LPO 8852: Regression II
Vanderbilt University
Midterm Exam
October 24, 2023

Name: _____

By signing below, I agree to the terms of Vanderbilt University's honor code. I attest that I have not collaborated with, or received any external assistance from other individuals on this at-home exam.

Signature: _____

Instructions: Read each question carefully and provide clear, concise responses in a separate document. Be sure to complete every part of every question. Partial credit will be given where appropriate. If you make any assumptions to answer a question, please state those explicitly. You may use lecture notes or other reference materials for this exam, but you must complete the exam **on your own**. Use of Chat GPT or other AI technology is not permitted. Please submit your responses via email to `sean.corcoran@vanderbilt.edu` before 9:00 am on Thursday, October 26. Good luck!

Question 1. Rebecca Jack et al. (2023) used district-level data to estimate the effects of COVID-related school closures during the 2020-21 school year on student test performance. Specifically, their team collected data on the percent of school days in 2020-21 that each district offered in-person, virtual, or hybrid instruction. (Hybrid is some combination of in-person and virtual). These measures were then used in a regression predicting the percent of students in grades 3-8 scoring proficient or better on statewide tests in mathematics and English Language Arts (ELA). (**40 points**)

(a) One approach to this question might be to compare mean test outcomes in 2020-21 in districts that primarily used virtual instruction to those in districts that remained in person. This could be done in a regression framework, in which other characteristics of the districts are used as covariates, such as the percent of students who are economically disadvantaged and the number of COVID-19 cases per 1,000 residents. What is the chief disadvantage of this design, if the aim is identify the causal effect of exposure to virtual instruction? Carefully explain, with reference to potential outcomes. (**10 points**)

(b) Rather than use the approach described in (a), Jack et al. used *panel data* for districts over five years: 2016-2019 and 2021. (Test results were not available from 2020 due to COVID test cancellations). They estimated the following regression:

$$pass_{ict} = \alpha + \beta_1(\%InPerson_{it}) + \beta_2(\%Hybrid_{it}) + \gamma_{ct} + \delta_t + \eta_i + \Pi\mathbf{X_{ict}} + \epsilon_{ict}$$

where i is a school district, c is a location, and t is the year. ("Location" is alternatively defined as state, commuting zone, or county). $pass_{ict}$ is the percent of students testing proficient in math or ELA in district i and year t. $\%InPerson_{it}$ and $\%Hybrid_{it}$ represent the percent of time district i spent in in-person and hybrid instruction, respectively, in year t. (The percent of time in virtual instruction is omitted). $\mathbf{X_{ict}}$ are time-varying district covariates.

Carefully explain what the terms η_i and δ_t represent, and what their roles are in this regression specification. (**10 points**)

(c) The regression model in (b) could be described as a type of "difference-in-differences" design. Explain why this model can be thought of in this way. (**5 points**)

(d) Under what assumption(s) can we interpret the coefficients β_1 and β_2 as the causal effects of exposure to in-person and hybrid instruction, respectively? Be as specific as you can to this example. (**5 points**)

(e) The authors used $\mathbf{X_{it}}$ and location-by-year effects γ_{ct} in an attempt to satisfy the assumptions for causal inference addressed in (d). (Location c was either state, county, or commuting zone. Like metropolitan statistical areas, commuting zones are areas typically larger than counties). What role do the γ_{ct} and $\mathbf{X_{it}}$ play in causal identification? Carefully explain. (**5 points**)

(f) Estimates for the regression in part (b) are shown in Panel A of Table 3 below. (Each panel and column in this table represents a separate regression). Carefully provide a written interpretation of the coefficient estimates in Panel A columns (1) and (5). (**5 points**)

