LPO 8852: Regression II
Vanderbilt University
Midterm Exam
October 11, 2022

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. Please submit your responses via email to `sean.corcoran@vanderbilt.edu` before 11:59 pm on Wednesday, October 12. Good luck!

Question 1. Kearney & Levine (2019) used a difference-in-differences design to estimate the impact of *Sesame Street* (the popular children's television program on PBS) on short- and long-run educational outcomes. *Sesame Street* was first televised in 1969, a time when few children—and especially low-income children—attended pre-kindergarten programs. *Sesame Street* emphasized basic literacy and numeracy, and aimed to reduce gaps in school readiness.

The authors used individual-level Census data, focusing on individuals born between 1959 and 1969. Children born in 1963 and earlier would have entered school before *Sesame Street* debuted, while those born after 1963 were potentially exposed to the show during their pre-kindergarten years. The 1980 Census includes members of these cohorts when they were aged 12-21; the 1990 Census includes cohort members when they were 22-31, and so on. The Census data provide educational attainment (e.g., HS or college completion), labor market outcomes (e.g., employment and earnings), and—for cohort members still of school age in 1980—an indication of whether they were enrolled in the appropriate grade for their age (i.e., not previously retained). (**50 points**)

(a) One approach to this study might be to compare the mean outcomes of younger cohorts (exposed to *Sesame Street*) to those of older cohorts (not exposed). This could be done in a regression framework, in which other characteristics of the individuals are used as covariates, such as race and county of residence. What is the chief disadvantage of this design, if the aim is identify the causal effect of exposure to *Sesame Street*? Assume that we are interested in the effect of exposure, not actual viewing, since we have no way of knowing who watched the program. (**10 points**)

(b) *Sesame Street* was broadcast nationwide, but Kearney & Levine argue that its availability was uneven. In those days, the handful of stations Americans could access were broadcast on VHF or UHF frequencies. PBS was frequently broadcast on UHF, which was harder for households to access. The authors created a "coverage" rate ($SSCov_j$) ranging from 0-1 representing the share of a county's households that could likely access PBS in 1969, based on whether it was on UHF or VHF, the average distance from homes to the TV tower, etc. They then estimated the following regression using individual Census data:

$$(1) \qquad \boldsymbol{Outcome}_{ijc} = \beta_0 + \beta_1 \times (\boldsymbol{preschool69}_{ic} \times \boldsymbol{SSCov}_j) + \beta_2 \boldsymbol{Policy}_{jc} + \beta_3 \boldsymbol{X}_{ijc} + \gamma_c \times \gamma_s + \delta_j + \varepsilon_{ijc},$$

i is individual, j is county, and c is age cohort. $preschool69_{ic}$ is equal to 1 if the individual was of preschool age in 1969 or later (i.e., the later birth cohort), and X_{ijc} and $Policy_{jc}$ are controls. $\gamma_c \times \gamma_s$ are age cohort-by-state effects, and δ_j are county dummies. $\boldsymbol{\beta_1}$ is the coefficient of interest here. Carefully explain what this coefficient represents, and explain why this should be thought of as a difference-in-differences. (**10 points**)

(c) Under what assumption(s) can we interpret the coefficient β_1 as the causal effect of exposure to *Sesame Street*? Be as specific as you can to this example. How might you validate these assumptions? Briefly explain. (**10 points**)

(d) Table 4 from Kearney & Levine (reproduced on the next page) reports their estimate of the effect of exposure to *Sesame Street* on grade-for-age status. The outcome is $= 1$ if the individual was "on track," that is, enrolled in the expected grade for their age, and $= 0$ otherwise. Carefully provide a written interpretation of the first coefficient in column 1 (*Aggregate effect* under *All*). (**10 pts**)

