive an expression for $\Pr(M_{it} = 1|M_{it-1}, \alpha_i)$.

(b) Explain the role of the individual fixed effects in this model. What is it that we are attempting to control for by the inclusion of individual fixed effects?

(c) Unlike panel regression models, here there is no obvious way to difference out the individual fixed-effect α_i from the expression you obtained in (a). After some algebra, you deduce

$$\Pr(M_{i2} = 1|M_{i4}, M_{i2} + M_{i3} = 1, M_{i1}, \alpha_i) = \frac{1}{1 + e^{-\beta_1(M_{i1} - M_{i4})}}\,, \tag{8}$$

$$\Pr(M_{i2} = 0|M_{i4}, M_{i2} + M_{i3} = 1, M_{i1}, \alpha_i) = \frac{e^{-\beta_1(M_{i1} - M_{i4})}}{1 + e^{-\beta_1(M_{i1} - M_{i4})}}\,. \tag{9}$$

Describe how you could use these expressions to estimate β_1. Be sure to clearly describe the model you would estimate. You should state what the dependent and explanatory variable/s are, the (subset of) data you would use, and the method you would use to estimate β_1.

<u>Hint:</u> You might want to consider only "switchers": these are individuals who change union membership status between dates 2 and 3 (i.e., for whom $M_{i2} + M_{i3} = 1$).

Intro to Econometrics **FINAL EXAM** Thursday 05/14/20

T. Christensen Time Allowed: 24 hours

Question 3. (15 points in total, each part is worth 5 points)

Two economists wish to investigate whether news about COVID-19 related hospitalizations triggers consumers to seek face masks, disinfectant, and the like, in response. Together, they assemble a data set of COVID-19 related hospitalizations in New York and a Google Trends index of searches for "face mask" in New York. The data are daily and span the period February 29 to May 7, 2020.

Figure 1: Google Trends index.

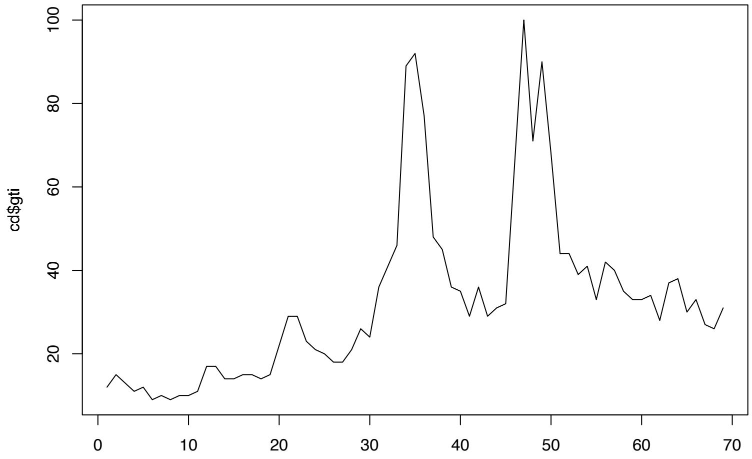
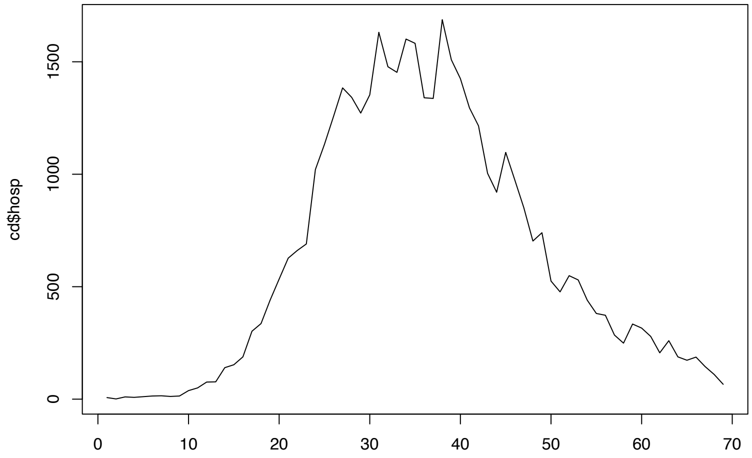


Figure 2: Hospitalizations.



The first economist runs a regression of the Google Trends index gti_t on the total number of hospitalizations the previous day $hosp_t$ (note: $hosp_t$ represents the total number of hospitalizations on day $t-1$, since this is only known at the end of date $t-1$) and obtains the following R output:

```
> fm0 <- lm(gti ~ hosp, data = cd)
> summary(fm0)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 21.284007   3.460304   6.151 4.84e-08 ***
hosp         0.018346   0.004189   4.380 4.27e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 18.82 on 67 degrees of freedom
Multiple R-squared:  0.2226,Adjusted R-squared:  0.211
F-statistic: 19.18 on 1 and 67 DF,  p-value: 4.272e-05

> coeftest(fm0, df = Inf, vcov = vcovHAC)

z test of coefficients:

              Estimate Std. Error z value Pr(>|z|)
(Intercept) 21.2840070  6.7613407  3.1479 0.001644 **
hosp         0.0183465  0.0059299  3.0939 0.001976 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The second economist runs a regression of $\Delta gti_t = gti_t - gti_{t-1}$ on the change in hospitalizations $\Delta hosp_t = hosp_t - hosp_{t-1}$ and obtains:

```
> fmd0 <- lm(diff(gti) ~ diff(hosp), data = cd)
> summary(fmd0)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.26005    1.35401   0.192   0.8483
diff(hosp)   0.02231    0.01230   1.814   0.0742 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11.17 on 66 degrees of freedom
Multiple R-squared:  0.04748,Adjusted R-squared:  0.03305
```

```
F-statistic:  3.29 on 1 and 66 DF,  p-value: 0.07425

> coeftest(fmd0, df = Inf, vcov = vcovHAC)

z test of coefficients:

            Estimate Std. Error z value Pr(>|z|)
(Intercept) 0.260051   1.416218  0.1836   0.8543
diff(hosp)  0.022314   0.015189  1.4690   0.1418
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

(a) How does the interpretation of the slope coefficient differ across the two economists' models? Which of the two interpretations seems more relevant to the economists' research question?

(b) Which of the two sets of results provides more reliable evidence of the causal effect of news about hospitalizations on consumer behavior? In answering, be sure to state whether the effect is significant or not.

The two economists notice that the second spike in the Google Trends index around day 47 coincides with the announcement by Governor Cuomo that face masks would be mandatory in New York. They define a dummy variable D_t that takes the value 0 before April 15 and 1 on and after April 15.

The first economist performs a Chow test for a structural break on April 15 and obtains:

```
> fm1 <- lm(gti ˜ hosp + D + D:hosp, data = cd)
> summary(fm1)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 10.208618   3.084119   3.310  0.00152 **
hosp         0.022826   0.003187   7.162 8.97e-10 ***
D            2.826095   6.360400   0.444  0.65828
hosp:D       0.060502   0.013699   4.417 3.88e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 13.09 on 65 degrees of freedom
Multiple R-squared:  0.635,Adjusted R-squared:  0.6181
```

```
F-statistic: 37.69 on 3 and 65 DF,  p-value: 3.122e-14

> coeftest(fm1, df = Inf, vcov = vcovHAC)

z test of coefficients:

              Estimate Std. Error z value  Pr(>|z|)
(Intercept) 10.2086176  1.2444053  8.2036 2.333e-16 ***
hosp         0.0228260  0.0042302  5.3960 6.815e-08 ***
D            2.8260955  5.9497960  0.4750    0.6348
hosp:D       0.0605016  0.0139306  4.3431 1.405e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> waldtest(fm0, fm1, test = "Chisq", vcov = vcovHAC)
Wald test

Model 1: gti ˜ hosp
Model 2: gti ˜ hosp + D + D:hosp
  Res.Df Df  Chisq Pr(>Chisq)
1     67
2     65  2 111.88  < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

(c) Explain what components of this output provide evidence of a structural break in the relation between search behavior for face masks and hospitalizations at April 15. In answering, be sure to state the null hypothesis you are testing and whether or not you reject the null hypothesis.

Question 4. (15 points in total, each part is worth 5 points)

You have been tasked with the following consulting project by a firm. The firm would like individuals to be remunerated for how they perform. However, the firm is worried that there may be unconscious bias, through which workers who currently earn high wages may be more likely to earn high wages in future, irrespective of their performance on the job.

The firm decides to run an experiment to investigate this issue. For an incoming cohort of graduates, the firm randomly assigns a wage W_{i1} to each individual i for their first year. Each individual's wages are then recorded for the subsequent two years. It is hypothesized that wages for the subsequent two years ($t = 2, 3$) evolve according to the model

$$W_{it} = \beta_1 W_{it-1} + \alpha_i + u_{it} \, , \tag{10}$$

where $\beta_1 < 1$ is an unknown parameter to be estimated, α_i is an individual fixed effect, and u_{it} is drawn independently each year.

The firm has given you a balanced panel of wages for years $t = 1, 2, 3$ for a large cohort of individuals.

(a) Explain whether or not you can estimate the model (10) by a panel regression of W_{it} on W_{it-1} using either of our two approaches for panel regression.

Hint: check if any of our assumptions for the fixed effects model are violated. If your answer is negative, you should provide some reasoning for why the relevant assumption fails.

(b) You notice that

$$\Delta W_{i3} = \beta_1 \Delta W_{i2} + \Delta u_{i3} \, , \tag{11}$$

where $\Delta W_{i3} = W_{i3} - W_{i2}$, $\Delta W_{i2} = W_{i2} - W_{i1}$, and $\Delta u_{i3} = u_{i3} - u_{i2}$.

Calculate $\text{Cov}(\Delta W_{i2}, W_{i1})$ and $\text{Cov}(\Delta u_{i3}, W_{i1})$.

(c) Using your answer to (b), propose an estimator of β_1 and show it is consistent.