TABLE 3—SCHOOLING MODE AND CHANGES IN PASS RATES

	Math			ELA		
	(1)	(2)	(3)	(4)	(5)	(6)
	Pass rate	Pass rate	Pass rate	Pass rate	Pass rate	Pass rate
Panel A. Main specifications						
% in-person	0.140	0.134	0.128	0.0813	0.0828	0.0872
	(0.0137)	(0.0147)	(0.0156)	(0.0102)	(0.0105)	(0.0105)
% hybrid	0.0776	0.0722	0.0743	0.0608	0.0537	0.0637
	(0.0143)	(0.0148)	(0.0161)	(0.0116)	(0.00949)	(0.00994)
Observations	11,041	11,041	11,041	11,064	11,064	11,064
Commute zone × year	No	Yes	No	No	Yes	No
County × year	No	No	Yes	No	No	Yes
Panel B. Demographic interactions						
% in-person × 2021	0.0960	0.156	0.0872	0.0686	0.0784	0.0729
	(0.0174)	(0.0196)	(0.0388)	(0.0138)	(0.0123)	(0.0276)
% hybrid × 2021	0.0379	0.0907	0.0280	0.0381	0.0409	0.0360
	(0.0169)	(0.0205)	(0.0388)	(0.0129)	(0.0123)	(0.0279)
% Black × % in-person × 2021	0.0943			0.0193		
	(0.0398)			(0.0240)		
% Black × % hybrid × 2021	0.0855			0.0508		
	(0.0472)			(0.0279)		
% Hispanic × % in-person × 2021		−0.135			−0.0178	
		(0.0680)			(0.0482)	
% Hispanic × % hybrid × 2021		−0.0664			0.0247	
		(0.0734)			(0.0421)	
% FRPL × % in-person × 2021			0.0810			0.000259
			(0.0582)			(0.0371)
% FRPL × % hybrid × 2021			0.0689			0.0153
			(0.0605)			(0.0380)
Observations	11,041	11,041	9,620	11,064	11,064	9,643
Commute zone × year	Yes	Yes	Yes	Yes	Yes	Yes

Question 2. The following questions ask you to interpret Stata results from two separate (unrelated) analyses. (**25 points**)

Analysis 1: A team of researchers is interested in the effect of an information campaign on the rate of e-cigarette use by teens aged 15-18. This information campaign was implemented in some states but not others. To estimate the effects of the campaign, the team has collected e-cigarette smoking rates by state, for 5 years before implementation and 5 years after. (Rates are the number of smokers per 100 persons). The Stata output below provides the team's initial estimates of the campaign's impact:

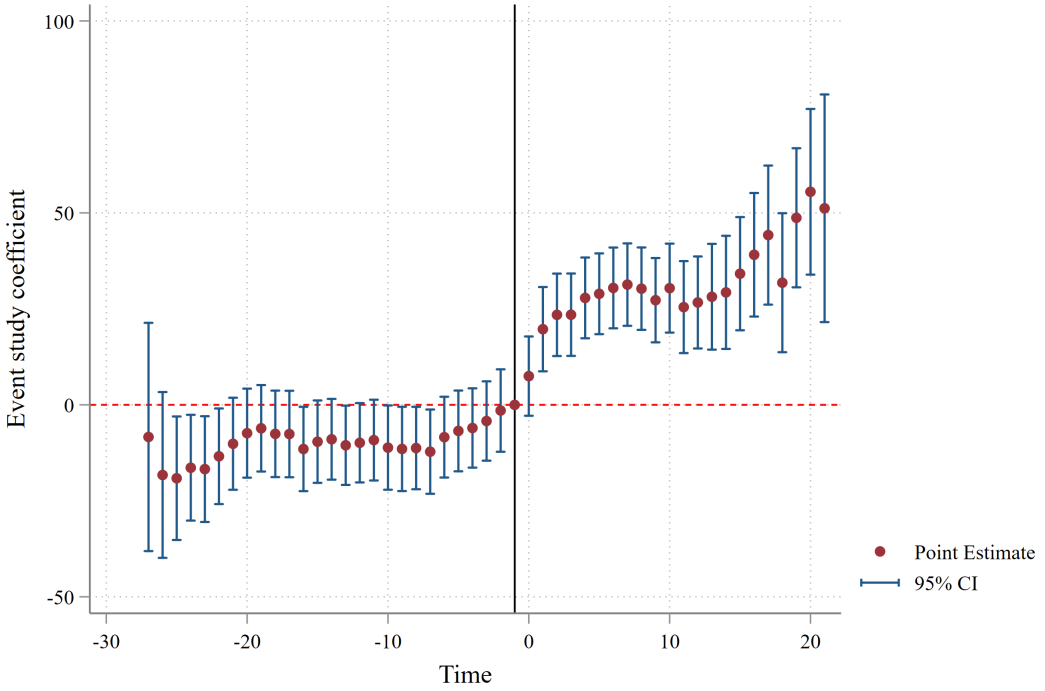
```
      Source |       SS           df       MS      Number of obs   =       500
-------------+----------------------------------   F(3, 496)       =    112.73
       Model |  50.1729863         3  16.7243288   Prob > F        =    0.0000
    Residual |  73.5829889       496    .1483528   R-squared       =    0.4054
-------------+----------------------------------   Adj R-squared   =    0.4018
       Total |  123.755975       499  .248007966   Root MSE        =    .38517

------------------------------------------------------------------------------
        ecigrate |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-----------------+------------------------------------------------------------
   1.evertreated |   .2306035   .0487201     4.73   0.000     .1348804    .3263267
         1.post |  -.6036862   .0487201   -12.39   0.000    -.6994093    -.507963
                |
evertreated#post |
            1 1 |   .4492737   .0689006     6.52   0.000     .3139007    .5846466
                |
          _cons |     7.4646   .0344503   216.68   0.000     7.396914    7.532287
------------------------------------------------------------------------------
```