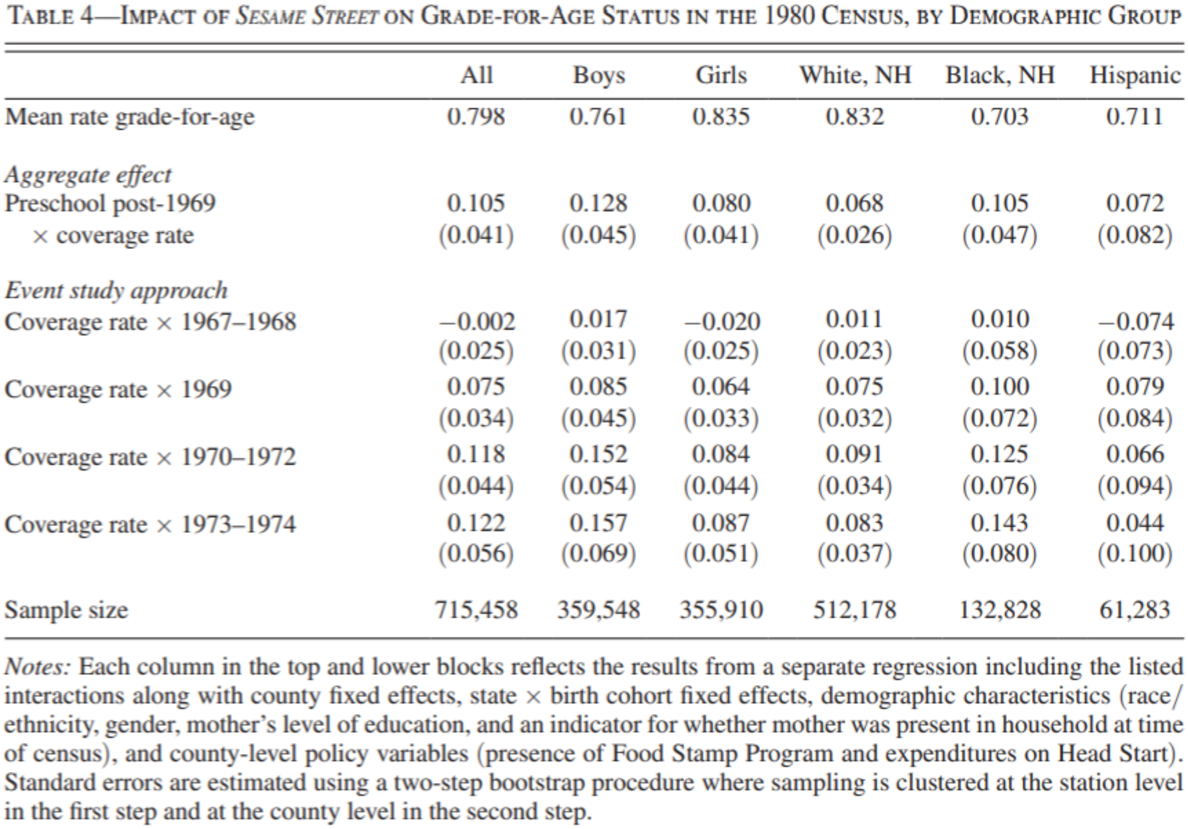

TABLE 4—IMPACT OF *SESAME STREET* ON GRADE-FOR-AGE STATUS IN THE 1980 CENSUS, BY DEMOGRAPHIC GROUP

	All	Boys	Girls	White, NH	Black, NH	Hispanic
Mean rate grade-for-age	0.798	0.761	0.835	0.832	0.703	0.711
Aggregate effect						
Preschool post-1969	0.105	0.128	0.080	0.068	0.105	0.072
× coverage rate	(0.041)	(0.045)	(0.041)	(0.026)	(0.047)	(0.082)
Event study approach						
Coverage rate × 1967–1968	−0.002	0.017	−0.020	0.011	0.010	−0.074
	(0.025)	(0.031)	(0.025)	(0.023)	(0.058)	(0.073)
Coverage rate × 1969	0.075	0.085	0.064	0.075	0.100	0.079
	(0.034)	(0.045)	(0.033)	(0.032)	(0.072)	(0.084)
Coverage rate × 1970–1972	0.118	0.152	0.084	0.091	0.125	0.066
	(0.044)	(0.054)	(0.044)	(0.034)	(0.076)	(0.094)
Coverage rate × 1973–1974	0.122	0.157	0.087	0.083	0.143	0.044
	(0.056)	(0.069)	(0.051)	(0.037)	(0.080)	(0.100)
Sample size	715,458	359,548	355,910	512,178	132,828	61,283

Notes: Each column in the top and lower blocks reflects the results from a separate regression including the listed interactions along with county fixed effects, state × birth cohort fixed effects, demographic characteristics (race/ethnicity, gender, mother's level of education, and an indicator for whether mother was present in household at time of census), and county-level policy variables (presence of Food Stamp Program and expenditures on Head Start). Standard errors are estimated using a two-step bootstrap procedure where sampling is clustered at the station level in the first step and at the county level in the second step.

(e) Finally, the bottom section of Table 4 is described as an "event study." The years here represent when a cohort was expected to enter 1st grade (e.g., 1969 means individuals in that cohort were likely to enter school in fall 1969). (1) Briefly explain how these coefficients should be interpreted, focusing on the *All* column, and (2) explain how the first event study coefficient (Coverage rate × 1967-1968) supports the main causal identification assumption of the DD design (or not). (**10 points**)

Question 2. 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 control 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 unconfoundedness, $E(Y_0|X, T = 1) = E(Y_0|X, T = 0)$, where Y_0 is an individual's value of Y in the untreated state.

(b) A test for unconfoundedness is to compare values of Y_0 in the treated and untreated groups.

(c) The omission of W from the matching procedure will result in 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) "Common support" requires that all untreated observations have propensity scores at least as high as the smallest propensity score observed among the treated observations.

(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.

VISTAVISION
YOU'LL LAUGH... YOU'LL LOVE...
PARAMOUNT PRESENTS THE DON HARTMAN PRODUCTION OF
The Matchmaker
STARRING SHIRLEY BOOTH · ANTHONY PERKINS
SHIRLEY MacLAINE · PAUL FORD
DIRECTED BY JOSEPH ANTHONY
PRODUCED BY DON HARTMAN
SCREENPLAY BY JOHN MICHAEL HAYES
FROM THE PLAY BY THORNTON WILDER

Question 3. You are interested in the effects of attending an "early college" high school on post-secondary educational attainment and earnings. Early college high schools allow students to complete both a high school 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 school attended. These records have been linked to post-secondary outcomes including 2- and 4-year college enrollment and degree completion, and labor market earnings 6 years after expected graduation. (**35 points**)

(a) You begin by fitting the OLS regression model below, where Y_i is individual i's college enrollment, degree completion, or earnings 6 years after expected graduation and $EC_i = 1$ if the individual attended an early college high school ($= 0$ otherwise). Does the population regression function being estimated in this case represent a conditional expectation function? Should the PRF (or CEF) be interpreted as *causal*? Briefly explain why or why not. (**5 points**)

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

(b) Your colleague suspects the regression in part (a) suffers from omitted variables bias. Speculate on the likely direction of OVB in this case, and carefully justify your answer using the "short" vs. "long" regression mnemonic used in class. To do this, you may choose an exemplar omitted variable as an illustration. (**10 points**)

(c) To mitigate OVB you opt for a nearest neighbor matching approach in which you match students who attended an early college high school to similar students in the state who did not attend an early college high school. Your matching model includes a long list of covariates measured prior to high school. (These include 8th grade test scores and other student variables such as gender, free or reduced price meals eligibility, special education status, limited English proficiency, etc.). You then compare the mean outcomes of early college high school attendees to that of the matched comparison group to estimate an ATT. Carefully explain the assumptions required for this method to provide plausibly causal estimates of the effect of attending an early college high school on later outcomes. (**10 points**)

(d) Do you think the assumptions described in part (c) are likely to hold here? Briefly explain why or why not. (**5 points**)

(e) You are considering a regression estimated using the matched sample that looks like the following, where X_i is a list of covariates used in the propensity score model and W_i is a set of covariates measured during high school, such as high school GPA, credits completed, and test scores. Would this be a sensible strategy? Why or why not? (**5 points**)

$$Y_i = \beta_0 + \beta_1 EC_i + \alpha X_i + \gamma W_i + u_i$$