(a) Carefully provide a written interpretation for the four coefficients in the output above. (**5 points**)

(b) What assumption(s) are required if one wants to interpret these results as providing the *causal* effect of the information campaign on smoking rates? How might you go about testing these assumptions? (**5 points**)

Analysis 2: A researcher is interested in the effects that eliminating school fees had on childrens' school enrollment in sub-Saharan Africa. She has annual country-level data on primary school enrollment rates and the year in which each country eliminated school fees. (The enrollment measure is "gross enrollment" which can be greater than 100% since it is the ratio of total enrollment in a grade level divided by the population of the age group typically served by that grade $\times 100$). The figure and Stata output below provide the event study from this analysis:




```
Fixed-effects (within) regression               Number of obs      =       490
Group variable: country2                        Number of groups   =        15

R-sq:  within  = 0.6278                         Obs per group: min =        27
       between = 0.0191                                        avg =      32.7
       overall = 0.2962                                        max =        35

                                                F(48,427)          =     15.01
corr(u_i, Xb)  = -0.1397                        Prob > F           =    0.0000

------------------------------------------------------------------------------
     primary |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      (lines omitted for space)
      lead23 |  -16.73508   7.007272    -2.39   0.017    -30.50812    -2.96204
      lead22 |  -13.39884   6.336948    -2.11   0.035    -25.85434   -.9433488
      lead21 |  -10.13638   6.096921    -1.66   0.097    -22.12009    1.847333
      lead20 |  -7.375621   5.897867    -1.25   0.212    -18.96809    4.216844
      lead19 |   -6.10891   5.729891    -1.07   0.287    -17.37121    5.153392
      lead18 |  -7.550735   5.729891    -1.32   0.188    -18.81304    3.711567
      lead17 |    -7.6034   5.729891    -1.33   0.185     -18.8657    3.658902
      lead16 |  -11.49576   5.585683    -2.06   0.040    -22.47461   -.5168984
      (lines omitted for space)
       lead5 |   -6.79485   5.351506    -1.27   0.205    -17.31342    3.723723
       lead4 |  -6.025198    5.25545    -1.15   0.252    -16.35497    4.304573
       lead3 |  -4.209601    5.25545    -0.80   0.424    -14.53937    6.120171
       lead2 |  -1.479151   5.460957    -0.27   0.787    -12.21285    9.254552
        lag0 |   7.484763    5.25545     1.42   0.155    -2.845008    17.81454
        lag1 |   19.71053    5.58498     3.53   0.000     8.733059    30.68801
        lag2 |   23.45178   5.460657     4.29   0.000     12.71866    34.18489
        lag3 |   23.48039   5.460788     4.30   0.000     12.74702    34.21376
```

```
        lag4 |   27.85973   5.352217     5.21   0.000     17.33975     38.3797
        (lines omitted for space)
       _cons |   84.90752   3.716437    22.85   0.000     77.60274    92.21231
-------------+----------------------------------------------------------------
     sigma_u |  21.346218
     sigma_e |  14.392643
         rho |  .68746938   (fraction of variance due to u_i)
------------------------------------------------------------------------------
F test that all u_i=0: F(14, 427) = 67.08                     Prob > F = 0.0000
```

(c) Carefully provide a written interpretation for the coefficient estimates `lead5` and `lag4` in the table above. (**5 points**)

(d) How would you interpret the results shown in the figure above? Do they support the conclusion that elimination of school fees had a causal effect on primary school enrollment? Briefly explain. (**5 points**)

(e) What does the F-test at the bottom of the Stata output represent? Briefly explain. What is the conclusion of this test? (**5 points**)

Question 3. A researcher is using matching to estimate the effect of treatment T on outcome Y. She has data on two sets of individual characteristics: X and W. She matches treatment and untreated observations on characteristics X but omits W from her model. For each of the statements below, indicate whether the statement is true or false. If the statement is false, carefully explain in 1-2 sentences what is wrong with it. (**14 points—2 points each**)

(a) If conditioning on X satisfies the conditional independence assumption, $E[Y_i(0)|X_i, D_i = 0] = E[Y_i(0)|X_i, D_i = 1]$, where $Y_i(0)$ is an individual's value of Y in the untreated state and D indicates treatment status.

(b) A good test for conditional independence is to compare mean values of $Y(0)$ in the treated and untreated groups.

(c) The omission of W from the matching procedure will cause omitted variables bias.

(d) Treated and untreated observations that are matched on their propensity score have the same X but not the same Y.

(e) If the conditional independence assumption holds when conditioning on X, it holds when conditioning on $P(D_i = 1|X_i)$.

(f) In nearest neighbor matching with replacement, untreated observations with high values of the propensity score are likely to be matched to multiple treated observations. Compared to matching without replacement, this reduces bias but increases standard errors.

(g) Mahalanobis matching identifies treated and untreated observations with a similar propensity to be treated.

Question 4. You are interested in the effects of completing a Career and Technical Education (CTE) dual enrollment pathway on post-secondary enrollment and persistence. This pathway offers students a CTE curriculum that enables them to complete both a diploma and a 2-year associate's degree during high school. Imagine you have collected a dataset containing records for all high school students in your state, including participation in a CTE dual enrolment pathway. These records have been linked to post-secondary outcomes including 2- and 4-year college enrollment and degree completion. (**21 points**)

(a) You are considering the OLS regression model below, where Y_i is student i's college enrollment or degree completion and $CTE_i = 1$ if the student participated in a CTE dual enrollment pathway (= 0 otherwise). Your colleague is concerned this regression suffers from omitted variables bias (OVB). Speculate on the likely direction of OVB, and carefully justify your answer using the "short" vs. "long" regression mnemonic used in class. You may choose an exemplar omitted variable as an illustration. (**10 points**)

$$Y_i = \beta_0 + \beta_1 CTE_i + u_i$$

(b) To mitigate OVB you opt for a nearest neighbor matching approach in which you match students who participated in a CTE dual enrollment pathway to similar students in the same school district who did not. Your matching approach includes a long list of covariates measured prior to high school (e.g., previous test scores and economic disadvantage). You then compare the mean outcomes of CTE dual enrollment students to those of their matched counterparts to estimate an ATT. Carefully explain the assumptions required for this method to provide plausibly causal estimates of the ATT of CTE dual enrollment on post-secondary outcomes. (**8 points**)

(c) Do you think the assumptions described in part (b) are likely to hold here? Briefly explain why or why not. (**3 points**)